A Weak Convergence Approach to the Theory of Large Deviations

Paul Dupuis
Lefschetz Center for Dynamical Systems
Division of Applied Mathematics
Brown University
Providence, RI 02912

and

Richard S. Ellis
Department of Mathematics and Statistics
University of Massachusetts
Amherst, MA 01003
## Contents

<table>
<thead>
<tr>
<th>Preface</th>
<th>v</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventions</td>
<td>viii</td>
</tr>
</tbody>
</table>

### 1 The Laplace Principle

- 1.1 Introduction .......................... 1
- 1.2 Equivalent Formulation .................. 4
- 1.3 Basic Results .......................... 14
- 1.4 Properties of the Relative Entropy ............. 26
- 1.5 Dynamic Programming .................. 32

### 2 Sanov’s Theorem

- 2.1 Introduction .......................... 39
- 2.2 Statement of Sanov’s Theorem ............. 40
- 2.3 The Representation Formula .................. 42
- 2.4 Proof of Lower Bound .................. 47
- 2.5 Proof of Upper Bound .................. 48

### 3 Mogulskii’s Theorem

- 3.1 Introduction .......................... 53
- 3.2 The Representation Formula .................. 55
- 3.3 Proof of Upper Bound .................. 61
- 3.4 Statement of Theorem and Completion of Proof .......... 66
- 3.5 Cramér’s Theorem .................. 70
- 3.6 Comments on the Proofs .................. 71

### 4 Other Representation Formulas

- 4.1 Introduction .......................... 75
- 4.2 Empirical Measures of a Markov Chain ............. 77
- 4.3 Random Walk Model .................. 80
- 4.4 Model with State–Dependent Noise .................. 83
- 4.5 Extensions to Unbounded Functions ............. 87
- 4.6 Continuous–Time Markov Processes ............. 90
  - 4.6.1 Introduction .................. 90
## 4.6.2 A Formal Derivation of Representation Formulas for Continuous-Time Markov Processes

92

## 4.6.3 Examples of Continuous-Time Representation Formulas

97

## 4.6.4 Remarks on the Proofs of the Representation Formulas

100

## 5 Compactness and Limit Properties

103

5.1 Introduction

103

5.2 Representation Formula

104

5.3 Compactness and Limit Properties

106

5.4 A Weaker Version of Condition 5.3.1

120

## 6 Random Walk with Continuous Statistics

123

6.1 Introduction

123

6.2 Proof of Upper Bound

124

6.3 Statement of the Laplace Principle

134

6.4 Strategy for Proof of Lower Bound

140

6.5 First Proof of Lower Bound

146

6.6 Second Proof of Lower Bound

152

6.7 An Extension of Theorem 6.3.3

169

## 7 Random Walk with Discontinuous Statistics

177

7.1 Introduction

177

7.2 Statement of the Laplace Principle

179

7.3 Another Laplace Principle and Examples

182

7.4 Proof of Upper Bound

185

7.5 Proof of the Laplace Principle Lower Bound

202

7.6 Compactness of the Level Sets of \( I_x \)

220

## 8 Empirical Measures of a Markov Chain

225

8.1 Introduction

225

8.2 Compactness and Limit Properties

227

8.3 Proof of Upper Bound

238

8.4 Statement of the Laplace Principle

243

8.5 Properties of the Rate Function

245

8.6 Proofs of Lower Bounds

249

## 9 Extensions of Laplace Principle

261

9.1 Introduction

261

9.2 Discontinuous Statistics

263

9.3 Empirical Measures in the \( \tau \)-Topology

271

## 10 Continuous-Time Markov Processes

285

10.1 Introduction

285

10.2 Statement of the Laplace Principle

285

10.3 Proof of the Laplace Principle

291
A Background Material 301
A.1 Introduction ......................................................... 301
A.2 Measure Theory ..................................................... 301
A.3 Weak Convergence of Probability Measures ..................... 302
A.4 Probability Theory .................................................. 312
A.5 Stochastic Kernels ................................................... 315
A.6 Analysis .............................................................. 322

B Representation Formulas 325
B.1 Introduction .......................................................... 325
B.2 Representation Formula for Sanov’s Theorem ....................... 325
B.3 Discussion ............................................................. 328

C Proofs of a Number of Results 329
C.1 Introduction .......................................................... 329
C.2 Donsker–Varadhan Variational Formula ........................... 329
C.3 Proof of the Chain Rule and Part (f) of Lemma 1.4.3 ............. 332
C.4 Proof of Part (g) of Lemma 1.4.3 .................................. 334
C.5 Proof of Part (f) of Lemma 6.2.3 ................................... 337
C.6 Proof of Part (g) of Lemma 6.2.3 ................................... 340
C.7 Proof of Proposition 6.3.4 ......................................... 343
C.8 A Continuity Property of Cramér Functions ....................... 345

D Convex Functions 349
D.1 Introduction .......................................................... 349
D.2 Background Material on Convex Functions ......................... 349
D.3 Legendre–Fenchel Transform of Compositions ..................... 353
D.4 Three Examples ....................................................... 361

E Proof of Theorem 5.3.5 Under Another Condition 365
E.1 Introduction .......................................................... 365
E.2 Proofs of Results ...................................................... 365

Bibliography 373
Preface

The theory of large deviations is one of the most active areas in probability today, having applications to many areas including statistics, communication networks, information theory, statistical mechanics, risk-sensitive control, and queueing systems. The richness of the theory of large deviations derives in part from the fact that it is nonlinear, and this in turn contributes to much of the difficulty of large deviation analysis. An intriguing question arises. Is it possible to apply to this nonlinear field the tools of the highly developed, linear theory of the weak convergence of probability measures? Although large deviation theory is formulated as an analogy to weak convergence theory and although a number of parallels in the structures of the two theories have been discovered, to date no one has applied weak convergence theory in a consistent way to prove large deviation results. To show how this can be done is the purpose of our book. We develop the weak convergence approach essentially from scratch, illustrate it via several basic examples, and then apply it to a number of sophisticated models. Much of the material in this book is being published here for the first time.

Traditionally the theory of large deviations is formulated in terms of the asymptotic behavior of normalized logarithms of probabilities of certain events. These probabilities, in turn, can be written as expectations involving discontinuous functions. Taking our cue from the theory of weak convergence, we introduce the technically useful innovation of replacing these discontinuous functions by continuous functions. This replacement greatly simplifies the asymptotic analysis. We refer to the asymptotic analysis of the normalized logarithms of such expectations as the Laplace principle, which in the setting of Polish spaces is equivalent to the large deviation principle. A basic step in our approach is to represent the normalized logarithms of the expectations as variational formulas that are then interpreted, in a natural way, as the minimal cost functions of associated stochastic optimal control problems. These minimal cost functions have a form to which the theory of weak convergence of probability measures can be applied. We consider the introduction of representation formulas to be one of our main contributions. They allow us to replace the exponential estimates of standard large deviation approaches by law of large numbers–type estimates, which are obtained by the theory of weak convergence.

This book is easily accessible to anyone who has taken courses in measure theory and measure–theoretic probability. We emphasize that no background in control theory is assumed or required. Indeed, control theory is used mainly to provide intuition and a convenient terminology.

To give an outline of the weak convergence approach, let us consider a sequence of random variables \( \{X^n, n \in \mathbb{N}\} \) taking values in a Polish space \( \mathcal{X} \) and let \( I \) be a rate
function on $\mathcal{X}$ (see Section 1.1). The sequence $\{X^n\}$ is said to satisfy the Laplace principle with rate function $I$ if for all bounded continuous functions $h$ mapping $\mathcal{X}$ into $\mathbb{R}$

$$
\lim_{n \to \infty} \frac{1}{n} \log E \left\{ e^{-n h(X^n)} \right\} = - \inf_{x \in \mathcal{X}} \{ h(x) + I(x) \}.
$$

Let $\mathcal{P}(\mathcal{X})$ denote the set of probability measures on $\mathcal{X}$ and for $\gamma$ and $\theta$ in $\mathcal{P}(\mathcal{X})$ write $R(\gamma \| \theta)$ for the relative entropy of $\gamma$ with respect to $\theta$. A key step in the weak convergence approach is to rewrite the normalized logarithm of the expectation in the last display as a variational formula over $\mathcal{P}(\mathcal{X})$. Namely, if $\theta^n$ represents the distribution of $X^n$, then as we will see in Proposition 1.4.2, it is extremely easy to prove that

$$
- \frac{1}{n} \log E \left\{ e^{-n h(X^n)} \right\} = - \frac{1}{n} \log \int_{\mathcal{X}} e^{-n h} \, d\theta^n = \inf_{\mu \in \mathcal{P}(\mathcal{X})} \left\{ \int_{\mathcal{X}} h \, d\mu + \frac{1}{n} R(\mu \| \theta^n) \right\}.
$$

The goal is to identify a rate function $I$ on $\mathcal{X}$ such that as $n \to \infty$ the last term in this display converges to $\inf_{x \in \mathcal{X}} \{ h(x) + I(x) \}$. The convergence is carried out by means of weak convergence methods. In the case where $h + I$ uniquely attains its infimum at a point $x^*$, $I(x^*)$ is identified as the limit of normalized relative entropies.

The representation formula given in the last display is a cornerstone of the weak convergence approach. One exploits structural properties of $\theta^n$ to obtain, in each problem, the most convenient form of the representation for evaluating the limit. An innovative use of dynamic programming allows us to carry this out in an efficient way and to reinterpret the infimum appearing in the representation formula as the minimal cost function of a stochastic optimal control problem. The use of representations and of ideas from stochastic optimal control theory in large deviation problems is originally due to Wendell Fleming in the context of diffusions with small noise.

We will develop the weak convergence approach by applying it to a wide variety of stochastic models, which fall into two broad classes. The first consists of what we call random–variable level problems. Our main examples in this class are the empirical measures of i.i.d. random variables and the empirical measures of Markov chains. These empirical measure problems are analyzed in Chapters 2, 8, and 9. The second class consists of process–level problems. Our main examples are a random walk model associated with small random perturbations of dynamical systems and continuous–time Markov processes that include diffusions and jump processes. The process–level problems are analyzed in Chapters 3, 5, 6, 7, and 10. Besides the detailed analysis of a number of large deviation problems, one of the goals of this book is to facilitate the application of the weak convergence approach in other contexts. In this regard, we develop in Chapter 4 the representation formulas for three other types of large deviation problems that we will not analyze but that are nevertheless interesting. This chapter also highlights the central role played by relative entropy in every representation formula. Properties of this function are developed in detail in Section 1.4. The remainder of Chapter 1 presents other background material and proves a number of general results in large deviation theory from the Laplace principle perspective.

The use of representation formulas and weak convergence methods has an aesthetic appeal and practical consequences. In the standard statement of the large deviation
principle, the limits superior and inferior of large deviation probabilities are characterized in terms of the solution of an associated variational problem involving the rate function. Typically, this limiting variational problem can be cast as an optimal control problem. In our approach, the normalized Laplace expectations are represented as the minimal cost functions of stochastic optimal control problems. We then prove the Laplace principle by establishing the convergence of the minimal cost functions of the prelimit control problems to the minimal cost function of the limiting control problem. The prelimit and limiting control problems typically have similar forms. As a consequence, the form of the rate function in the Laplace principle is almost never a surprise and in fact is usually easy to guess early on.

Our approach has the following additional features:

- It is a unified method applying to a wide range of problems.
- Similar compactness and convergence techniques are used to derive the upper and lower limits that give the Laplace principle. The proofs of both limits start with the same representation formula.
- In all cases that we consider, the proof that the rate function has compact level sets, and thus is lower semicontinuous, is based on a deterministic analogue of the method used to show the Laplace principle upper bound.
- Large deviation theory to date has mostly been concerned with the formulation of sufficient conditions under which a large deviation principle can be proved. Since necessary conditions for convergence of integrals are well known in the theory of weak convergence of probability measures, the use of weak convergence methods allows one to understand more easily which assumptions are needed for a large deviation principle to hold with a given rate function.
- The representation formulas are a natural starting point for the construction of numerical approximations to the limiting variational problem involving the rate function. Numerical analysis of large deviation problems is a topic that has largely been ignored. The compactness and convergence results developed to prove the Laplace principle should prove useful in this regard.
- The use of representations and weak convergence methods allow one to avoid awkward discretizations that appear in other approaches.
- The representation formulas provide a link to other topics of interest, such as risk-sensitive control and robust control, Doob’s $h$-transform, and the construction of conditioned processes.

The power of the weak convergence approach can be seen in the study of Markov processes with discontinuous statistics, which arise when modeling queueing systems, in the study of computer and communication networks, in the analysis of controlled diffusions with discontinuous feedback controls, and in other areas. Markov processes with discontinuous statistics are characterized by the fact that the components of their generators
do not depend smoothly on the spatial parameter. As a result of this nonsmooth dependence, classical methods for studying these processes are not applicable. An example in which the weak convergence approach is particularly useful is given in Chapter 7 of the book, where we study a random walk model with discontinuous statistics. Markov processes with continuous statistics will also be analyzed. These processes are characterized by the fact that the components of their generators depend smoothly on the spatial parameter. In the cases of both continuous statistics and discontinuous statistics, the weak convergence approach enables us not only to reproduce known results, often under weaker conditions, but also to treat new problems.

We would like to thank a number of people for their help. Christopher Boucher, Michelle Boué, Gordon Kieffer, and Kavita Ramanan each read parts of the manuscript closely and suggested improvements. Andrew Barron, Alejandro de Acosta, Amir Dembo, Wendell Fleming, Joseph Horowitz, Steven Shreve, and Srinivasa Varadhan made useful comments that were incorporated into the book. We gratefully acknowledge the support of the National Science Foundation and thank the Division of Applied Mathematics at Brown University and the Department of Mathematics and Statistics at the University of Massachusetts for providing excellent working environments. One of us (P.D.) also thanks the Army Research Office for its support. Our greatest debt of gratitude is owed to Ofer Zeitouni, whose many insightful remarks significantly affected the final form of this book.

Conventions

We adopt the following conventions throughout the book.

- Subsequences are labeled by the same index as the original sequence.
- The $\sigma$–field on a Polish space is always taken to be the Borel $\sigma$–field.
- By a probability measure on a Polish space we mean a Borel probability measure.
- A function mapping a Polish space into a Polish space is called measurable if it is measurable with respect to the underlying Borel $\sigma$–fields.
- Since all control problems that we consider are optimal control problems, we will drop the term “optimal.”
- The usual rules for arithmetic involving $\infty$ are adopted; e.g., $0 \cdot \infty = 0$.
- We abbreviate “almost surely” as “a.s.” and “with probability 1” as “w.p.1.”

For historical background on large deviations the reader is referred to the texts of Dembo and Zeitouni and of Deuschel and Stroock.
Chapter 1

A Formulation of Large Deviation Theory in Terms of the Laplace Principle

1.1 Introduction

This book presents an approach to large deviation theory which is based on evaluating the asymptotics of certain expectations. We will begin the exposition after making a few remarks about large deviations.

Throughout this chapter \( \{X^n, n \in \mathbb{N}\} \) is a sequence of random variables defined on a probability space \( (\Omega, \mathcal{F}, P) \) and taking values in a complete separable metric space \( \mathcal{X} \). As is usual, we will refer to such a space as a Polish space. The metric of \( \mathcal{X} \) is denoted by \( d(x, y) \) and expectation with respect to \( P \) by \( E \). The theory of large deviations focuses on random variables \( \{X^n\} \) for which the probabilities \( P\{X^n \in A\} \) converge to 0 exponentially fast for a class of Borel sets \( A \). The exponential decay rate of these probabilities is expressed in terms of a function \( I \) mapping \( \mathcal{X} \) into \([0, \infty]\). A function \( I \) on \( \mathcal{X} \) is called a rate function on \( \mathcal{X} \), or simply a rate function, if \( I \) maps \( \mathcal{X} \) into \([0, \infty]\) and if for each \( M < \infty \) the level set \( \{x \in \mathcal{X} : I(x) \leq M\} \) is a compact subset of \( \mathcal{X} \). We summarize the last property by saying that \( I \) has compact level sets. A function having compact level sets is automatically lower semicontinuous and it attains its infimum on any nonempty closed set.

We next define the concept of a large deviation principle. For \( A \) a subset of \( \mathcal{X} \), we denote \( \inf_{x \in A} I(x) \) by \( I(A) \).

**Definition 1.1.1.** Let \( I \) be a rate function on \( \mathcal{X} \). The sequence \( \{X^n\} \) is said to satisfy the large deviation principle on \( \mathcal{X} \) with rate function \( I \) if the following two conditions hold.

(a) **Large deviation upper bound.** For each closed subset \( F \) of \( \mathcal{X} \)

\[
\limsup_{n \to \infty} \frac{1}{n} \log P\{X^n \in F\} \leq -I(F).
\]
(b) **Large deviation lower bound.** For each open subset $G$ of $\mathcal{X}$

$$\liminf_{n \to \infty} \frac{1}{n} \log P\{X^n \in G\} \geq -I(G).$$

It is well known that if a sequence of random variables satisfies the large deviation principle with some rate function, then the rate function is unique. We will give a new proof of this fact in Theorem 1.3.1.

A number of authors such as [28] and [29] use the following slightly different terminology, which we do not adopt. A “rate function” on $\mathcal{X}$ is a function that maps $\mathcal{X}$ into $[0, \infty]$ and is lower semicontinuous. A “good rate function” is a function that maps $\mathcal{X}$ into $[0, \infty]$ and has compact level sets. While a few of our results hold without the assumption of compact level sets, for ease of exposition we will not point out which ones.

Here are two examples of large deviation principles.

**Example 1.1.2.** The first example is a special case of Cramér’s Theorem, a general form of which is stated in Theorem 3.5.1. Given positive numbers $p$ and $q$ summing to 1, let $\{v_j, j \in \mathbb{N}_0\}$ be a sequence of independent, identically distributed (i.i.d.) random variables taking values in $\mathbb{R}$ and having the common distribution $P\{v_j = 0\} = q$ and $P\{v_j = 1\} = p$. For $n \in \mathbb{N}$ we define the sample means

$$S^n_n = \frac{1}{n} \sum_{j=0}^{n-1} v_j.$$ 

According to Theorem 3.5.1, the sequence $\{S^n/n\}$ satisfies the large deviation principle on $\mathbb{R}$ with rate function

$$I(x) \doteq \sup_{\alpha \in \mathbb{R}} \left\{\alpha x - \log \int_{\mathbb{R}} e^{\alpha y} \rho(dy)\right\},$$

where $\rho = q\delta_0 + p\delta_1$. The supremum can be explicitly evaluated to give

$$I(x) \doteq \begin{cases} x \log \left(\frac{x}{p}\right) + (1-x) \log \left(\frac{1-x}{q}\right) & \text{if } x \in [0,1] \\ \infty & \text{if } x \in \mathbb{R} \setminus [0,1]. \end{cases}$$

This completes the first example. ■

**Example 1.1.3.** The second example is known as Schilder’s Theorem [81], which is a special case of Theorem 10.2.6. Another proof is given, for example, in Theorem 5.2.3 in [28]. Let $\{w(t), t \in [0,1]\}$ denote a standard Brownian motion taking values in $\mathbb{R}^d$ and for $n \in \mathbb{N}$ consider the process $Y^n = \{Y^n(t), t \in [0,1]\}$ defined by

$$Y^n(t) \doteq \frac{1}{\sqrt{n}} w(t).$$

$Y^n$ takes values in the space $C([0,1]: \mathbb{R}^d)$ consisting of all continuous functions $\varphi$ that map $[0,1]$ into $\mathbb{R}^d$, and $Y^n$ satisfies $Y^n(0) = 0$. When equipped with the supremum
norm, $C([0, 1] : \mathbb{R}^d)$ is a Banach space and thus a Polish space. Let $\mathcal{A}_0([0, 1] : \mathbb{R}^d)$ denote the subset consisting of all absolutely continuous functions $\varphi$ satisfying $\varphi(0) = 0$. Schilder’s Theorem states that the sequence $\{Y^n\}$ satisfies the large deviation principle on $C([0, 1] : \mathbb{R}^d)$ with rate function

$$I(\varphi) \triangleq \begin{cases} 
\frac{1}{2} \int_0^1 \| \dot{\varphi}(t) \|^2 dt & \text{if } \varphi \in \mathcal{A}_0([0, 1] : \mathbb{R}^d) \\
\infty & \text{if } \varphi \in C([0, 1] : \mathbb{R}^d) \setminus \mathcal{A}_0([0, 1] : \mathbb{R}^d).
\end{cases}$$

This completes the second example. ■

Suppose we would like to prove that a sequence $\{X^n\}$ satisfies the large deviation principle. As we will show in the next section, the large deviation principle will follow provided we can evaluate, for all bounded continuous functions $h$ mapping $\mathcal{X}$ into $\mathbb{R}$, the asymptotics of quantities of the form

$$\frac{1}{n} \log E\{\exp[-n h(X^n)]\} \quad (1.1)$$

as $n \to \infty$. The weak convergence approach is ideally suited to such evaluations. The main purpose of this chapter is to introduce some of the machinery that will enable us to implement it. In a nutshell, the basic step is to associate with the given large deviation problem a sequence of stochastic control problems, the minimal cost functions of which will give a variational representation for the quantities in the last display. In turn, the asymptotics of the minimal cost functions will be determined by weak convergence methods. Two important tools needed to prove the variational representation are given in Sections 1.4 and 1.5. They are a variational formula involving the relative entropy and the dynamic programming equation. In the subsequent two chapters we will give two basic examples that will illustrate in detail how the procedure works. Other methods for proving representations are briefly discussed at the end of Chapter 4.

The evaluation of the asymptotics of quantities of the form given in equation (1.1) leads to the concept of the Laplace principle. In a sense to be made precise in Section 1.2, the Laplace principle is equivalent to the large deviation principle. In Section 1.3 we make a small but pleasant detour as we reformulate, in terms of the Laplace principle, a number of basic results in the theory. Although most of the theorems in Section 1.3 will not be applied in the sequel, we have included this section in order to illustrate the naturalness of the reformulation and the relative ease with which such theorems can be proved. The results to be considered in this section include the contraction principle and beautiful theorems of O’Brien and Verwaal [76], Pukhalskii [78], and Bryc [16] on the existence of a large deviation principle for sequences of random variables that are exponentially tight.

The techniques used in Section 1.2 to prove the equivalence of the Laplace principle and the large deviation principle as well as the techniques used in Section 1.3 are basic in large deviation theory. However, they will not be used in the remainder of the book. In contrast, the techniques used in Section 1.4 to establish properties of the relative entropy will be applied many times.
Notation. We will implement the weak convergence approach by studying the asymptotics of
\[ W^n \doteq - \frac{1}{n} \log E\{\exp[-n h(X^n)]\} \]
as \( n \to \infty \). \( W^n \) is the negative of the quantity given in equation (1.1). By inserting the two annoying minus signs in the formula for \( W^n \), we avoid the more annoying minus signs that would arise in the asymptotic analysis.

1.2 An Equivalent Formulation of the Large Deviation Principle

In [87] Varadhan proved an important consequence of the large deviation principle which involves the asymptotic behavior of certain expectations. It generalizes the well known method of Laplace for studying the asymptotics of certain integrals on \( IR \). Given \( h \) a bounded continuous function mapping \([0,1]\) into \( IR \), Laplace’s method states that

\[
\lim_{n \to \infty} \frac{1}{n} \log \int_0^1 \exp[-n h(x)] \, dx = - \min_{x \in [0,1]} h(x).
\]

The proof is a straightforward exercise, and if \( h \) is smooth, then further analysis yields an asymptotic expansion for the integral as \( n \to \infty \). After stating Varadhan’s result in the next theorem, we will explain its relationship to the weak convergence approach. We remind the reader that throughout this chapter \( \{X^n, n \in \mathbb{N}\} \) is a sequence of random variables defined on a probability space \( (\Omega, \mathcal{F}, P) \) and taking values in a Polish space \( \mathcal{X} \). The metric of \( \mathcal{X} \) is denoted by \( d(x,y) \) and expectation with respect to \( P \) by \( E \).

**Theorem 1.2.1 (Varadhan).** Assume that the sequence \( \{X^n\} \) satisfies the large deviation principle on \( \mathcal{X} \) with rate function \( I \). Then for all bounded continuous functions \( h \) mapping \( \mathcal{X} \) into \( IR \)

\[
\lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\} = - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}. \tag{1.2}
\]

More precisely, the following conclusions hold.

(a) The large deviation upper bound implies that

\[
\limsup_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\} \leq - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}.
\]

(b) The large deviation lower bound implies that

\[
\liminf_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\} \geq - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}.
\]
1.2. EQUIVALENT FORMULATION

**Proof.** (a) There exists $M \in (0, \infty)$ such that $-M \leq h(x) \leq M$ for all $x \in \mathcal{X}$. For $N$ a positive integer and $j \in \{1, 2, \ldots, N\}$, consider the closed sets

$$F_{N,j} \doteq \left\{ x \in \mathcal{X} : -M + \frac{2(j - 1)M}{N} \leq -h(x) \leq -M + \frac{2jM}{N} \right\}.$$ 

The large deviation upper bound yields

$$\limsup_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \ h(X^n)]\} \leq \max_{j \in \{1, 2, \ldots, N\}} \left\{ -M + \frac{2jM}{N} - I(F_{N,j}) \right\}$$

$$\leq \max_{j \in \{1, 2, \ldots, N\}} \sup_{x \in F_{N,j}} \{ -h(x) - I(x) \} + \frac{2M}{N}$$

$$\leq \sup_{x \in \mathcal{X}} \{ -h(x) - I(x) \} + \frac{2M}{N}.$$ 

Sending $N \to \infty$, we obtain

$$\limsup_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \ h(X^n)]\} \leq \sup_{x \in \mathcal{X}} \{ -h(x) - I(x) \} = - \inf_{x \in \mathcal{X}} \{ h(x) + I(x) \},$$

as claimed.

(b) Given $x$ an arbitrary point in $\mathcal{X}$ and $\varepsilon$ an arbitrary positive number, we apply the large deviation lower bound to the open set $G \doteq \{ y \in \mathcal{X} : h(y) < h(x) + \varepsilon \}$, obtaining

$$\liminf_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \ h(X^n)]\} \geq \liminf_{n \to \infty} \frac{1}{n} \log E\{1_{G}(X^n) \ \exp[-n \ h(X^n)]\}$$

$$\geq -h(x) - \varepsilon + \liminf_{n \to \infty} \frac{1}{n} \log P\{X^n \in G\}$$

$$\geq -h(x) - \varepsilon - I(G)$$

$$\geq -h(x) - I(x) - \varepsilon.$$ 

Since $x \in \mathcal{X}$ and $\varepsilon > 0$ are arbitrary,

$$\liminf_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \ h(X^n)]\} \geq - \inf_{x \in \mathcal{X}} \{ h(x) + I(x) \}.$$

This completes the proof of the theorem. ■

If we summarize the large deviation principle by the formal notation

$$P\{X^n \in dx\} \propto \exp[-n \ I(x)] \ dx,$$
then we can write
\[
E\{\exp[-n \, h(X^n)]\} = \int_{\mathcal{X}} \exp[-n \, h(x)] \, P\{X^n \in dx\} \propto \int_{\mathcal{X}} \exp[-n(h(x) + I(x))] \, dx.
\]

As in Laplace’s method, Varadhan’s Theorem states that to exponential order the main contribution to the integral is due to the largest value of the exponent.

It is convenient to coin phrases to refer to the validity of the limit (1.2) for all bounded continuous functions \( h \) as well as to the validity of the upper and lower bounds in parts (a) and (b) of the theorem.

**Definition 1.2.2.** Let \( I \) be a rate function on \( \mathcal{X} \). The sequence \( \{X^n\} \) is said to satisfy the Laplace principle on \( \mathcal{X} \) with rate function \( I \) if for all bounded continuous functions \( h \)
\[
\lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \, h(X^n)]\} = - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}.
\]

The term **Laplace principle upper bound** refers to the validity of
\[
\limsup_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \, h(X^n)]\} \leq - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}
\]
for all bounded continuous functions \( h \) while the term **Laplace principle lower bound** refers to the validity of
\[
\liminf_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \, h(X^n)]\} \geq - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}
\]
for all bounded continuous functions \( h \).

Evaluating the Laplace limit for the zero function on \( \mathcal{X} \) shows that the infimum of the rate function \( I \) on \( \mathcal{X} \) equals 0. Since a function with compact level sets attains its infimum on a closed set, it follows that there exists a point \( x_0 \in \mathcal{X} \) for which \( I(x_0) = 0 \).

With the last definition, we can express the content of Varadhan’s Theorem by saying that the large deviation principle implies the Laplace principle with the same rate function. The next theorem, Theorem 1.2.3, proves the converse, highlighting a basic feature of the weak convergence approach. In many cases the weak convergence approach directly yields the Laplace principle and thus through Theorem 1.2.3 can be used to derive the large deviation principle. We will use this technique for proving the large deviation principle throughout the book. The equivalence between the Laplace principle and the large deviation principle as expressed by Theorems 1.2.1 and 1.2.3 can be regarded as an analogue of the Portmanteau Theorem [Theorem A.3.4]. The latter states the equivalence between the weak convergence of probability measures and limits involving closed and open sets. An examination of the proof of the next theorem reveals that given a rate function \( I \), the Laplace principle upper bound implies the large deviation upper bound and the Laplace principle lower bound implies the large deviation lower bound. This and other features of the theorem will be discussed later in the section.
1.2. EQUIVALENT FORMULATION

**Theorem 1.2.3.** The Laplace principle implies the large deviation principle with the same rate function. More precisely, if $I$ is a rate function on $\mathcal{X}$ and the limit

$$\lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\} = -\inf_{x \in \mathcal{X}} \{h(x) + I(x)\}$$

is valid for all bounded continuous functions $h$, then $\{X^n\}$ satisfies the large deviation principle on $\mathcal{X}$ with rate function $I$.

We will prove Theorem 1.2.3 in a moment. It is related to another converse to Varadhan’s Theorem due to Bryc. The sequence $\{X^n\}$ is said to be **exponentially tight** if for each $M \in (0, \infty)$ there exists a compact set $K$ such that

$$\limsup_{n \to \infty} \frac{1}{n} \log P\{X^n \in K^c\} \leq -M.$$

Let $C_b(\mathcal{X})$ denote the space of all bounded continuous functions mapping $\mathcal{X}$ into $\mathbb{R}$. Bryc proved that if the sequence $\{X^n\}$ is exponentially tight and if the limit

$$\Lambda(h) \doteq \lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\}$$

exists for all $h \in C_b(\mathcal{X})$, then $\{X^n\}$ satisfies the large deviation principle on $\mathcal{X}$ with rate function

$$I(x) \doteq -\inf_{h \in C_b(\mathcal{X})} \{h(x) + \Lambda(h)\}$$

and formula (1.2) is valid; i.e.,

$$\Lambda(h) = -\inf_{x \in \mathcal{X}} \{h(x) + I(x)\}.$$

See the original paper [16] and the nice exposition in Section 4.4 of [28]. Special cases and variations are given by a number of authors including [30, 44, 55, 62]. We will prove Bryc’s Theorem in the next section [Theorem 1.3.8].

Theorem 1.2.3 states that the large deviation principle follows once we have a suitable asymptotic evaluation of the expectations $E\{\exp[-n h(X^n)]\}$ for all bounded continuous functions $h$. In the cases that we will treat, we could easily modify our method and show the large deviation principle by obtaining bounds on the asymptotic behavior of expectations that involve a class of discontinuous functions rather than continuous functions. This class is large enough to contain suitable approximations to the characteristic functions of closed sets and of open balls in $\mathcal{X}$. However, just as with weak convergence theory, it is somewhat easier to deal with bounded continuous functions. For this reason we will use Theorem 1.2.3 throughout the book. We now prove the latter result.

**Proof of Theorem 1.2.3.** We assume that $I$ is a rate function on $\mathcal{X}$ and that for all bounded continuous functions $h$

$$\lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\} = -\inf_{x \in \mathcal{X}} \{h(x) + I(x)\}. $$
We want to prove that for each closed set $F$ the sequence $\{X^n\}$ satisfies the large deviation upper bound
\[
\limsup_{n \to \infty} \frac{1}{n} \log P\{X^n \in F\} \leq -I(F)
\]
and that for each open set $G$ the sequence $\{X^n\}$ satisfies the large deviation lower bound
\[
\liminf_{n \to \infty} \frac{1}{n} \log P\{X^n \in G\} \geq -I(G).
\]

**Proof of the large deviation upper bound.** Given a closed set $F$, we define the nonnegative, lower semicontinuous function
\[
\varphi(x) = \begin{cases} 
0 & \text{if } x \in F \\
\infty & \text{if } x \in F^c.
\end{cases}
\]

Let $d(x, F)$ denote the distance from $x$ to $F$ and for $j \in \mathbb{N}$ define
\[
h_j(x) = j(d(x, F) \land 1). \tag{1.3}
\]
Then $h_j$ is a bounded continuous function and $h_j \uparrow \varphi$ as $j \to \infty$. Hence
\[
\frac{1}{n} \log P\{X^n \in F\} = \frac{1}{n} \log E\{\exp[-n \varphi(X^n)]\} \leq \frac{1}{n} \log E\{\exp[-n h_j(X^n)]\},
\]
and so
\[
\limsup_{n \to \infty} \frac{1}{n} \log P\{X^n \in F\} \leq \lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h_j(X^n)]\} = -\inf_{x \in \mathcal{X}} \{h_j(x) + I(x)\}.
\]

We complete the proof by showing that
\[
\lim_{j \to \infty} \inf_{x \in \mathcal{X}} \{h_j(x) + I(x)\} = I(F). \tag{1.4}
\]
Half of this is easy. Since $h_j \leq \varphi$,
\[
\inf_{x \in \mathcal{X}} \{h_j(x) + I(x)\} \leq \inf_{x \in \mathcal{X}} \{\varphi(x) + I(x)\} = \inf_{x \in \mathcal{X}} I(x) = I(F),
\]
and thus
\[
\limsup_{j \to \infty} \inf_{x \in \mathcal{X}} \{h_j(x) + I(x)\} \leq I(F).
\]
The final step is to prove that
\[
\liminf_{j \to \infty} \inf_{x \in \mathcal{X}} \{h_j(x) + I(x)\} \geq I(F).
\]
We can assume that $I(F) > 0$ since if $I(F) = 0$ we are done. We proceed under the further assumption that $I(F) < \infty$. The routine modifications needed to handle the case $I(F) = \infty$ are omitted.
1.2. EQUIVALENT FORMULATION

Since $h_j = 0$ on $F$,

$$\inf_{x \in \mathcal{X}} \{h_j(x) + I(x)\} = \min \left( \inf_{x \in F} \{h_j(x) + I(x)\}, \inf_{x \in F^c} \{h_j(x) + I(x)\} \right)$$

$$= \min \left( I(F), \inf_{x \in F^c} \{h_j(x) + I(x)\} \right).$$

It suffices to show that

$$\liminf_{j \to \infty} \inf_{x \in F^c} \{h_j(x) + I(x)\} \geq I(F),$$

which we prove by contradiction. Thus assume that

$$\liminf_{j \to \infty} \inf_{x \in F^c} \{h_j(x) + I(x)\} < I(F).$$

Then there exists a subsequence of $j \in \mathbb{N}$ and $\varepsilon \in (0, I(F)/2)$ such that for all $j$ in this subsequence

$$\inf_{x \in F^c} \{h_j(x) + I(x)\} \leq I(F) - 2\varepsilon.$$

In addition, for each $j$ there exists $x_j \in F^c$ such that

$$h_j(x_j) + I(x_j) \leq I(F) - \varepsilon.$$

We claim that $d(x_j, F) \to 0$. Indeed, otherwise for some subsequence we would have $h_j(x_j) \geq j(d(x_j, F) \wedge 1) \to \infty$. Since $I(F) - \varepsilon < \infty$, this would contradict the last display. The convergence $d(x_j, F) \to 0$ implies that there exists a sequence $\{y_j\}$ in $F$ such that $d(x_j, y_j) \to 0$. We now use the fact that $\sup_{x \in \mathcal{X}} I(x) \leq I(F) - \varepsilon$, which is a consequence of the last display. Since $I$ has compact level sets, it follows that there exists a further subsequence and a point $x^* \in \{x \in \mathcal{X} : I(x) \leq I(F) - \varepsilon\}$ such that $d(x_j, x^*) \to 0$. Of course, $d(x_j, y_j) \to 0$ implies that $d(y_j, x^*) \to 0$. But since the subsequence $\{y_j\}$ lies in $F$, which is closed, $x^*$ must lie in $F$ and so $I(x^*) \geq I(F)$. This contradicts the fact that $x^* \in \{x \in \mathcal{X} : I(x) \leq I(F) - \varepsilon\}$. The contradiction completes the proof of the large deviation upper bound.

**Proof of the large deviation lower bound.** Let $G$ be an open set. If $I(G) = \infty$, then there is nothing to prove, so we may assume that $I(G) < \infty$. Let $x$ be any point in $G$ such that $I(x) < \infty$ and choose a real number $M > I(x)$. There exists $\delta > 0$ such that $B(x, \delta) \doteq \{y \in \mathcal{X} : d(y, x) < \delta\}$ is a subset of $G$. In terms of $M$, $x$, and $\delta$, we define

$$h(y) \doteq M \left( \frac{d(y, x)}{\delta} \wedge 1 \right). \quad (1.5)$$

This function is bounded and continuous and satisfies $h(x) = 0$, $h(y) = M$ for $y \in B(x, \delta)^c$ and $0 \leq h(z) \leq M$ for all $z \in \mathcal{X}$. We then have

$$E \{\exp[-n \, h(X^n)]\}$$

$$\leq e^{-nM} P\{X^n \in B(x, \delta)^c\} + P\{X^n \in B(x, \delta)\} \leq e^{-nM} + P\{X^n \in B(x, \delta)\},$$
and therefore
\[
\max \left( \liminf_{n \to \infty} \frac{1}{n} \log P\{X^n \in B(x, \delta)\}, -M \right) \geq \lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\} = - \inf_{y \in X} \{h(y) + I(y)\} \\
\geq -h(x) - I(x) \geq -I(x).
\]

Since \( M > I(x) \) and \( B(x, \delta) \subset G \), it follows that
\[
\liminf_{n \to \infty} \frac{1}{n} \log P\{X^n \in G\} \geq \liminf_{n \to \infty} \frac{1}{n} \log P\{X^n \in B(x, \delta)\} \geq -I(x)
\]
and thus
\[
\liminf_{n \to \infty} \frac{1}{n} \log P\{X^n \in G\} \geq - \inf\{I(x) : x \in G, I(x) < \infty\} = -I(G).
\]
This proves the large deviation lower bound. The proof of Theorem 1.2.3 is complete.

Let \( \xi \) be any point in \( X \). If we evaluate the limit (1.4) for the closed set \( F = \{\xi\} \), then it follows that knowing the Laplace limit for all bounded continuous functions yields the value of the associated rate function at any point. We record this observation in the next lemma. It leads immediately to the fact that a rate function in a Laplace principle is unique [Theorem 1.3.1].

**Lemma 1.2.4.** Let \( I \) be a rate function on \( X \). For any \( \xi \in X \) and \( j \in \mathbb{N} \) we define the bounded continuous function \( h_j \) on \( X \) by \( h_j(x) = j(d(x, \xi) \wedge 1) \). Then
\[
\lim_{j \to \infty} \inf_{x \in X} \{h_j(x) + I(x)\} = I(\xi).
\]

With the proof of Theorem 1.2.3 we have completed our basic exposition of the equivalence between the Laplace principle and the large deviation principle. We next present a technical refinement of the Laplace principle which will be extremely useful in later chapters. In order to state it, a definition is needed. Let \( f \) be a function mapping \( X \) into \( \mathbb{R} \) for which there exists \( M < \infty \) such that
\[
|f(x) - f(y)| \leq M d(x, y)
\]
for all \( x \) and \( y \) in \( X \). Such an \( f \) is called **Lipschitz continuous** with constant \( M \). The message of Theorem 1.2.3 is that if the Laplace limit
\[
\lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\} = - \inf_{x \in X} \{h(x) + I(x)\}
\]
is valid for all bounded continuous functions \( h \), then the sequence \( \{X^n\} \) satisfies the large deviation principle on \( X \) with rate function \( I \). This can easily be strengthened. Indeed the functions \( h_j \) defined in equation (1.3) in the proof of the large deviation upper bound
are bounded and Lipschitz continuous as is the function $h$ defined in equation (1.5) in the proof of the large deviation lower bound. Thus it suffices if the Laplace limit is valid merely for all bounded, Lipschitz continuous functions. As we point out in the next corollary, this can be strengthened even further by considering separately the Laplace principle upper and lower bounds.

**Corollary 1.2.5.** Let $I$ be a rate function on $\mathcal{X}$. If the Laplace limit

$$\lim_{n \to \infty} \frac{1}{n} \log E \{ \exp[-n h(X^n)] \} = - \inf_{x \in \mathcal{X}} \{ h(x) + I(x) \}$$

is valid for all bounded, Lipschitz continuous functions $h$ mapping $\mathcal{X}$ into $\mathbb{R}$, then the sequence $\{X^n\}$ satisfies the Laplace principle on $\mathcal{X}$ with rate function $I$ and the large deviation principle on $\mathcal{X}$ with rate function $I$. More precisely, the following implications hold.

(a) If the upper bound

$$\limsup_{n \to \infty} \frac{1}{n} \log E \{ \exp[-n h(X^n)] \} \leq - \inf_{x \in \mathcal{X}} \{ h(x) + I(x) \}$$

is valid for all bounded, Lipschitz continuous functions $h$, then both the large deviation upper bound and the Laplace principle upper bound are valid with rate function $I$.

(b) If the lower bound

$$\liminf_{n \to \infty} \frac{1}{n} \log E \{ \exp[-n h(X^n)] \} \geq - \inf_{x \in \mathcal{X}} \{ h(x) + I(x) \}$$

is valid for all bounded, Lipschitz continuous functions $h$, then both the large deviation lower bound and the Laplace principle lower bound are valid with rate function $I$.

**Proof.** As we have just pointed out, the large deviation bounds are valid simply because the functions $h_j$ and $h$ defined in equations (1.3) and (1.5) are bounded and Lipschitz continuous. The Laplace principle bounds are then obtained by applying Theorem 1.2.1. 

There is an analogy between Corollary 1.2.5 and an aspect of the theory of weak convergence of probability measures which is worth mentioning. A consequence of the corollary is that the limit

$$\lim_{n \to \infty} \frac{1}{n} \log E \{ \exp[-n h(X^n)] \} = - \inf_{x \in \mathcal{X}} \{ h(x) + I(x) \}$$

is valid for all bounded continuous function $h$ if and only if it is valid for all bounded, Lipschitz continuous functions $h$. This is an obvious analogy to the fact that if $\{\theta_n, n \in \mathbb{N}\}$ and $\theta$ are probability measures on $\mathcal{X}$, then

$$\lim_{n \to \infty} \int_{\mathcal{X}} g \, d\theta_n = \int_{\mathcal{X}} g \, d\theta$$
Lipschitz continuous functions $g$ and $y$ and only if the same limit holds for all bounded, Lipschitz continuous functions $g$ and $y$. Remark A.3).

In order to introduce the last topic of this section, consider the problem of proving a Laplace principle upper and lower bounds. It is elementary to show that together the uniform Laplace principle upper and lower bounds yield the uniform limit (1.6). The following proposition gives a useful criterion for showing these uniform bounds.

The term uniform Laplace principle upper bound refers to the validity of

$$\limsup_{y \in J} \left( \log \mathbb{E} \left( \exp \left(-n M(x)\right) \right) - F(y,h) \right) = 0.$$

where $F(y,h) = -\inf_{x \in \mathcal{X}} \{ h(x) + I_h(x) \}$.

The term uniform Laplace principle lower bound refers to the validity of

$$\liminf_{y \in J} \sup_{h \in H} \left( -\log \mathbb{E} \left( \exp \left(-n M(x)\right) \right) - F(y,h) \right) \geq 0.$$

Definition 1.2.6. Let $I$ be a family of rate functions on $\mathcal{X}$ parametrized by $y$ in a Polish space $\mathcal{X}$ and $X_n$ a sequence of stochastic processes. Typically the associated rate functions will depend on the parameter $y$. In such cases it is often useful to show that the Laplace principle holds uniform with respect to $y$. We denote by $E_y$ expectation with respect to $\{X^n\}$, a family of rate functions $I_y$ parameterized by $y \in J$.

A family of rate functions $I_y$ (parametrized by $y \in J$) is said to have compact level sets on compact if for all compact subsets $K$ of $\mathcal{X}$ and all bounded continuous functions $h$ mapping $X$ into $\mathbb{R}$.

$E_y \left( \exp \left(-n M(x)\right) \right) - F(y,h) \leq 0.$

Assume that the distributions of $\{X^n\}$ depend on a parameter $y$ which is a point in $\mathcal{X}$. In such cases it is often useful to show that the Laplace principle holds uniformly with respect to $y$. In this section we formulate the concept of a Laplace principle uniform with respect to a parameter $y$.

Typically the associated rate functions will depend on the parameter $y$. The examples to be considered in Chapters 6, 8, and 9 will illustrate the ease and naturality with which this issue of uniformity with respect to $y$ is handled by the Laplace principle.

The examples to be considered in Chapters 6, 8, and 9 will illustrate the ease and naturality with which this issue of uniformity with respect to $y$ is handled by the Laplace principle.

$E_y \left( \exp \left(-n M(x)\right) \right) - F(y,h) \leq 0.$

Assume that the distributions of $\{X^n\}$ depend on a parameter $y$ which is a point in $\mathcal{X}$. In such cases it is often useful to show that the Laplace principle holds uniformly with respect to $y$. In this section we formulate the concept of a Laplace principle uniform with respect to a parameter $y$.
1.2. EQUIVALENT FORMULATION

Proposition 1.2.7. Let $I_y$ be a family of rate functions on $\mathcal{X}$ parametrized by $y$ in a Polish space $\mathcal{Y}$ and assume that this family has compact level sets on compacts. Let $h$ be any bounded continuous function mapping $\mathcal{X}$ into $\mathbb{R}$. Assume that whenever $\{y_n, n \in \mathbb{N}\}$ is a sequence in $\mathcal{Y}$ converging to a point $y \in \mathcal{Y}$,

$$\lim_{n \to \infty} \frac{1}{n} \log E_{y_n} \{\exp[-n \cdot h(X^n)]\} = F(y, h),$$

(1.7)

where $F(y, h) \doteq - \inf_{x \in \mathcal{Y}} \{h(x) + I_y(x)\}$. Then the sequence $\{X^n\}$ satisfies the Laplace principle on $\mathcal{X}$ with rate function $I_y$ uniformly on compacts.

Proof. The key observation is that the function mapping $y \in \mathcal{Y} \mapsto F(y, h)$ is continuous. In order to show this, we define for $y \in \mathcal{Y}$

$$F^n(y, h) \doteq \frac{1}{n} \log E_y \{\exp[-n \cdot h(X^n)]\}$$

and fix $\bar{y} \in \mathcal{Y}$. We claim that for any $\delta > 0$ there exists $N \in \mathbb{N}$ such that for all $n \geq N$ and all $y \in \mathcal{Y}$ satisfying $d(y, \bar{y}) < 1/N$ the inequality $|F^n(y, h) - F(\bar{y}, h)| < \delta$ is valid. Sending $n \to \infty$ yields the continuity of $F(\cdot, h)$. We prove the claim by contradiction. Thus suppose that there exists $\delta > 0$ such that for all $N \in \mathbb{N}$ there exists $n_N \geq N$ and a point $y_N \in \mathcal{Y}$ satisfying $d(y_N, \bar{y}) < 1/N$ and $|F^n(y_N, h) - F(\bar{y}, h)| \geq \delta$. The limit (1.7) yields the inequality

$$0 = \lim_{N \to \infty} |F^n_N(y_N, h) - F(\bar{y}, h)| \geq \delta,$$

which is nonsense. The claim is proved.

Now let $\{y_n, n \in \mathbb{N}\}$ be any sequence in $\mathcal{Y}$ converging to a point $y \in \mathcal{Y}$. Then (1.7) implies that

$$\limsup_{n \to \infty} F^n(y_n, h) \leq F(y, h)$$

(1.8)

and

$$\liminf_{n \to \infty} F^n(y_n, h) \geq F(y, h).$$

(1.9)

We prove that (1.8) yields the uniform Laplace principle upper bound. We omit the similar proof that (1.9) yields the uniform Laplace principle lower bound.

If the uniform Laplace principle upper bound is not valid, then for some compact set $K$

$$\limsup_{n \to \infty} \sup_{y \in K} (F^n(y, h) - F(y, h)) > 0.$$

Thus there would exist a sequence $\{y_n, n \in \mathbb{N}\}$ in $K$ satisfying

$$\limsup_{n \to \infty} (F^n(y_n, h) - F(y_n, h)) > 0.$$

Without loss of generality we can assume that $y_n \to y$ for some point $y \in K$. Hence by the continuity of $F(\cdot, h)$ we would have

$$\limsup_{n \to \infty} (F^n(y_n, h) - F(y, h)) > 0.$$

Since this contradicts (1.8), the uniform Laplace principle upper bound is proved.

In the next section we explore the Laplace principle in somewhat more detail, presenting a number of results that are basic to the theory.
1.3 Basic Results in the Theory

The naturalness of formulating the large deviation principle in terms of a Laplace principle can be seen by the relative ease of proof of a number of basic results in the theory. Of the following results to be considered in this section, in the sequel we will apply only the contraction principle, stated second, and the result on superexponential approximation, stated third. The others may be omitted with no loss in continuity.

1. If a sequence of random variables satisfies the Laplace principle with some rate function, then the rate function is unique [Theorem 1.3.1].

2. The continuous image of a sequence of random variables satisfying the Laplace principle also satisfies the Laplace principle [Theorem 1.3.2].

3. The Laplace principle is preserved under superexponential approximation [Theorem 1.3.3].

4. If the Laplace principle is valid, then the Laplace limit holds for certain unbounded continuous functions [Theorem 1.3.4].

5. If the Laplace principle is valid, then the Laplace limit holds for certain lower semicontinuous functions satisfying a continuity condition [Theorem 1.3.6].

6. Any exponentially tight sequence of random variables has a subsequence that satisfies the Laplace principle, and thus the large deviation principle, with some rate function $I$ [Theorem 1.3.7].

7. Any exponentially tight sequence of random variables $\{X^n, n \in \mathbb{N}\}$ having the property that the limit $\lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-nh(X^n)]\}$ exists for every bounded continuous function $h$ satisfies the Laplace principle, and thus the large deviation principle, with some rate function $I$ [Theorem 1.3.8].

The first four of these results are standard. Generalizations of the fifth result can be found in [87]. The sixth result is due to O’Brien and Verwaat [76] and separately to Pukhalskii [78]; a generalization is given in [26]. Although we will not use it, this theorem nicely complements our weak convergence approach since it is a large deviation analogue of Prohorov’s Theorem [A.3.15], of which we will make numerous applications in the sequel. The last result, due to Bryc [16], has already been discussed just after the statement of Theorem 1.2.3.

We start by proving that a rate function in a Laplace principle must be unique.

**Theorem 1.3.1.** We assume that $\{X^n\}$ satisfies the Laplace principle on $\mathcal{X}$ with rate function $I$ and with rate function $J$. Then $I(\xi) = J(\xi)$ for all $\xi \in \mathcal{X}$.

**Proof.** For $j \in \mathbb{N}$ and any point $\xi \in \mathcal{X}$ we define the bounded continuous function $h_j(x) = j(d(x, \xi) \wedge 1)$. By Lemma 1.2.4

$$\lim_{j \to \infty} \lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-nh_j(X^n)]\} = \lim_{j \to \infty} \inf_{x \in \mathcal{X}} \{h_j(x) + I(x)\} = I(\xi)$$
1.3. BASIC RESULTS

and

$$\lim_{j \to \infty} \lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h_j(X^n)]\} = \lim_{j \to \infty} \inf_{x \in \mathcal{X}} \{h_j(x) + J(x)\} = J(\xi).$$

Thus $I(\xi) = J(\xi)$. ■

The next result is a standard tool in the theory.

**Theorem 1.3.2 (contraction principle).** Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, $I$ a rate function on $\mathcal{X}$, and $f$ a continuous function mapping $\mathcal{X}$ into $\mathcal{Y}$. The following conclusions hold.

(a) For each $y \in \mathcal{Y}$

$$J(y) \doteq \inf\{I(x) : x \in f^{-1}(y)\}$$

is a rate function on $\mathcal{Y}$.

(b) If $\{X^n\}$ satisfies the Laplace principle on $\mathcal{X}$ with rate function $I$, then $\{f(X^n)\}$ satisfies the Laplace principle on $\mathcal{Y}$ with rate function $J$.

**Proof.** (a) Given $M < \infty$ we define the level sets

$$L_J(M) \doteq \{y \in \mathcal{Y} : J(y) \leq M\} \quad \text{and} \quad L_I(M) \doteq \{x \in \mathcal{X} : I(x) \leq M\}.$$ 

The definition of $J$ implies that $L_J(M) \supset f(L_I(M))$. On the other hand, since $I$ is a rate function, for each $y \in f(\mathcal{X})$ the infimum in the definition of $J$ is attained at some $x$ in the closed set $f^{-1}(y)$. It follows that $L_J(M) \subset f(L_I(M))$, which when coupled with the opposite inclusion yields that $L_J(M) = f(L_I(M))$. Since $f$ is continuous and the level sets of $I$ are compact, this formula shows that the level sets of $J$ are compact. Since $J$ is obviously nonnegative, we have shown that $J$ is a rate function on $\mathcal{Y}$.

(b) For any bounded continuous function $h$ mapping $\mathcal{Y}$ into $\mathbb{R}$, the composition $h \circ f$ is a bounded continuous function mapping $\mathcal{X}$ into $\mathbb{R}$. Hence

$$\lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(f(X^n))]\} = -\inf_{x \in \mathcal{X}} \{h(f(x)) + I(x)\} = -\inf_{y \in \mathcal{Y}} \{h(y) + J(y)\}.$$ 

Since we have already checked that $J$ is a rate function on $\mathcal{Y}$, the proof of part (b) is complete. ■

In Chapter 10 we will have a sequence of random variables which satisfies the Laplace principle with some rate function $I$ and a second sequence of random variables which is superexponentially close to the first sequence. The next theorem will allow us to conclude that the second sequence also satisfies the Laplace principle with the same rate function $I$.

**Theorem 1.3.3.** For $n \in \mathbb{N}$ let $X^n$ and $Y^n$ be random variables that are defined on the same probability space $(\Omega, \mathcal{F}, P)$ and take values in $\mathcal{X}$. We assume that $\{X^n\}$ satisfies the Laplace principle on $\mathcal{X}$ with rate function $I$ and that $\{Y^n\}$ is superexponentially close to $\{X^n\}$ in the following sense: for each $\delta > 0$

$$\limsup_{n \to \infty} \frac{1}{n} \log P\{d(Y^n, X^n) > \delta\} = -\infty.$$ 

Then $\{Y^n\}$ satisfies the Laplace principle on $\mathcal{X}$ with the same rate function $I$. 

**Proof.** By Corollary 1.2.5 it suffices to verify the Laplace limit
\[
\lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \, h(Y^n)]\} = - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}
\]
for all bounded, Lipschitz continuous functions $h$ mapping $\mathcal{X}$ into $\mathcal{R}$. Let $h$ be any such function and let $M$ denote its Lipschitz constant. Then for any $\delta > 0$
\[
\limsup_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \, h(Y^n)]\}
\]
\[
= \limsup_{n \to \infty} \frac{1}{n} \log \left( E\left\{\mathbf{1}_{d(Y^n, X^n) \leq \delta} \, \exp[-n \, h(Y^n)]\right\} + E\left\{\mathbf{1}_{d(Y^n, X^n) > \delta} \, \exp[-n \, h(Y^n)]\right\} \right)
\]
\[
\leq \limsup_{n \to \infty} \frac{1}{n} \log \left( E\{\exp[-n \, h(X^n) + nM\delta]\} + e^{n\|h\|_{\infty}} P\{d(Y^n, X^n) > \delta\} \right)
\]
\[
= - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\} + M\delta.
\]
Sending $\delta \to 0$ gives the Laplace principle upper bound
\[
\limsup_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \, h(Y^n)]\} \leq - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}.
\]

Similarly, for any $\delta > 0$
\[
\liminf_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \, h(Y^n)]\}
\]
\[
\geq \liminf_{n \to \infty} \frac{1}{n} \log E\left\{\mathbf{1}_{d(Y^n, X^n) \leq \delta} \, \exp[-n \, h(Y^n)]\right\}
\]
\[
\geq \liminf_{n \to \infty} \frac{1}{n} \log E\left\{\mathbf{1}_{d(Y^n, X^n) \leq \delta} \, \exp[-n \, h(X^n) - nM\delta]\right\}
\]
\[
\geq \liminf_{n \to \infty} \frac{1}{n} \log \left( E\{\exp[n \, h(X^n) - nM\delta]\} - e^{n\|h\|_{\infty} - nM\delta} P\{d(Y^n, X^n) > \delta\} \right)
\]
\[
= - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\} - M\delta.
\]
Sending $\delta \to 0$ gives the Laplace principle lower bound
\[
\liminf_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \, h(Y^n)]\} \geq - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}.
\]

This completes the proof of the theorem. 

If the Laplace principle is valid, then a natural question is whether the Laplace limit can be evaluated for certain unbounded continuous functions mapping $\mathcal{X}$ into $\mathcal{R}$. We next point out a class of such functions for which this can be carried out. The following theorem is due to Varadhan [87].

**Theorem 1.3.4.** Assume that $\{X^n\}$ satisfies the Laplace principle on $\mathcal{X}$ with rate function $I$. Let $h$ be a continuous function mapping $\mathcal{X}$ into $\mathcal{R}$ for which
\[
\lim_{C \to \infty} \limsup_{n \to \infty} \frac{1}{n} \log E\left\{\mathbf{1}_{h \leq -C} (X^n) \, \exp[-n \, h(X^n)]\right\} = -\infty.
\]
Then
\[
\lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\} = - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\} \tag{1.11}
\]
and the limit is finite. In particular, if \( h \) is bounded below on the union of the supports of the \( \{X^n\} \), then condition (1.10) is satisfied and the limit (1.11) holds and is finite.

**Proof.** To streamline the proof, we introduce the notation
\[
\Lambda^n(A, \varphi) = E\{1_A(X^n) \exp[-n \varphi(X^n)]\}
\]
for \( A \) a Borel subset of \( \mathcal{X} \) and \( \varphi \) a measurable function mapping \( \mathcal{X} \) into \( \mathbb{R} \). Our goal is to prove that
\[
\lim_{n \to \infty} \frac{1}{n} \log \Lambda^n(\mathcal{X}, h) = - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}
\]
and that the limit is finite.

We first evaluate the limit inferior of \( \frac{1}{n} \log \Lambda^n(\mathcal{X}, h) \) as \( n \to \infty \). Given \( x \) an arbitrary point in \( \mathcal{X} \) and \( \varepsilon \) an arbitrary positive number, we define the open set \( G = \{y \in \mathcal{X} : h(y) < h(x) + \varepsilon\} \). Since the Laplace principle implies the large deviation principle with the same rate function \( I \) [Theorem 1.2.3], the large deviation lower bound yields
\[
\liminf_{n \to \infty} \frac{1}{n} \log \Lambda^n(\mathcal{X}, h) \geq \liminf_{n \to \infty} \frac{1}{n} \log \Lambda^n(G, h)
\]
\[
\geq -h(x) - \varepsilon + \liminf_{n \to \infty} \frac{1}{n} \log P\{X^n \in G\}
\]
\[
\geq -h(x) - \varepsilon - I(G)
\]
\[
\geq -h(x) - I(x) - \varepsilon.
\]
Since \( x \in \mathcal{X} \) and \( \varepsilon > 0 \) are arbitrary, it follows that
\[
\liminf_{n \to \infty} \frac{1}{n} \log \Lambda^n(\mathcal{X}, h) \geq - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}.
\]

Clearly \( \inf_{x \in \mathcal{X}} \{h(x) + I(x)\} < \infty \). Condition (1.10) guarantees that as \( n \to \infty \) the limit superior of \( \frac{1}{n} \log \Lambda^n(\mathcal{X}, h) \) is finite. Hence the last display implies that \( \inf_{x \in \mathcal{X}} \{h(x) + I(x)\} > -\infty \). We now prove that
\[
\limsup_{n \to \infty} \frac{1}{n} \log \Lambda^n(\mathcal{X}, h) \leq - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}.
\]
According to (1.10), there exists \( C \in (0, \infty) \) satisfying both \( C > \inf_{x \in \mathcal{X}} \{h(x) + I(x)\} \) and
\[
\limsup_{n \to \infty} \frac{1}{n} \log \Lambda^n(\{h \leq -C\}, h) \leq - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}.
\]
In terms of \( C \), we define the bounded continuous function
\[
h_C(x) = \begin{cases} h(x) & \text{if } -C \leq h(x) \leq C \\ C & \text{if } h(x) \geq C \\ -C & \text{if } h(x) \leq -C. \end{cases}
\]
By the choice of $C$ and the nonnegativity of $I$

$$\inf_{x \in \mathcal{X}} \{h_C(x) + I(x)\} \geq \inf_{x \in \mathcal{X}} \{h(x) + I(x)\},$$

and since the Laplace principle is valid,

$$\lim_{n \to \infty} \frac{1}{n} \log \Lambda^n(\mathcal{X}, h_C) = - \inf_{x \in \mathcal{X}} \{h_C(x) + I(x)\}.$$ 

Therefore

$$\limsup_{n \to \infty} \frac{1}{n} \log \Lambda^n(\mathcal{X}, h)$$

$$= \limsup_{n \to \infty} \frac{1}{n} \log \left[ \Lambda^n\left(\{-C \leq h \leq C\}, h\right) + \Lambda^n\left(\{h > C\}, h\right) + \Lambda^n\left(\{h < -C\}, h\right) \right]$$

$$\leq \limsup_{n \to \infty} \frac{1}{n} \log \left[ \Lambda^n(\mathcal{X}, h_C) + e^{-nC} + \Lambda^n(\mathcal{X}, h) \right]$$

$$\leq \left( - \inf_{x \in \mathcal{X}} \{h_C(x) + I(x)\} \right) \vee (-C) \vee \left( - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\} \right)$$

$$= - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}.$$ 

This is what we want to show. The proof of the theorem is complete. \(\blacksquare\)

In a number of applications Laplace-type expectations arise which involve discontinuous functions. Theorem 1.3.6 shows that the Laplace limit can be evaluated if the function is bounded and lower semicontinuous and satisfies a continuity condition. One encounters such functions, for example, in the study of the exit times of processes from smooth regions. Before stating the theorem, we need to know that a bounded, lower semicontinuous function can be suitably approximated by a sequence of uniformly bounded, Lipschitz continuous functions.

**Lemma 1.3.5.** Let $g$ be a bounded, lower semicontinuous function mapping $\mathcal{X}$ into $\mathbb{R}$. Then there exists a sequence $\{g_j, j \in \mathbb{N}\}$ of uniformly bounded, Lipschitz continuous functions mapping $\mathcal{X}$ into $\mathbb{R}$ with the properties that $g_j \uparrow g$ and that if $\{x_j, j \in \mathbb{N}\}$ is a sequence of points in $\mathcal{X}$ converging to some point $x^*$, then $\liminf_{j \to \infty} g_j(x_j) \geq g(x^*)$.

**Proof.** We follow the proofs of Lemmas 7.7 and 7.14 in [8]. For $j \in \mathbb{N}$ and $x \in \mathcal{X}$ we define

$$g_j(x) \doteq \inf_{y \in \mathcal{X}} \{g(y) + j \, d(x, y)\}.$$ 

Clearly $g_j \leq g_{j+1}$ and

$$\inf_{y \in \mathcal{X}} g(y) \leq g_j(x) \leq g(x) + j \, d(x, x) = g(x) \leq \sup_{y \in \mathcal{X}} g(y).$$

Thus the sequence $\{g_j, j \in \mathbb{N}\}$ is uniformly bounded and monotonically nondecreasing, and $\lim_{j \to \infty} g_j \leq g$. For any points $x$, $y$, and $z$ in $\mathcal{X}$

$$g(y) + j \, d(x, y) \leq g(y) + j \, d(z, y) + j \, d(x, z),$$

$$g(z) + j \, d(x, z) \leq g(z) + j \, d(x, y) + j \, d(x, z),$$

$$g(y) + j \, d(y, x) \leq g(z) + j \, d(x, y) + j \, d(x, z),$$

$$g(x) + j \, d(x, y) \leq g(z) + j \, d(x, y) + j \, d(x, z).$$

Therefore $\liminf_{j \to \infty} g_j(x) \geq g(x)$. Since $g_j \to g$ uniformly, the conclusion follows.

\(\blacksquare\)
1.3. BASIC RESULTS

which yields \( g_j(x) \leq g_j(z) + j \, d(x, z) \), and interchanging \( x \) and \( z \) gives

\[
|g_j(x) - g_j(z)| \leq j \, d(x, z).
\]

This inequality shows that \( g_j \) is Lipschitz continuous on \( \mathcal{X} \). Now let \( \{x_j, j \in \mathbb{N}\} \) be any sequence of points in \( \mathcal{X} \) converging to a point \( x^* \) and set \( A = \sup_{x \in \mathcal{X}} g(x) \). For \( j \in \mathbb{N} \) and \( \epsilon > 0 \) there exists \( y_j \in \mathcal{X} \) such that

\[
A + \epsilon \geq g_j(x_j) + \epsilon \geq g(y_j) + j \, d(x_j, y_j) \geq g(y_j).
\]

This inequality is violated unless \( d(x_j, y_j) \to 0 \) as \( j \to \infty \), and since \( x_j \to x^* \), it follows that \( y_j \to x^* \). The lower semicontinuity of \( g \) yields

\[
\liminf_{j \to \infty} g_j(x_j) + \epsilon \geq \liminf_{j \to \infty} g(y_j) \geq g(x^*).
\]

Since \( \epsilon \) is an arbitrary positive number, we have proved the claim that \( \liminf_{j \to \infty} g_j(x_j) \geq g(x^*) \). In particular, for each fixed \( x \in \mathcal{X} \) \( \liminf_{j \to \infty} g_j(x) \geq g(x) \). Since the sequence \( \{g_j\} \) is monotonically nondecreasing and \( \lim_{j \to \infty} g_j \leq g \), it follows that \( g_j \uparrow g \). This completes the proof. \( \blacksquare \)

We next prove that the Laplace limit is valid for a bounded, lower semicontinuous function satisfying a continuity condition. This result can easily be extended to an unbounded function satisfying condition (1.10) in Theorem 1.3.4. Generalizations for sequences of functions are given in Section 3 of [87].

**Theorem 1.3.6.** Assume that \( \{X^n\} \) satisfies the Laplace principle on \( \mathcal{X} \) with rate function \( I \). Let \( g \) be a bounded, lower semicontinuous function mapping \( \mathcal{X} \) into \( \mathbb{R} \). The following conclusions hold.

(a) The upper bound

\[
\limsup_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \, g(X^n)]\} \leq - \inf_{x \in \mathcal{X}} \{g(x) + I(x)\}
\]

is valid.

(b) Assume in addition that for each \( \epsilon > 0 \) there exists a point \( x_\epsilon \in \mathcal{X} \) such that \( g \) is continuous at \( x_\epsilon \) and

\[
g(x_\epsilon) + I(x_\epsilon) \leq \inf_{x \in \mathcal{X}} \{g(x) + I(x)\} + \epsilon.
\]

Then the Laplace limit

\[
\lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \, g(X^n)]\} = - \inf_{x \in \mathcal{X}} \{g(x) + I(x)\}
\]

is valid.
Proof. (a) Let \( \{g_j, j \in \mathbb{N}\} \) be the sequence of functions in Lemma 1.3.5. Since each function \( g_j \) is bounded and continuous and \( g_j \leq g \),
\[
\limsup_{n \to \infty} \frac{1}{n} \log E \{\exp[-n g(X^n)]\} \leq \limsup_{n \to \infty} \frac{1}{n} \log E \{\exp[-n g_j(X^n)]\} = - \inf_{x \in \mathcal{X}} \{g_j(x) + I(x)\}.
\]
In order to complete the proof we must show that
\[
\liminf_{j \to \infty} \inf_{x \in \mathcal{X}} \{g_j(x) + I(x)\} = \inf_{x \in \mathcal{X}} \{g(x) + I(x)\}.
\]
The sequence of infima \( \inf_{x \in \mathcal{X}} \{g_j(x) + I(x)\} \) is nondecreasing and
\[
\limsup_{j \to \infty} \inf_{x \in \mathcal{X}} \{g_j(x) + I(x)\} \leq \inf_{x \in \mathcal{X}} \{g(x) + I(x)\} < \infty.
\]
We now prove that
\[
\liminf_{j \to \infty} \inf_{x \in \mathcal{X}} \{g_j(x) + I(x)\} \geq \inf_{x \in \mathcal{X}} \{g(x) + I(x)\}.
\]
Since \( I \) is a rate function and each \( g_j \) is bounded and continuous, there exists a sequence \( \{x_j, j \in \mathbb{N}\} \) such that
\[
g_j(x_j) + I(x_j) = \inf_{x \in \mathcal{X}} \{g_j(x) + I(x)\} \leq \inf_{x \in \mathcal{X}} \{g(x) + I(x)\}.
\]
The uniform boundedness of the sequence \( \{g_j\} \) implies that \( \sup_{j \in \mathbb{N}} I(x_j) < \infty \), and since \( I \) has compact level sets there exists a subsequence of \( \{x_j\} \) converging to some \( x^* \in \mathcal{X} \). A property of \( \{g_j\} \) stated at the end of Lemma 1.3.5 gives the required lower limit
\[
\liminf_{j \to \infty} \inf_{x \in \mathcal{X}} \{g_j(x) + I(x)\} = \liminf_{j \to \infty} [g_j(x_j) + I(x_j)] \geq g(x^*) + I(x^*) \geq \inf_{x \in \mathcal{X}} \{g(x) + I(x)\}.
\]
This completes the proof of the upper bound.

(b) In order to prove part (b) we must show that
\[
\liminf_{n \to \infty} \frac{1}{n} \log E \{\exp[-n g(X^n)]\} \geq \inf_{x \in \mathcal{X}} \{g(x) + I(x)\}.
\]
By the hypothesis on \( g \), for each \( \varepsilon > 0 \) there exists \( x_\varepsilon \in \mathcal{X} \) and \( \delta > 0 \) such that
\[
g(x_\varepsilon) + I(x_\varepsilon) \leq \inf_{x \in \mathcal{X}} \{g(x) + I(x)\} + \varepsilon
\]
and \( |g(y) - g(x_\varepsilon)| < \varepsilon \) whenever \( y \) lies in the open ball \( B(x_\varepsilon, \delta) \ni \{y \in \mathcal{X} : d(y, x_\varepsilon) < \delta\} \). Since the Laplace principle implies the large deviation principle with the same rate function \( I \) [Theorem 1.2.3], the large deviation lower bound applied to \( B(x_\varepsilon, \delta) \) yields the lower limit
\[
\liminf_{n \to \infty} \frac{1}{n} \log E \{\exp[-n g(X^n)]\} \geq \liminf_{n \to \infty} \frac{1}{n} \log E \{1_{B(x_\varepsilon, \delta)}(X^n) \exp[-n g(X^n)]\}
\]
\[
\geq - g(x_\varepsilon) - \varepsilon + \liminf_{n \to \infty} \frac{1}{n} \log P\{X^n \in B(x_\varepsilon, \delta)\}
\]
\[
\geq - g(x_\varepsilon) - I(B(x_\varepsilon, \delta)) - \varepsilon
\]
\[
\geq - g(x_\varepsilon) - I(x_\varepsilon) - \varepsilon
\]
\[
\geq - \inf_{x \in \mathcal{X}} \{g(x) + I(x)\} - 2\varepsilon.
\]
1.3. BASIC RESULTS

Sending \( \varepsilon \to 0 \) completes the proof. ■

O’Brien and Verwaal [76] and separately Pukhalskii [78] proved that any exponentially tight sequence of random variables has a subsequence that satisfies a large deviation principle. We give a new proof based on the Laplace principle. As a bonus, our method of proof will also yield Bryc’s Theorem [16], which is stated at the end of this section. We recall that \( \{X^n\} \) is exponentially tight if for each \( M \in (0, \infty) \) there exists a compact set \( K \) such that
\[
\limsup_{n \to \infty} \frac{1}{n} \log P\{X^n \in K^c\} \leq -M.
\]

**Theorem 1.3.7.** If the sequence \( \{X^n\} \) is exponentially tight, then there exists a subsequence satisfying the Laplace principle on \( \mathcal{X} \), and thus the large deviation principle on \( \mathcal{X} \), with some rate function \( I \).

**Proof.** We start by appealing to a standard separability result in analysis which is given in Theorem A.6.1. This result states there exists an equivalent metric \( m(x, y) \) on \( \mathcal{X} \) with the property that the space \( \mathcal{U}_b(\mathcal{X}, m) \) of bounded, uniformly continuous functions mapping \( \mathcal{X} \) into \( \mathbb{R} \) is separable with respect to the uniform metric. Let \( \Xi \) be a countable dense subset of \( \mathcal{U}_b(\mathcal{X}, m) \). For each \( f \in \Xi \) the sequence of numbers \( \left\{ \frac{1}{n} \log E\{\exp[-n f(X^n)]\}, n \in \mathbb{N} \right\} \) is bounded. Hence a diagonalization argument guarantees the existence of a subsequence of \( n \in \mathbb{N} \) for which the limit
\[
\Lambda(f) \doteq \lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n f(X^n)]\}
\]
exists for each \( f \in \Xi \). Given any \( h \in \mathcal{U}_b(\mathcal{X}, m) \), for each \( k \in \mathbb{N} \) there exists \( f_k \in \Xi \) satisfying \( \|h - f_k\|_\infty \leq 1/k \). It follows that
\[
\left| \frac{1}{n} \log E\{\exp[-n h(X^n)]\} - \frac{1}{n} \log E\{\exp[-n f_k(X^n)]\} \right| \leq \|h - f_k\|_\infty \leq \frac{1}{k},
\]
and thus that
\[
\Lambda(f_k) - \frac{1}{k} \leq \liminf_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\} \leq \limsup_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\} \leq \Lambda(f_k) + \frac{1}{k}.
\]
This in turn implies that
\[
\limsup_{k \to \infty} \Lambda(f_k) \leq \liminf_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\} \leq \limsup_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\} \leq \liminf_{k \to \infty} \Lambda(f_k).
\]
We conclude that both of the limits \( \lim_{k \to \infty} \Lambda(f_k) \) and
\[
\Lambda(h) \doteq \lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\}
\]
exist and are finite and that
\[ \Lambda(h) = \lim_{k \to \infty} \Lambda(f_k). \]
In this way we extend the definition of \( \Lambda \) from a functional on \( \Xi \) to a functional on all of \( U'_b(\mathcal{X}, m) \). It is easy to check that the definition of \( \Lambda(h) \) is independent of the approximating sequence \( \{f_k, k \in \mathbb{N}\} \) and thus that \( \Lambda(h) \) is well defined.

For \( x \in \mathcal{X} \) we now define
\[
I(x) = -\inf_{h \in U'_b(\mathcal{X}, m)} \{h(x) + \Lambda(h)\}. \tag{1.13}
\]
Since for any \( h \in U'_b(\mathcal{X}, m) \) and any real number \( c \) \( \Lambda(h - c) = \Lambda(h) + c \), \( I(x) \) may be written in the alternate form
\[
I(x) = -\inf \{\Lambda(\varphi) : \varphi \in U'_b(\mathcal{X}, m), \varphi(x) = 0\}.
\]
We claim that \( I \) is a rate function and that along the subsequence of \( n \in \mathbb{N} \) for which the limit in equation (1.12) exists, \( \{X^n\} \) satisfies the Laplace principle with rate function \( I \). Theorem 1.2.3 then implies that along the same subsequence of \( n \in \mathbb{N} \) \( \{X^n\} \) satisfies the large deviation principle with rate function \( I \).

\( I \) maps \( \mathcal{X} \) into \([0, \infty]\) since for each \( x \in \mathcal{X} \) \( I(x) \geq -\Lambda(0) = 0 \). We now prove that \( I \) has compact level sets. Given \( M \in (0, \infty) \), the exponential tightness of \( \{X^n\} \) allows us to choose a compact set \( K \) such that
\[
\limsup_{n \to \infty} \frac{1}{n} \log P\{X^n \in K^c\} \leq -2M.
\]
Fix a point \( x \in K^c \) and define a continuous function \( \psi \) on \([0, \infty)\) by \( \psi(t) = 1 - t \) for \( t \in [0, 1] \) and \( \psi(t) = 0 \) for \( t \geq 1 \). Setting \( \delta = m(x, K)/2 > 0 \), we then define for \( y \in \mathcal{X} \)
\[
\varphi(y) \doteq 2M \psi(m(y, K)/\delta).
\]
This function is uniformly continuous and satisfies \( \varphi(x) = 0, \varphi(y) = 2M \) for all \( y \in K \), and \( 0 \leq \varphi(y) \leq 2M \) for all \( y \in \mathcal{X} \). Therefore
\[
\Lambda(\varphi) \doteq \lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n \varphi(X^n)]\}
\]
\[= \lim_{n \to \infty} \frac{1}{n} \log \left( E\{1_K(X^n) \exp[-n \varphi(X^n)]\} + E\{1_{K^c}(X^n) \exp[-n \varphi(X^n)]\}\right)\]
\[\leq \limsup_{n \to \infty} \frac{1}{n} \log \left( e^{-2mM} + P\{X^n \in K^c\}\right)\]
\[= -2M,
\]
and thus
\[
I(x) \doteq -\inf \{\Lambda(\varphi) : \varphi \in U'_b(\mathcal{X}, m), \varphi(x) = 0\} \geq 2M.
\]
Hence for any \( x \in K^c \) \( I(x) \geq 2M \). This shows that the level set \( \{y \in \mathcal{X} : I(y) \leq M\} \) is a subset of the compact set \( K \).
In order to complete the proof that \( I \) has compact level sets, we must show that \( I \) is
lower semicontinuous. Take any \( x \in \mathcal{X} \) for which \( I(x) < \infty \). Given \( \varepsilon > 0 \) there exists
\( \varphi \in \mathcal{U}_b(\mathcal{X}, m) \) satisfying \( \varphi(x) = 0 \) and \( I(x) \leq -\Lambda(\varphi) + \varepsilon \). Since \( \varphi \) is continuous, \( \varphi(y) \leq \varepsilon \)
for all \( y \in \mathcal{X} \) that are sufficiently close to \( x \). For such points \( y \)
\[
I(y) \geq -\Lambda(\varphi - \varphi(y)) = -\Lambda(\varphi) - \varphi(y) \geq -\Lambda(\varphi) - \varepsilon \geq I(x) - 2\varepsilon.
\]
Since \( \varepsilon > 0 \) is arbitrary, it follows that
\[
\liminf_{y \to x} I(y) \geq I(x).
\]
The proof of the lower semicontinuity of \( I \) at points \( x \in \mathcal{X} \) for which \( I(x) = \infty \) is similar,
and we omit it. This completes the proof that \( I \) has compact level sets. We conclude that
the function \( I \) defined in equation (1.13) is a rate function.

All that remains to prove is that for every \( h \in \mathcal{U}_b(\mathcal{X}, m) \) the Laplace limit \( \Lambda(h) \) is
given by
\[
\Lambda(h) = - \inf_{x \in \mathcal{X}} \{ h(x) + I(x) \}.
\]
Since every bounded, Lipschitz continuous function mapping \( \mathcal{X} \) into \( \mathbb{R} \) is an element of
\( \mathcal{U}_b(\mathcal{X}, m) \), it will then follow from Corollary 1.2.5 that along the subsequence of \( n \in \mathbb{N} \)
for which the limit in equation (1.12) exists, \( \{ X^n \} \) satisfies the Laplace principle with
rate function \( I \) defined in formula (1.13).

By the definition of \( I \), for each \( x \in \mathcal{X} \) and \( h \in \mathcal{U}_b(\mathcal{X}, m) \) \( I(x) \geq -h(x) - \Lambda(h) \), which
implies that
\[
\Lambda(h) \geq - \inf_{x \in \mathcal{X}} \{ h(x) + I(x) \}.
\]
We now prove that the reverse inequality. By replacing \( h \) with the nonnegative function
\( h - \inf_{y \in \mathcal{X}} h(y) \), without loss of generality it can be assumed that \( h \) is nonnegative. Given
\( M \in (0, \infty) \), the exponential tightness of \( \{ X^n \} \) allows us to choose a compact set \( K \) such that
\[
\limsup_{n \to \infty} \frac{1}{n} \log P\{ X^n \in K^c \} \leq -M.
\]
We temporarily fix \( \varepsilon > 0 \). For each \( x \in K \) for which \( I(x) < \infty \) there exists \( \varphi_x \in \mathcal{U}_b(\mathcal{X}, m) \)
satisfying \( \varphi_x(x) = 0 \) and
\[
\Lambda(\varphi_x) \leq -I(x) + \varepsilon.
\]
For each \( x \in K \) for which \( I(x) = \infty \) there \( \varphi_x \in \mathcal{U}_b(\mathcal{X}, m) \) satisfying \( \varphi_x(x) = 0 \) and
\[
\Lambda(\varphi_x) \leq -M - \| h \|_\infty - 2\varepsilon.
\]
For each \( x \in K \) we now choose \( \delta(x) > 0 \) such that \( \varphi_x(y) \leq \varepsilon \) and \( h(y) \geq h(x) - \varepsilon \)
whenever \( y \) lies in the open ball \( B(x, \delta(x)) \). Since the open balls \( \{ B(x, \delta(x)), x \in K \} \)
are an open cover of the compact set \( K \), there exists an integer \( r \) and points \( x_1, x_2, \ldots, x_r \) in
\( K \) such that \( K \subset \bigcup_{i=1}^r B(x_i, \delta(x_i)) \). Hence if \( y \in K \), then \( y \) lies in some \( B(x_i, \delta(x_i)) \), and
so
\[
h(y) \geq h(x_i) - \varepsilon \geq \varphi_{x_i}(y) + h(x_i) - 2\varepsilon.
\]
It follows that for all \( y \in K \)
\[
h(y) \geq \bigwedge_{i=1}^{r} (\varphi_{x_i}(y) + h(x_i)) - 2\varepsilon.
\]
Therefore
\[
\Lambda(h) = \lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\}
\]
\[
= \lim_{n \to \infty} \frac{1}{n} \log \left( E\{1_K(X^n) \exp[-n h(X^n)]\} + E\{1_{K^c}(X^n) \exp[-n h(X^n)]\} \right)
\]
\[
\leq \limsup_{n \to \infty} \frac{1}{n} \log \left( E\left[\exp\left[-n \bigwedge_{i=1}^{r} (\varphi_{x_i}(X^n) + h(x_i) - 2\varepsilon)\right]\right] + P\{X^n \in K^c\} \right)
\]
\[
\leq \limsup_{n \to \infty} \frac{1}{n} \log \left( \sum_{i=1}^{r} E\{\exp[-n (\varphi_{x_i}(X^n) + h(x_i) - 2\varepsilon)]\} + P\{X^n \in K^c\} \right)
\]
\[
\leq \max_{i=1,2,\ldots,r} \{\Lambda(\varphi_{x_i}) - h(x_i) + 2\varepsilon\} \vee (-M)
\]
\[
\leq \max_{i=1,2,\ldots,r} \{-I(x_i) - h(x_i) + 3\varepsilon\} \vee (-\varepsilon)
\]
\[
\leq \sup_{x \in \mathcal{X}} \{-I(x) - h(x) + 3\varepsilon\} \vee (-M).
\]
Sending \( M \to \infty \) and \( \varepsilon \to 0 \) gives the required upper bound
\[
\Lambda(h) \leq \sup_{x \in \mathcal{X}} \{-I(x) - h(x)\} = -\inf_{x \in \mathcal{X}} \{h(x) + I(x)\}.
\]
This completes the proof of the theorem. \( \blacksquare \)

We end this section with Bryc’s Theorem.

**Theorem 1.3.8 (Bryc).** Assume that \( \{X^n\} \) is exponentially tight and that for all functions \( h \in \mathcal{C}_b(\mathcal{X}) \) the limit
\[
\Lambda(h) = \lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(X^n)]\}
\]
exists. Then
\[
I(x) = -\inf_{h \in \mathcal{C}_b(\mathcal{X})} \{h(x) + \Lambda(h)\}
\]
(1.14)
is a rate function on \( \mathcal{X} \), and for all \( h \in \mathcal{C}_b(\mathcal{X}) \) the Laplace limit \( \Lambda(h) \) has the form
\[
\Lambda(h) = -\inf_{x \in \mathcal{X}} \{h(x) + I(x)\}.
\]
It follows that \( \{X^n\} \) satisfies the Laplace principle on \( \mathcal{X} \), and thus the large deviation principle on \( \mathcal{X} \), with rate function \( I \).
1.3. BASIC RESULTS

Proof. Much of the work needed to prove this theorem has been done in the proof of the previous theorem. As in that proof, we endow \( \mathcal{X} \) with an equivalent metric \( m(x, y) \) such that \( \mathcal{U}_b(\mathcal{X}, m) \) is separable with respect to the uniform metric. Then

\[
J(x) \doteq - \inf_{h \in \mathcal{U}_b(\mathcal{X}, m)} \{ h(x) + \Lambda(h) \}
\]

is a rate function on \( \mathcal{X} \) and for all \( h \in \mathcal{U}_b(\mathcal{X}, m) \)

\[
\Lambda(h) = - \inf_{x \in \mathcal{X}} \{ h(x) + J(x) \}.
\]

Since \( \mathcal{U}_b(\mathcal{X}, m) \) contains all bounded, Lipschitz continuous functions, Corollary 1.2.5 implies that \( \{ X^n \} \) satisfies the Laplace principle and the large deviation principle on \( \mathcal{X} \) with rate function \( J \). The first of these means that for all \( h \in \mathcal{C}_b(\mathcal{X}) \)

\[
\Lambda(h) = - \inf_{x \in \mathcal{X}} \{ h(x) + J(x) \}.
\]

Hence the proof of Bryc’s Theorem is complete once we prove that for each \( x \in \mathcal{X} \)

\[
J(x) \doteq - \inf_{h \in \mathcal{U}_b(\mathcal{X}, m)} \{ h(x) + \Lambda(h) \} \text{ equals } I(x) \doteq - \inf_{h \in \mathcal{C}_b(\mathcal{X})} \{ h(x) + \Lambda(h) \}.
\]

Fix \( x \in \mathcal{X} \). Since \( \mathcal{U}_b(\mathcal{X}, m) \subset \mathcal{C}_b(\mathcal{X}) \),

\[
\inf_{h \in \mathcal{U}_b(\mathcal{X}, m)} \{ h(x) + \Lambda(h) \} \geq \inf_{h \in \mathcal{C}_b(\mathcal{X})} \{ h(x) + \Lambda(h) \}.
\]

To prove the reverse inequality, let \( \varepsilon > 0 \) be given and choose \( \bar{h} \in \mathcal{C}_b(\mathcal{X}) \) such that

\[
\inf_{h \in \mathcal{C}_b(\mathcal{X})} \{ h(x) + \Lambda(h) \} + \varepsilon \geq \bar{h}(x) + \Lambda(\bar{h}).
\]

By Lemma 1.3.5 applied to \( g \doteq -\bar{h} \), there exists a sequence of uniformly bounded, Lipschitz continuous functions \( \{ h_j, j \in \mathbb{N} \} \) satisfying \( h_j \downarrow \bar{h} \). For each \( j \), since \( h_j \geq \bar{h} \),

\[
\frac{1}{n} \log E \{ \exp[-n \bar{h}(X^n)] \} \geq \frac{1}{n} \log E \{ \exp[-n h_j(X^n)] \}
\]

and thus \( \Lambda(\bar{h}) \geq \Lambda(h_j) \). If \( j \) is chosen so large that \( \bar{h}(x) \geq h_j(x) - \varepsilon \), then

\[
\inf_{h \in \mathcal{C}_b(\mathcal{X})} \{ h(x) + \Lambda(h) \} + \varepsilon \geq \bar{h}(x) + \Lambda(\bar{h}) \geq h_j(x) + \Lambda(h_j) - \varepsilon \geq \inf_{h \in \mathcal{U}_b(\mathcal{X}, m)} \{ h(x) + \Lambda(h) \} - \varepsilon.
\]

Sending \( \varepsilon \to 0 \) completes the proof. ■

In the next section we introduce the relative entropy and show how to represent, via a variational formula involving the relative entropy, the logarithms of the expectations appearing in the Laplace limit

\[
\lim_{n \to \infty} \frac{1}{n} \log E \{ \exp[-n h(X^n)] \} = - \inf_{x \in \mathcal{X}} \{ h(x) + I(x) \}.
\]

This will be a key step in representing these quantities in terms of the minimum cost functions of associated stochastic control problems. The representation in terms of minimal cost functions will be the basis for analyzing the asymptotic behavior of the expectations and thus through Theorem 1.2.3 will be the basis for proving the Laplace principle.
1.4 Properties of the Relative Entropy

The relative entropy will play a key role in the definition of every rate function that we will consider. It arises in the weak convergence approach to large deviations via the variational formula given in part (a) of Proposition 1.4.2. The derivation of this formula requires only a measurable space \((\mathcal{V}, \mathcal{A})\). We denote by \(\mathcal{P}(\mathcal{V})\) the set of probability measures on \((\mathcal{V}, \mathcal{A})\).

For \(\theta \in \mathcal{P}(\mathcal{V})\), the relative entropy \(R(\cdot \| \theta)\) is a mapping from \(\mathcal{P}(\mathcal{V})\) into the extended real numbers. It is defined by

\[
R(\gamma \| \theta) \equiv \int_{\mathcal{V}} \left( \log \frac{d\gamma}{d\theta} \right) d\gamma
\]

whenever \(\gamma \in \mathcal{P}(\mathcal{V})\) is absolutely continuous with respect to \(\theta\). Otherwise, we set \(R(\gamma \| \theta) = \infty\). Occasionally we will refer to \(R(\gamma \| \theta)\) as the relative entropy of \(\gamma\) with respect to \(\theta\). For \(t \in \mathbb{R}\) define \(t^- \equiv -t \wedge 0\). Since \(s(\log s)^-\) is bounded for \(s \in [0, \infty)\), whenever \(\gamma \in \mathcal{P}(\mathcal{V})\) is absolutely continuous with respect to \(\theta\)

\[
\int_{\mathcal{V}} \left( \log \frac{d\gamma}{d\theta} \right)^- d\gamma = \int_{\mathcal{V}} \frac{d\gamma}{d\theta} \left( \log \frac{d\gamma}{d\theta} \right)^- d\theta < \infty.
\]

It follows that \(\int_{\mathcal{V}} \left( \log \frac{d\gamma}{d\theta} \right) d\gamma\) is well defined and that

\[
R(\gamma \| \theta) = \int_{\mathcal{V}} \frac{d\gamma}{d\theta} \left( \log \frac{d\gamma}{d\theta} \right) d\theta.
\]

The proof of the variational formula in Proposition 1.4.2 requires the two properties of relative entropy given in the next lemma.

**Lemma 1.4.1.** Let \((\mathcal{V}, \mathcal{A})\) be a measurable space and \(\gamma\) and \(\theta\) probability measures on \(\mathcal{V}\). Then \(R(\gamma \| \theta) \geq 0\) and \(R(\gamma \| \theta) = 0\) if and only if \(\gamma = \theta\).

**Proof.** In order to prove the nonnegativity, it suffices to consider the case where \(R(\gamma \| \theta) < \infty\). Since \(s \log s \geq s - 1\) with equality if and only if \(s = 1\),

\[
R(\gamma \| \theta) = \int_{\mathcal{V}} \frac{d\gamma}{d\theta} \left( \log \frac{d\gamma}{d\theta} \right) d\theta \geq \int_{\mathcal{V}} \left( \frac{d\gamma}{d\theta} - 1 \right) d\theta = 0.
\]

In this display equality holds if and only if \(d\gamma/d\theta = 1\) \(\theta\)-a.e.; i.e., if and only if \(\gamma = \theta\). This completes the proof. \(\blacksquare\)

Part (a) of the next proposition states the variational formula, and part (b) indicates where the infimum in the variational formula is attained. The proposition, though completely elementary, is a cornerstone of the weak convergence approach and will be applied on numerous occasions throughout the book. Its first applications will be in Chapters 2 and 3, where we will prove the Laplace principle in two basic examples: the empirical measures of i.i.d. random variables (Sanov’s Theorem) and certain continuous-time stochastic processes constructed from i.i.d. random variables (Mogulskii’s Theorem).
1.4. PROPERTIES OF THE RELATIVE ENTROPY

Proposition 1.4.2. Let $(\mathcal{V}, \mathcal{A})$ be a measurable space, $k$ a bounded measurable function mapping $\mathcal{V}$ into $\mathbb{R}$, and $\theta$ a probability measure on $\mathcal{V}$. The following conclusions hold.

(a) We have the variational formula

$$- \log \int_\mathcal{V} e^{-k} d\theta = \inf_{\gamma \in \mathcal{P}(\mathcal{V})} \left\{ R(\gamma \| \theta) + \int_\mathcal{V} k d\gamma \right\}. \quad (1.15)$$

(b) Let $\gamma_0$ denote the probability measure on $\mathcal{V}$ which is absolutely continuous with respect to $\theta$ and satisfies

$$\frac{d\gamma_0}{d\theta}(x) = e^{-k(x)} \cdot \frac{1}{\int_\mathcal{V} e^{-k} d\theta}.$$

Then the infimum in the variational formula (1.15) is uniquely attained at $\gamma_0$.

Proof. For part (a) it suffices to prove that

$$- \log \int_\mathcal{V} e^{-k} d\theta = \inf \left\{ R(\gamma \| \theta) + \int_\mathcal{V} k d\gamma : \gamma \in \mathcal{P}(\mathcal{V}), R(\gamma \| \theta) < \infty \right\}.$$

If $R(\gamma \| \theta) < \infty$, then $\gamma$ is absolutely continuous with respect to $\theta$, and since $\theta$ is absolutely continuous with respect to $\gamma_0$, $\gamma$ is also absolutely continuous with respect to $\gamma_0$. Thus

$$R(\gamma \| \theta) + \int_\mathcal{V} k d\gamma$$

$$= \int_\mathcal{V} \left( \log \frac{d\gamma}{d\theta} \right) d\gamma + \int_\mathcal{V} k d\gamma$$

$$= \int_\mathcal{V} \left( \log \frac{d\gamma}{d\gamma_0} \right) d\gamma + \int_\mathcal{V} \left( \log \frac{d\gamma_0}{d\theta} \right) d\gamma + \int_\mathcal{V} k d\gamma$$

$$= R(\gamma \| \gamma_0) - \log \int_\mathcal{V} e^{-k} d\theta.$$

We now use the facts that $R(\gamma \| \gamma_0) \geq 0$ and $R(\gamma \| \gamma_0) = 0$ if and only if $\gamma = \gamma_0$. These not only complete the proof of the variational formula in part (a) but also show that the infimum in the variational formula is uniquely attained at $\gamma_0$, as claimed in part (b). ■

We spend the rest of this section investigating six extremely pleasant properties of relative entropy: convexity, lower semicontinuity, compactness of level sets, uniform integrability of sequences of measures satisfying a suitable relative entropy bound, factorization, and approximation by sums. These properties will be used repeatedly throughout the book. Several of the proofs rely on the Donsker–Varadhan variational formula for the relative entropy [32], which is stated in part (a) of the next lemma. This formula is dual to the variational formula (1.15). Relative entropy has been the subject of deep study by numerous authors in probability, statistics, information theory, and other areas. In the present section we develop only those properties of the relative entropy that we need.

Let $\mathcal{X}$ be a Polish space and $\{\theta_n, n \in \mathbb{N}\}$ a sequence in $\mathcal{P}(\mathcal{X})$. We say that $\{\theta_n\}$ converges weakly to a probability measure $\theta$ on $\mathcal{X}$, and write $\theta_n \Rightarrow \theta$, if for each bounded continuous function $g$ mapping $\mathcal{X}$ into $\mathbb{R}$

$$\lim_{n \to \infty} \int_{\mathcal{X}} g d\theta_n = \int_{\mathcal{X}} g d\theta.$$
CHAPTER 1. THE LAPLACE PRINCIPLE

We endow $\mathcal{P}(\mathcal{X})$ with the weak topology, which is defined to be the topology corresponding to the weak convergence of probability measures. As we discuss in Section A.3, there exists a metric on $\mathcal{P}(\mathcal{X})$ which is compatible with the weak topology, and with respect to this metric $\mathcal{P}(\mathcal{X})$ is also a Polish space.

The proof that relative entropy has compact level sets relies on Prohorov’s Theorem, a key result in weak convergence theory which we now state. A family $\Phi$ of probability measures on $\mathcal{X}$ is said to be **tight** if for each $\varepsilon > 0$ there exists a compact set $K$ such that

$$\inf_{\gamma \in \Phi} \gamma(K) \geq 1 - \varepsilon.$$  

According to Prohorov’s Theorem [Theorem A.3.15], $\Phi$ is tight if and only if it is relatively compact with respect to weak convergence.

Part (f) of the lemma states a factorization property of the relative entropy. In order to formulate it, we introduce a fundamental concept. Let $(\mathcal{Y}, \mathcal{A})$ be a measurable space and $\mathcal{Y}$ a Polish space and let $\tau(dy|x)$ be a family of probability measures on $\mathcal{Y}$ parametrized by $x \in \mathcal{V}$. We call $\tau(dy|x)$ a **stochastic kernel** on $\mathcal{Y}$ given $\mathcal{V}$ if for every Borel subset $E$ of $\mathcal{Y}$ the function mapping $x \in \mathcal{V} \mapsto \tau(E|x) \in [0, 1]$ is measurable. A basic result is that a family $\tau(dy|x)$ of probability measures on $\mathcal{Y}$ parametrized by $x \in \mathcal{Y}$ is a stochastic kernel if and only if the function mapping $x \in \mathcal{V} \mapsto \tau(x|\cdot) \in \mathcal{P}(\mathcal{Y})$ is measurable [Theorem A.5.2]; i.e., if and only if $\tau(x|\cdot)$ is a random variable mapping $\mathcal{V}$ into $\mathcal{P}(\mathcal{Y})$. We have the following useful fact. If $\tau(dy|x)$ is a stochastic kernel on $\mathcal{Y}$ given $\mathcal{V}$ and $f$ is a measurable function mapping a measurable space $(\mathcal{Z}, \mathcal{D})$ into $\mathcal{V}$, then $\sigma(dy|z) = \tau(dy|f(z))$ is a stochastic kernel on $\mathcal{Y}$ given $\mathcal{Z}$. We adopt the convention that if $\mathcal{V}$ is empty, then a stochastic kernel on $\mathcal{Y}$ given $\mathcal{V}$ is a probability measure on $\mathcal{Y}$.

Stochastic kernels are commonly encountered in probability. Indeed, let $X$ and $Y$ be random variables taking values in $\mathcal{V}$ and $\mathcal{Y}$, respectively. Then a regular conditional distribution for $Y$ given $X = x$ defines a stochastic kernel on $\mathcal{Y}$ given $\mathcal{V}$. In Section A.5 regular conditional distributions will be used to derive two key decompositions of stochastic kernels. A basic example of a stochastic kernel on $\mathcal{Y}$ given $\mathcal{V}$ is a transition probability function $\tau(x, dy)$ of a Markov chain taking values in $\mathcal{Y}$. In this example or in cases when we want to suggest an interpretation of a particular stochastic kernel as a transition probability function, we will deviate from the notation $\tau(dy|x)$. In this context, if $X$ is a random variable mapping a probability space $(\Omega, \mathcal{F}, P)$ into $\mathcal{V}$, then $\tau(X(\omega), dy)$ is a stochastic kernel on $\mathcal{Y}$ given $\Omega$. Similar examples will arise in the book. Since in general probability spaces are not Polish spaces, this example motivates our general definition of a stochastic kernel.

Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, $\sigma(dy|x)$ and $\tau(dy|x)$ stochastic kernels on $\mathcal{Y}$ given $\mathcal{X}$, and $\theta$ a probability measure on $\mathcal{X}$. We define a probability measure $\theta \otimes \sigma$ on $\mathcal{X} \times \mathcal{Y}$ by

$$\theta \otimes \sigma(A \times B) \triangleq \int_{A \times B} \theta(dx) \sigma(dy|x) = \int_A \sigma(B|x) \theta(dx)$$

for Borel subsets $A$ of $\mathcal{X}$ and $B$ of $\mathcal{Y}$. This formula is summarized by the notation $\theta \otimes \sigma(dx \times dy) = \theta(dx) \otimes \sigma(dy|x)$. If, for example, $\sigma(dy|x)$ is independent of $x$, thus defining a probability measure $\sigma(dy)$ on $\mathcal{Y}$, then $\theta \otimes \sigma$ equals the product measure $\theta \times \sigma$.
1.4. PROPERTIES OF THE RELATIVE ENTROPY

We also define by the analogous formula the probability measure \( \theta \otimes \tau \) on \( \mathcal{X} \times \mathcal{Y} \). Part (f) of the next lemma states that the \( \theta(dx) \)-integral of \( R(\sigma \cdot |x|) \) equals the relative entropy \( R(\theta \otimes \sigma \cdot \theta \otimes \tau) \). A generalization is given in Theorem C.3.1, which is well known in information theory as the chain rule.

A \textbf{finite measurable partition} of \( \mathcal{X} \) is a finite sequence \( \pi \equiv \{ A_i, i = 1, 2, \ldots, r \} \) consisting of disjoint Borel sets whose union is \( \mathcal{X} \). Part (g) of the next lemma gives an approximation property of the relative entropy in terms of sums over finite measurable partitions.

All of the results in the lemma are formulated for arbitrary Polish spaces except parts (d) and (e), which are stated for \( \mathbb{R}^d \). A number of the results are valid for more general spaces, but for ease of exposition we will not point out which ones. Parts (b)-(e) of the lemma are proved in the present section since similar techniques will be used throughout the book. The compactness property given in part (c) is generalized in Proposition 9.3.6. Parts (a), (f), and (g) are proved in Appendix C.

\textbf{Lemma 1.4.3.} Let \( \mathcal{X} \) and \( \mathcal{Y} \) be Polish spaces. The relative entropy \( R(\cdot|\cdot) \) has the following properties.

(a) (Donsker–Varadhan variational formula.) We denote by \( \mathcal{C}_b(\mathcal{X}) \) the space of bounded continuous functions mapping \( \mathcal{X} \) into \( \mathbb{R} \) and by \( \Psi_\theta(\mathcal{X}) \) the space of bounded \( \mathcal{B}_0 \)-measurable functions mapping \( \mathcal{X} \) into \( \mathbb{R} \). Then for each \( \gamma \) and \( \theta \) in \( \mathcal{P}(\mathcal{X}) \)

\[
R(\gamma||\theta) = \sup_{g \in \mathcal{C}_b(\mathcal{X})} \left\{ \int_{\mathcal{X}} g \, d\gamma - \log \int_{\mathcal{X}} e^g \, d\theta \right\} = \sup_{\psi \in \Psi_\theta(\mathcal{X})} \left\{ \int_{\mathcal{X}} \psi \, d\gamma - \log \int_{\mathcal{X}} e^\psi \, d\theta \right\}.
\]

(b) \( R(\gamma||\theta) \) is a convex, lower semicontinuous function of \( (\gamma, \theta) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \). In particular, \( R(\gamma||\theta) \) is a convex, lower semicontinuous function of each variable \( \gamma \) or \( \theta \) separately. In addition, for fixed \( \theta \in \mathcal{P}(\mathcal{X}) \), \( R(\cdot||\theta) \) is strictly convex on the set \( \{ \gamma \in \mathcal{P}(\mathcal{X}) : R(\gamma||\theta) < \infty \} \).

(c) For each \( \theta \in \mathcal{P}(\mathcal{X}) \), \( R(\cdot||\theta) \) has compact level sets. That is, for each \( M < \infty \) the set \( \{ \gamma \in \mathcal{P}(\mathcal{X}) : R(\gamma||\theta) \leq M \} \) is a compact subset of \( \mathcal{P}(\mathcal{X}) \).

(d) Let \( \mathcal{X} = \mathbb{R}^d \) and let \( \{ \gamma_n, n \in \mathbb{N} \} \) and \( \{ \theta_n, n \in \mathbb{N} \} \) be sequences in \( \mathcal{P}(\mathbb{R}^d) \). Assume that for each \( \alpha \in \mathbb{R}^d \)

\[
\sup_{n \in \mathbb{N}} \int_{\mathbb{R}^d} \exp(\alpha, y) \theta_n(dy) < \infty \quad \text{and} \quad \sup_{n \in \mathbb{N}} R(\gamma_n||\theta_n) \equiv \Delta < \infty.
\]

Then \( \{ \gamma_n \} \) is both tight and uniformly integrable in the sense that

\[
\lim_{C \to \infty} \sup_{n \in \mathbb{N}} \int_{\{ y \in \mathbb{R}^d : \|y\| > C \}} \|y\| \gamma_n(dy) = 0.
\]

(e) Let \( \mathcal{X} = \mathbb{R}^d \) and let \( \gamma \) and \( \theta \) be probability measures on \( \mathbb{R}^d \). Assume that

\[
\int_{\mathbb{R}^d} \exp(\alpha, y) \theta(dy) < \infty
\]

and that \( R(\gamma||\theta) < \infty \). Then \( \int_{\mathbb{R}^d} \|y\| \gamma(dy) < \infty \).
(f) Let $\sigma(dy|x)$ and $\tau(dx|x)$ be stochastic kernels on $\mathcal{Y}$ given $\mathcal{X}$ and $\theta$ a probability measure on $\mathcal{X}$. Then the function mapping $x \in \mathcal{X} \mapsto R(\sigma(.|x)||\tau(.|x))$ is measurable and

$$\int_{\mathcal{X}} R(\sigma(.|x)||\tau(.|x)) \theta(dx) = R(\theta \otimes \sigma || \theta \otimes \tau).$$

(g) We denote by $\Pi$ the class of all finite measurable partitions of $\mathcal{X}$. Then for each $\gamma$ and $\theta$ in $\mathcal{P}(\mathcal{X})$

$$R(\gamma||\theta) = \sup_{\pi \in \Pi} \sum_{A \in \pi} \gamma(A) \log \frac{\gamma(A)}{\theta(A)},$$

where the summand equals $0$ if $\gamma(A) = 0$ and equals $\infty$ if $\gamma(A) > 0$ and $\theta(A) = 0$. In addition, if $A$ is any Borel subset of $\mathcal{X}$, then

$$R(\gamma||\theta) \geq \gamma(A) \log \frac{\gamma(A)}{\theta(A)} - 1.$$

**Proof.** (a) This is proved in Section C.2.

(b) To prove the first assertion, we use the variational formula stated in part (a). For each fixed $g \in C_b(\mathcal{X})$, the function mapping

$$(\gamma, \theta) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \mapsto \int_{\mathcal{X}} g d\gamma - \log \int_{\mathcal{X}} e^g d\theta$$

is convex and continuous. As the supremum over $g \in C_b(\mathcal{X})$ of such functions, $R(\gamma||\theta)$ is a convex, lower semicontinuous function of $(\gamma, \theta) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$. To prove the strict convexity, we use

$$R(\gamma||\theta) = \int_{\mathcal{X}} \frac{d\gamma}{d\theta} \left( \log \frac{d\gamma}{d\theta} \right) d\theta,$$

which is valid for any $\gamma \in \mathcal{P}(\mathcal{X})$ satisfying $R(\gamma||\theta) < \infty$. The strict convexity of $R(\cdot||\theta)$ on the set $\{\gamma \in \mathcal{P}(\mathcal{X}) : R(\gamma||\theta) < \infty\}$ follows from the strict convexity of $s \log s$ for $s \in [0, \infty)$. This completes the proof of part (b).

(c) We follow the proof of Lemma 5.1 in [34]. Let $\{\gamma_n, n \in \mathbb{N}\}$ be any sequence in $\mathcal{P}(\mathcal{X})$ satisfying $\sup_{n \in \mathbb{N}} R(\gamma_n||\theta) \leq M < \infty$. According to the variational formula stated in part (a), for any bounded measurable function $\psi$ mapping $\mathcal{X}$ into $\mathbb{R}$ we have for each $n \in \mathbb{N}$

$$\int_{\mathcal{X}} \psi d\gamma_n - \log \int_{\mathcal{X}} e^\psi d\theta \leq R(\gamma_n||\theta) \leq M.$$

Let $\delta > 0$ and $\varepsilon > 0$ be given. The tightness of $\theta$ guarantees that there exists a compact set $K$ such that $\theta(K^c) \leq \varepsilon$ [Theorem A.2.3]. Substituting into the last display the function $\psi$ that equals $0$ on $K$ and equals $\log(1 + 1/\varepsilon)$ on $K^c$, we have for each $n \in \mathbb{N}$

$$\gamma_n(K^c) \leq \frac{1}{\log(1 + 1/\varepsilon)} \left( M + \log \left[ \theta(K) + \left(1 + \frac{1}{\varepsilon}\right) \theta(K^c) \right] \right)$$

$$= \frac{1}{\log(1 + 1/\varepsilon)} \left( M + \log \left[ 1 + \frac{1}{\varepsilon} \theta(K^c) \right] \right)$$

$$\leq \frac{1}{\log(1 + 1/\varepsilon)} (M + \log 2).$$
1.4. PROPERTIES OF THE RELATIVE ENTROPY

Since $\varepsilon > 0$ can be chosen so that $\frac{1}{\log(1+1/\varepsilon)}(M + \log 2) \leq \delta$, this formula implies that $\{\gamma_n\}$ is tight. By Prohorov’s Theorem there exists $\gamma \in \mathcal{P}(\mathcal{X})$ and a subsequence of $n \in \mathbb{N}$ such that $\gamma_n \implies \gamma$. The lower semicontinuity of $R(\cdot\|\theta)$ yields

$$R(\gamma\|\theta) \leq \liminf_{n \to \infty} R(\gamma_n\|\theta) \leq M.$$  

This completes the proof that $\{\gamma \in \mathcal{P}(\mathcal{X}) : R(\gamma\|\theta) \leq M\}$ is compact.

(d) As in the proof of part (c), we again follow the proof of Lemma 5.1 in [34]. Let $\sigma$ be any positive number. By considering each coordinate direction separately, we obtain from the assumed bound

$$\sup_{n \in \mathbb{N}} \int_{\mathbb{R}^d} \exp\langle \alpha, y \rangle \theta_n(dy) < \infty$$

for each $\alpha \in \mathbb{R}^d$ the limit

$$\lim_{C \to \infty} \sup_{n \in \mathbb{N}} \int_{\{y \in \mathbb{R}^d : \|y\| > C\}} e^{\sigma\|y\|} \theta_n(dy) = 0. \quad (1.16)$$

Using this together with the bound $\sup_{n \in \mathbb{N}} R(\gamma_n\|\theta_n) \equiv \Delta < \infty$, we will prove that $\{\gamma_n\}$ is uniformly integrable. Since the latter property implies that $\{\gamma_n\}$ is tight, the proof of part (d) will be done.

Since for each $n \in \mathbb{N}$ $R(\gamma_n\|\theta_n)$ is finite, $\gamma_n$ is absolutely continuous with respect to $\theta_n$. Thus we can consider the Radon–Nikodym derivative

$$f_n \doteq \frac{d\gamma_n}{d\theta_n},$$

in terms of which

$$R(\gamma_n\|\theta_n) = \int_{\mathcal{X}} f_n \log f_n \, d\theta_n. \quad (1.17)$$

We need the inequality

$$ab \leq e^{\sigma a} + \frac{1}{\sigma} (b \log b - b + 1), \quad (1.18)$$

valid for $a \geq 0$, $b \geq 0$, and $\sigma \geq 1$. This follows from the fact that

$$\sup_{a \in \mathbb{R}} \{ab - e^{\sigma a}\} = \frac{b}{\sigma} \left(\log \frac{b}{\sigma} - 1\right) \leq \frac{1}{\sigma} (b \log b - b + 1).$$

Since $b \log b - b + 1 \geq 0$ for all $b \geq 0$, we find that for $\sigma \geq 1$ and $C > 0$

$$\sup_{n \in \mathbb{N}} \int_{\{y \in \mathbb{R}^d : \|y\| > C\}} \|y\| \gamma_n(dy)
\leq \sup_{n \in \mathbb{N}} \int_{\{y \in \mathbb{R}^d : \|y\| > C\}} \|y\| f_n(y) \theta_n(dy)
\leq \sup_{n \in \mathbb{N}} \int_{\{y \in \mathbb{R}^d : \|y\| > C\}} e^{\sigma\|y\|} \theta_n(dy)
\leq \frac{1}{\sigma} \sup_{n \in \mathbb{N}} \int_{\mathbb{R}^d} (f_n \log f_n - f_n + 1) \, d\theta_n.$$

Equation (1.17) yields
\[
\frac{1}{\sigma} \sup_{n \in \mathbb{N}} \int_{\mathbb{R}^d} (f_n \log f_n - f_n + 1) \, d\theta_n
= \frac{1}{\sigma} \sup_{n \in \mathbb{N}} \int_{\mathbb{R}^d} f_n \log f_n \, d\theta_n
= \frac{1}{\sigma} \sup_{n \in \mathbb{N}} R(\gamma_n \| \theta_n) = \frac{\Delta}{\sigma}.
\]
Combining this with (1.16), we obtain
\[
\lim \sup_{C \to \infty} \sup_{n \in \mathbb{N}} \int_{\{y \in \mathbb{R}^d : \|y\| > C\}} \|y\| \gamma_n(dy) \leq \frac{\Delta}{\sigma}.
\]
Since \(\sigma \geq 1\) can be taken arbitrarily large, we have completed the proof that \(\{\gamma_n\}\) is uniformly integrable.

(e) Part (d) implies that
\[
\lim_{C \to \infty} \int_{\{y \in \mathbb{R}^d : \|y\| > C\}} \|y\| \gamma(dy) = 0,
\]
which in turn implies that \(\int_{\mathbb{R}^d} \|y\| \gamma(dy) < \infty\).

(f) This is proved in Corollary C.3.2.

(g) This is proved in Section C.4. \(\blacksquare\)

Throughout the remainder of the book we will be making frequent use of relative entropy. It will sometimes be necessary to work simultaneously with the relative entropy associated with a number of different spaces. We will abuse notation and simply write \(R(\cdot \| \cdot)\) to denote the relative entropy in all the different cases. The particular space involved in each circumstance will be obvious.

So far in this chapter we have developed a number of tools that are basic in the weak convergence approach to large deviations. The next section introduces another important tool, the dynamic programming equation.

### 1.5 Stochastic Control Theory and Dynamic Programming

In this section we present some standard material in a somewhat general setting. An essential step in our approach to the study of large deviation phenomena for a sequence of random variables \(\{X^n, n \in \mathbb{N}\}\) will be to obtain variational representation formulas for normalized logarithms of certain expectations. An example of such a quantity is

\[
W^n \doteq -\frac{1}{n} \log E\{\exp[-n \, h(X^n)]\},
\]
which arises in the study of the Laplace principle in Section 1.2. The representation formulas will express large deviation quantities such as $W^n$ in terms of the minimal cost functions of associated stochastic control problems. The Laplace principle will then be proved by using weak convergence theory to study the limit of the minimal cost functions appearing in the representation formulas.

The purpose of the present section is to introduce needed terminology and notation and to recapitulate the well known relationship between minimal cost functions and dynamic programming equations. Dynamic programming is a standard technique in operations research, optimization, and other fields and is discussed in many undergraduate texts such as [57, 80]. Readers familiar with the technique can safely omit this section. Starting in Chapter 2, we will explain in detail how the relationship between minimal cost functions and the dynamic programming equation can be used as a quick and convenient method for deriving the representation formula that is appropriate for each of the large deviation problems that we treat. Numerous examples of representation formulas for large deviation quantities such as $W^n$ will be given throughout the text.

As we will see in each example, the derivation of the representation formula via dynamic programming is more or less automatic and seems to give the most convenient form of the representation. In some cases the representation formula can also be derived by a measure-theoretic argument, which however has some disadvantages. For example, the measure-theoretic derivation can easily give an awkward form of the representation if one happens to pick an inappropriate description of the underlying process. An instance of the measure-theoretic derivation is given in Appendix B.

We begin by defining a stochastic control problem, which involves a controlled process and a cost structure. The problem is defined on the discrete time set $\{0, 1, \ldots, n\}$, where $n \in \mathbb{N}$ is fixed. The controlled process is denoted by $\{\hat{Z}_j, j = 0, 1, \ldots, n\}$. $\hat{Z}_j$ is called the state variable of the controlled process at time $j$. It takes values in a Polish space $\mathcal{Z}_j$. The time evolution of $\hat{Z}_j$, to be specified below, involves a control $u_j$ applied at time $j$. The controls $u_j$ take values in a Polish space $\mathcal{U}$. We will consider controls that are of feedback type; namely, for each $j \in \{0, 1, \ldots, n - 1\}$ $u_j$ is a measurable function mapping $\mathcal{Z}_j$ into $\mathcal{U}$. By an admissible control sequence we will mean a sequence $\{u_j, j = 0, 1, \ldots, n - 1\}$ of feedback controls. In all of the examples to be treated in this book, the space $\mathcal{U}$ will always be the space of probability measures on some Polish space $\mathcal{Y}$ and each control $u_j$ will be a stochastic kernel on $\mathcal{Y}$ given $\mathcal{Z}_j$.

There are a wide variety of different stochastic control problems that are considered in the literature. The particular control problem that we consider is the simplest formulation that covers all our needs. In order to specify the control problem, we must describe the dynamics—that is, the evolution of the state variables $\hat{Z}_j$—and the cost structure.

**Dynamics.** The evolution of the state variable $\mathcal{Z}_j$ is defined in terms of a family of probability measures $p_j(z, d\xi|u)$ on $\mathcal{Z}_{j+1}$ indexed by $(z, u) \in \mathcal{Z}_j \times \mathcal{U}$. Specifically, we assume that for each $j \in \{0, 1, \ldots, n - 1\}$ $p_j(z, d\xi|u)$ is a stochastic kernel on $\mathcal{Z}_{j+1}$ given $\mathcal{Z}_j \times \mathcal{U}$. Let $\{u_j, j = 0, 1, \ldots, n - 1\}$ be any admissible control sequence. Then

$$q_j(z, d\xi) = p_j(z, d\xi|u_j(z))$$
defines a stochastic kernel on \( \mathcal{Z}_{j+1} \) given \( \mathcal{Z}_j \). Although in a strict sense \( q_j(z, d\xi) \) is not a transition probability function unless \( \mathcal{Z}_j \) and \( \mathcal{Z}_{j+1} \) are equal, it is useful in general to think of \( q_j(z, d\xi) \) in this way, and we will do so. The controlled process \( \{\tilde{Z}_j, j = 0, 1, \ldots, n\} \) is defined to be the Markov chain that has the respective state spaces \( \{\mathcal{Z}_j, j = 0, 1, \ldots, n\} \) and the nonstationary transition probability functions \( \{q_j(z, d\xi), j = 0, 1, \ldots, n - 1\} \). In other words, we define the evolution of the controlled random variables by

\[
\tilde{P}\{\tilde{Z}_{j+1} \in d\xi | \tilde{Z}_j = z\} \triangleq p_j(z, d\xi | u_j(z))
\]  

(1.19)

for \( z \in \mathcal{Z}_j \). Depending on the problem, the state variable \( \tilde{Z}_0 \) at time 0 will equal a fixed value \( z \in \mathcal{Z}_0 \) or, more generally, will have a given distribution on \( \mathcal{Z}_0 \).

In the sequel many such Markov chains with time-dependent state spaces will be encountered, and for each time \( i \) it will be useful to have a version of the expectation conditioned on \( \tilde{Z}_i = z \) which is defined for all \( z \in \mathcal{Z}_i \). In each case we adapt to the time-dependent setting the standard definition that is used when all the state spaces coincide. Namely, this version is the one defined by iterated integration of the transition probability functions. Thus in the present case, given a measurable function \( F \) mapping \( \mathcal{Z}_i \times \mathcal{Z}_{i+1} \times \cdots \times \mathcal{Z}_n \) into \( \mathbb{R} \) we define the expectation of \( F(\tilde{Z}_i, \tilde{Z}_{i+1}, \ldots, \tilde{Z}_n) \) conditioned on \( \tilde{Z}_i = z \) by

\[
\mathbb{E}_{i, z}\{F(Z_i, Z_{i+1}, \ldots, Z_n)\} \\
\triangleq \int_{\mathcal{Z}_{i+1} \times \mathcal{Z}_{i+2} \times \cdots \times \mathcal{Z}_n} F(z, z_{i+1}, \ldots, z_n) q_i(z, dz_{i+1}) q_{i+1}(z_{i+1}, dz_{i+2}) \cdots q_{n-1}(z_{n-1}, dz_n).
\]

**Cost structure.** For each \( j \in \{0, 1, \ldots, n-1\} \) let \( c_j \) be a nonnegative measurable function mapping \( \mathcal{Z}_j \times \mathcal{U} \) into \([0, \infty]\) with the added property that \( c_j(z, \cdot) \) is not identically \( \infty \) for each \( z \in \mathcal{Z}_j \). Also let \( f \) be a bounded measurable function mapping \( \mathcal{Z}_n \) into \( \mathbb{R} \). For each \( i \in \{0, 1, \ldots, n\} \) and \( z \in \mathcal{Z}_i \), we define

\[
V(i, z) \triangleq \inf_{\{u_j\}} \mathbb{E}_{i, z}\left\{ \sum_{j=i}^{n-1} c_j(\tilde{Z}_j, u_j(\tilde{Z}_j)) + f(\tilde{Z}_n) \right\}.
\]  

(1.20)

In this formula the infimum is taken over all admissible control sequences \( \{u_j, j = 0, 1, \ldots, n - 1\} \), \( \{\tilde{Z}_j, j = 0, 1, \ldots, n\} \) is the controlled process that is associated with a particular admissible control sequence \( \{u_j\} \), and \( \mathbb{E}_{i, z} \) denotes expectation conditioned on \( \tilde{Z}_i = z \). When \( i = n \) we define the sum in (1.20) to be 0. Although the functions \( c_j \) may be unbounded, the nonnegativity of each \( c_j \) guarantees that \( V(i, z) \) is well defined. In formula (1.20), the sum \( \sum_{j=i}^{n-1} c_j(\tilde{Z}_j, u_j(\tilde{Z}_j)) \) is referred to as the **running cost** while \( f(\tilde{Z}_n) \) is called the **terminal cost**. In every large deviation problem that we consider, the functions \( c_j \) will be defined in terms of relative entropy. The quantities \( V(i, z) \) are called **minimal cost functions**. It is convenient to use the term “running cost” also to refer to the expected value of the sum in formula (1.20).

In summary, the definition of the stochastic control problem involves the following five items.
1.5. **DYNAMIC PROGRAMMING**

- A sequence of Polish spaces \( \{ \mathcal{Z}_j : j = 0, 1, \ldots, n \} \), \( \mathcal{Z}_j \) being the state space of the controlled process at time \( j \). We call \( \mathcal{Z}_j \) simply the **state space at time \( j \)**.

- A Polish space \( \mathcal{U} \), which is the space in which the controls \( \{ u_j : j = 0, 1, \ldots, n - 1 \} \) take values. We call \( \mathcal{U} \) the **control space**.

- Stochastic kernels \( p_j(z, d\xi | u) \) on \( \mathcal{Z}_{j+1} \) given \( \mathcal{Z}_j \times \mathcal{U} \), in terms of which the evolution of the controlled process is defined. We call \( \{ p_j(z, d\xi | u) : j = 0, \ldots, n - 1 \} \) the sequence of **controlled transition probability functions**.

- Nonnegative measurable functions \( c_j : \mathcal{Z}_j \times \mathcal{U} \rightarrow [0, \infty] \) such that \( c_j(z, \cdot) \) is not identically \( \infty \) for each \( z \in \mathcal{Z}_j \). The running cost in the minimal cost functions is defined in terms of these functions. We call \( c_j \) the **running cost at time \( j \)** or simply the **running cost**.

- A bounded measurable function \( f : \mathcal{Z}_n \rightarrow \mathbb{R} \), in terms of which the terminal cost in the minimal cost functions is defined. We also call \( f \) the **terminal cost**.

In order to determine the variational representation that is appropriate for a particular large deviation problem, one must understand how, starting from the large deviation problem, one defines the spaces \( \mathcal{Z}_j \) and \( \mathcal{U} \) and the functions \( p_j(z, d\xi | u) \), \( c_j \), and \( f \). This is carried out by means of the dynamic programming equation, which we now introduce.

**The Dynamic Programming Equation.** As we will see just below, the Markov property of the controlled process under an admissible control sequence turns out to imply a recursive equation relating the minimal cost functions \( V(i, \cdot) \) and \( V(i + 1, \cdot) \). Specifically, for any \( i \in \{0, 1, \ldots, n - 1\} \) and any \( z \in \mathcal{Z}_i \)

\[
V(i, z) = \inf_{u \in \mathcal{U}} \left\{ c_i(z, u) + \int_{\mathcal{Z}_{i+1}} V(i + 1, \xi) p_i(z, d\xi | u) \right\}. \tag{1.21}
\]

The form of this equation is easy to motivate. Indeed, the quantity within the brackets on the right hand side equals the cost that one obtains, given that one is in state \( z \) at time \( i \), if one uses the control \( u \) at time \( i \) and proceeds optimally thereafter. The minimal cost function \( V(i, z) \) equals the infimum of this quantity over all controls in the control space \( \mathcal{U} \).

The recursive equations (1.21) are called collectively the **dynamic programming equation** for the control problem. It follows from the definition of the minimal cost function \( V(n, z) \) that

\[
V(n, z) = f(z). \tag{1.22}
\]

We call this the **terminal condition** for the control problem. In a moment we will verify that under Condition 1.5.1 the minimal cost functions \( V(i, z) \) are bounded and measurable and they satisfy the dynamic programming equation (1.21). Backward induction on \( i \) shows that for a given terminal cost \( f \) equations (1.21) and (1.22) have a unique solution. Thus for all \( i \in \{0, 1, \ldots, n\} \) and all \( z \in \mathcal{Z}_i \) these equations uniquely characterize the minimal cost functions \( V(i, z) \) defined in (1.20). An important consequence is that any
other solution of (1.21) and (1.22) equals these minimal cost functions and thus can be represented in terms of them.

In most situations equations having the form (1.21) and (1.22) are encountered in the context of analyzing a given stochastic control problem. Our application of the dynamic programming equation will be completely different. Namely, we will use it to represent a large deviation quantity such as

\[ W^n = -\frac{1}{n} \log E\{\exp[-n h(X^n)]\} \]

in terms of the minimal cost function of an associated stochastic control problem. The procedure will be explained in detail in a basic example to be given in the next chapter and will be used later in the book. Hence we content ourselves here with a loose description. Starting with \( W^n \) and related functions, we will derive equations having the form (1.21) and (1.22). From them we will read off the data defining a stochastic control problem; viz., the state spaces \( Z_j \), the control space \( U \), the sequence of controlled transition probability functions \( p_j(z, d\xi|u) \), the running costs \( c_j \), and the terminal cost \( f \). The minimal cost functions then defined in equation (1.20) will be used as a representation for the solution to the equations having the form (1.21) and (1.22). In particular, the large deviation quantity \( W^n \) will be represented in terms of \( V \). The precise form of the representation will be determined by the specific spaces and functions that play the roles of \( Z_j \), \( U \), \( p_j(z, d\xi|u) \), \( c_j \), and \( f \).

We close this section by proving that the minimal cost functions \( V(i, z) \) are bounded and measurable and that they satisfy the dynamic programming equation

\[ \bar{V}(i, z) = \inf_{u \in U} \left\{ c_i(z, u) + \int_{Z_{i+1}} \bar{V}(i + 1, \xi) p_{i}(z, d\xi|u) \right\} \]  \hspace{1cm} (1.23)

and the terminal condition \( \bar{V}(n, z) = f(z) \). We do this under the following assumption, which we call the **Attainment Condition**.

**Condition 1.5.1.** There exists an admissible control sequence \( \{\bar{u}_i, i = 0, 1, \ldots, n - 1\} \) such that the infimum in each of the formulas (1.23) is attained at \( \bar{u}_i \); i.e.,

\[ \bar{V}(i, z) = c_i(z, \bar{u}_i(z)) + \int_{Z_{i+1}} \bar{V}(i + 1, \xi) p_{i}(z, d\xi|\bar{u}_i(z)). \] \hspace{1cm} (1.24)

Such an infimizing admissible control sequence \( \{\bar{u}_i\} \) is called an **optimal control**. The existence of an optimal control does not automatically hold under the assumptions that we have stated so far. However, the Attainment Condition is valid for all of the stochastic control problems that are associated with the large deviation problems to be considered in subsequent chapters. The general case is discussed in [8]. The following theorem is an example of a verification theorem in control theory; see, e.g., [53].

**Theorem 1.5.2.** We define the sequence of minimal cost functions \( \{V(i, z), i = 0, 1, \ldots, n\} \) by formula (1.20). If the Attainment Condition given in Condition 1.5.1 is valid, then these minimal cost functions are bounded and measurable and are the unique solution of the dynamic programming equation (1.23) and the terminal condition \( V(n, z) = f(z) \).
1.5. DYNAMIC PROGRAMMING

Proof. That the dynamic programming equation (1.23) and the terminal condition have a unique solution \( \{ \hat{V}(i, z) \} \) is immediate. The crux of the proof is to show that the minimal cost functions \( V(i, z) \) are this unique solution.

The terminal cost \( f \) is a bounded measurable function mapping \( \mathcal{Z}_n \) into \( \mathbb{R} \). Since \( c_i(z, \cdot) \) is not identically \( \infty \) for each \( i \in \{0, 1, \ldots, n - 1\} \) and \( z \in \mathcal{Z}_i \), backward induction on \( i \) and the form of the dynamic programming equation show that the functions \( \hat{V}(i, z) \) are bounded functions of \( z \in \mathcal{Z}_i \). By hypothesis there exists an admissible control sequence \( \{u_i\} \) such that equation (1.24) in the Attainment Condition is valid. It follows again by backward induction on \( i \) that the functions \( \hat{V}(i, z) \) are measurable functions of \( z \in \mathcal{Z}_i \).

We complete the proof of the theorem by showing that \( V(i, z) \) and \( \hat{V}(i, z) \) are equal. For \( i = n \) this is obvious since \( V(n, z) = f(z) = \hat{V}(n, z) \).

We next prove that for \( i \in \{0, 1, \ldots, n-1\} \) \( V(i, z) \geq \hat{V}(i, z) \). Let \( \{u_j, j = 0, 1, \ldots, n-1\} \) be any admissible control sequence and let \( \{\hat{Z}_j, j = 0, 1, \ldots, n\} \) be the associated controlled process. Then we obtain from the dynamic programming equation relating \( \hat{V}(j, \cdot) \) and \( \hat{V}(j + 1, \cdot) \) the inequality

\[
c_j(z, u_j(z)) \geq \hat{V}(j, z) - \int_{\mathcal{Z}_{j+1}} \hat{V}(j + 1, \xi) p_j(z, d\xi | u_j(z)).
\]

Since \( p_j(z, d\xi | u) \) is a stochastic kernel on \( \mathcal{Z}_{j+1} \) given \( \mathcal{Z}_j \times \mathcal{U} \), equation (1.19) implies that \( p_j(\hat{Z}_j, d\xi | u_j(\hat{Z}_j)) \) is a regular conditional distribution for \( \hat{Z}_{j+1} \) given \( \hat{Z}_j \). The equality \( V(n, z) = f(z) \) and the Markov property yield

\[
\begin{align*}
E_{i,z} \left\{ \sum_{j=i}^{n-1} c_j(Z_j, u_j(Z_j)) + f(Z_n) \right\} \\
\geq E_{i,z} \left\{ \sum_{j=i}^{n-1} \left[ V(j, Z_j) - \int_{\mathcal{Z}_{j+1}} \hat{V}(j + 1, \xi) p_j(Z_j, d\xi | u_j(Z_j)) \right] + V(n, Z_n) \right\} \\
= E_{i,z} \left\{ \sum_{j=i}^{n-1} \left[ V(j, Z_j) - E\{ V(j + 1, Z_{j+1}) | Z_j \} \right] + V(n, Z_n) \right\} \\
= E_{i,z} \left\{ \sum_{j=i}^{n-1} \left[ \hat{V}(j, \hat{Z}_j) - \hat{V}(j + 1, \hat{Z}_{j+1}) \right] + V(n, \hat{Z}_n) \right\} \\
= \hat{V}(i, z).
\end{align*}
\]

By the definition of \( V(i, z) \) as an infimum over all admissible control sequences, it follows that \( V(i, z) \geq \hat{V}(i, z) \).

We now prove the reverse inequality, assuming that \( \{\bar{u}_j\} \) is an admissible control sequence that satisfies equation (1.24) in the Attainment Condition. Then repeating the steps given in the previous paragraph, we see that with \( u_j \equiv \bar{u}_j \) the inequality in the last display becomes an equality; i.e.,

\[
E_{i,z} \left\{ \sum_{j=i}^{n-1} c_j(Z_j, \bar{u}_j(Z_j)) + f(Z_n) \right\} = \hat{V}(i, z).
\]
Again by the definition of $V(i, z)$ as an infimum over all admissible control sequences, it follows that $V(i, z) \leq \bar{V}(i, z)$. This completes the proof that $V(i, z) = \bar{V}(i, z)$. We conclude that the minimal cost functions $V(i, z)$ are bounded and measurable and that they satisfy the dynamic programming equation (1.23) and the terminal equation $V(n, z) = f(z)$. ■

In the next chapter we will illustrate the procedure of the present section by deriving a representation formula for large deviation quantities arising in a basic example; namely, the empirical measures of i.i.d. random variables (Sanov’s Theorem). We will then use weak convergence methods to study the asymptotic behavior of the minimal cost functions appearing in this representation. In this way we will obtain the Laplace principle for the empirical measures.
Chapter 2

A First Example: Sanov’s Theorem

2.1 Introduction

In order to introduce the ideas of the weak convergence approach, we first treat a well known, classical problem involving independent, identically distributed (i.i.d.) random variables. This problem is Sanov’s Theorem, which states the large deviation principle for the empirical measures and can be proved by completely standard methods. The ideas of the weak convergence approach to be presented here will be generalized later on to treat much more difficult problems. A second classical problem for i.i.d. random variables is Mogulskii’s Theorem, which is the large deviation principle for a sequence of continuous-time stochastic processes constructed from the i.i.d. random variables. We will discuss Mogulskii’s Theorem in Chapter 3.

Let \( \{X_j, j \in \mathbb{N}_0\} \) be a sequence of i.i.d. random variables that take values in a Polish space \( \mathcal{S} \) and let \( \{L^n, n \in \mathbb{N}\} \) be the corresponding sequence of empirical measures. The empirical measures take values in the space \( \mathcal{P}(\mathcal{S}) \) of probability measures on \( \mathcal{S} \). As we discuss in the next section, Sanov’s theorem will be proved by verifying the Laplace principle for the empirical measures. This entails calculating the asymptotic behavior of the large deviation quantities

\[
W^n = -\frac{1}{n} \log E\{\exp[-n h(L^n)]\},
\]

where \( h \) is any bounded continuous function mapping \( \mathcal{P}(\mathcal{S}) \) into \( \mathbb{R} \). A key step in this calculation is to represent \( W^n \) in terms of the minimal cost function of an associated stochastic control problem. We present two distinct derivations of this representation formula, the first via the dynamic programming formalism of Section 1.5 and the second via purely measure-theoretic considerations. While in the present case the second derivation is rather straightforward, it has the disadvantage, in contrast to the dynamic programming approach, of not readily generalizing to the more complicated large deviation problems to be considered later in the book. We present the dynamic programming derivation in Section 2.3 of this chapter and the measure-theoretic derivation in Appendix B. There is a large literature on Sanov’s Theorem and its applications; see, e.g., [3, 5, 24, 28, 29, 34, 56].

After the representation formula is proved, we turn to the verification of the Laplace
principle in Sections 2.4 and 2.5. In these sections we evaluate the asymptotic behavior of the minimal cost functions that appear in the representation formula for \( W^n \), using the theory of weak convergence of probability measures.

2.2 Statement of Sanov’s Theorem

Let \( S \) be a Polish space and \( \rho \) a probability measure on \( S \). We consider a sequence of i.i.d. random variables \( \{X_j, j \in \mathbb{N}_0\} \) that are defined on a probability space \( (\Omega, \mathcal{F}, P) \), take values in \( S \), and have the common distribution \( \rho \). For each \( n \in \mathbb{N}, \omega \in \Omega, \) and Borel subset \( A \) of \( S \), we define the empirical measure, or the normalized occupation measure up to time \( n - 1 \), by

\[
L^n(\omega, A) = L^n(A) = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{X_j(\omega)}(A),
\]

where for \( x \in S \) \( \delta_x \) denotes the unit point measure at \( x \). \( L^n(A) \) counts the relative frequency with which the random variables \( \{X_j, j = 0, 1, \ldots, n-1\} \) lie in \( A \). The empirical measure \( L^n \) takes values in \( \mathcal{P}(S) \). In Section A.3 we define on this space the well-known Lévy–Prohorov metric \( \mathcal{L}(\cdot, \cdot) \). This metric is compatible with the weak topology in the sense that \( \theta_n \Rightarrow \theta \) in \( \mathcal{P}(S) \) if and only if \( \mathcal{L}(\theta_n, \theta) \to 0 \). Furthermore, with respect to this metric \( \mathcal{P}(S) \) is a Polish space [Theorem A.3.2].

We claim that w.p.1 the sequence of empirical measures \( \{L^n, n \in \mathbb{N}\} \) converges weakly to \( \rho \). The law of large numbers takes us most of the way. In fact, for each bounded continuous function \( g \) mapping \( S \) into \( \mathbb{R} \), there exists a null set depending on \( g \) such that if \( \omega \) lies in the complement of this null set, then

\[
\lim_{n \to \infty} \int_S g(x) L^n(\omega, dx) = \lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} g(X_j(\omega)) = \int_S g(x) \rho(dx).
\]

If there exists a fixed null set such that this limit holds for all bounded continuous \( g \) whenever \( \omega \) lies in the complement of the null set, then the probability–1 weak convergence \( L^n \Rightarrow \rho \) will be proved. The existence of such a null set can be proved by a standard separability argument, which is a consequence of Theorem A.6.1. According to this theorem, there exists an equivalent metric \( m(x, y) \) on \( S \) with the property that the space \( \mathcal{U}_b(S, m) \) of bounded, uniformly continuous functions mapping \( S \) into \( \mathbb{R} \) is separable with respect to the uniform metric.

We now complete the proof that \( L^n \Rightarrow \rho \) w.p.1. Let \( \Xi \) be a countable dense subset of \( \mathcal{U}_b(S, m) \). As we have seen, for each \( g \in \Xi \) there exists a null set \( N(g) \) such that if \( \omega \notin N(g) \), then

\[
\lim_{n \to \infty} \int_S g(x) L^n(\omega, dx) = \int_S g(x) \rho(dx).
\]

It follows that for any \( f \in \mathcal{U}_b(S, m) \) and any \( \omega \) in the complement of the null set \( \cup_{g \in \Xi} N(g) \)

\[
\lim_{n \to \infty} \int_S f(x) L^n(\omega, dx) = \int_S f(x) \rho(dx).
\]
2.2. STATEMENT OF SANOV’S THEOREM

The Portmanteau Theorem [Theorem A.3.4] now implies that the previous display holds for any bounded continuous function $f$ mapping $S$ into $\mathcal{R}$ and any $\omega \not\in \bigcup_{g \in \mathbb{E}} N(g)$. This completes the proof that $L^n \Rightarrow \rho$ w.p.1.

Before proceeding, we recall the relative entropy, a number of properties which are given in Section 1.4. The following properties are needed in this chapter. If $\gamma$ and $\theta$ are two probability measures on $S$, then $R(\gamma\|\theta) \geq 0$ and $R(\gamma\|\theta) = 0$ if and only if $\gamma = \theta$. In addition, $R(\cdot\|\theta)$ is convex and lower semicontinuous and it has compact level sets. It follows that $R(\cdot\|\theta)$ is a rate function on $\mathcal{P}(S)$ and that if $\Phi$ is a closed subset of $\mathcal{P}(S)$ not containing $\theta$, then $\lim_{\gamma \in \Phi} R(\gamma\|\theta) > 0$.

Let us now return to the empirical measures and Sanov’s Theorem. If $\Phi$ is any closed subset of $\mathcal{P}(S)$ not containing $\rho$, then the probability–1 weak convergence $L^n \Rightarrow \rho$ implies that

$$\lim_{n \to \infty} P\{L^n \in \Phi\} = 0. \quad (2.1)$$

Sanov’s Theorem states that the sequence $\{L^n\}$ satisfies the large deviation principle with rate function $R(\cdot\|\rho)$. Since the closed subset $\Phi$ does not contain $\rho$,

$$\limsup_{n \to \infty} \frac{1}{n} \log P\{L^n \in \Phi\} \leq - \inf_{\gamma \in \Phi} R(\gamma\|\rho) < 0.$$

Hence the convergence to 0 in formula (2.1) is exponentially fast.

The main purpose of this chapter is to introduce the weak convergence approach in the context of proving the following theorem. The theorem states the Laplace principle [Definition 1.2.2], which by Theorems 1.2.1 and 1.2.3 is equivalent to the large deviation principle with the same rate function.

**Theorem 2.2.1 (Sanov).** Let $S$ be a Polish space, $\rho$ a probability measure on $S$, and $\{L^n, n \in \mathbb{N}\}$ the sequence of empirical measures of i.i.d. random variables taking values in $S$ and having common distribution $\rho$. Then $\{L^n\}$ satisfies the Laplace principle on $\mathcal{P}(S)$ with rate function $R(\cdot\|\rho)$. In other words, for all bounded continuous functions $h$ mapping $\mathcal{P}(S)$ into $\mathcal{R}$

$$\lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(L^n)]\} = - \inf_{\nu \in \mathcal{P}(S)} \{R(\nu\|\rho) + h(\nu)\}.$$

Even the case of a finite state space $S$ is of interest. It arises in the study of the Curie–Weiss–Potts model in statistical mechanics [47, 48]. For an appropriate choice of the function $h$ the limit in Theorem 2.2.1 defines a quantity that is proportional to the specific Gibbs free energy of the model.

We have already remarked that $R(\cdot\|\rho)$ is a rate function on $\mathcal{P}(S)$. In order to prove that $\{L^n\}$ satisfies the Laplace principle on $\mathcal{P}(S)$ with this rate function, we define

$$W^n = - \frac{1}{n} \log E\{\exp[-n h(L^n)]\}.$$

As noted in the introduction, a basic step in the proof of the Laplace principle is to represent $W^n$ in terms of the minimal cost function of an associated stochastic control problem. This is carried out in the next section.
2.3 The Representation Formula

Let \( M(\mathcal{S}) \) denote the space of subprobability measures on \( \mathcal{S} \); i.e., measures on \( \mathcal{S} \) having total mass not exceeding 1. \( P(\mathcal{S}) \) is a subset of \( M(\mathcal{S}) \), and like \( P(\mathcal{S}) \) \( M(\mathcal{S}) \) is topologized with the topology corresponding to weak convergence of measures. In Section A.3 we introduce a metric on \( M(\mathcal{S}) \) which is compatible with the weak topology on this space and with respect to which \( M(\mathcal{S}) \) is a Polish space. For \( c \in [0, 1] \) \( M_c(\mathcal{S}) \) is defined to be the closed set \( \{ \mu \in M(\mathcal{S}) : \mu(\mathcal{S}) = c \} \).

In order to represent \( W^n \) in terms of the minimal cost function of an associated stochastic control problem, it is useful to introduce a sequence of random subprobability measures that are closely related to the empirical measures. For \( n \in \mathbb{N} \) and \( j \in \{0, 1, \ldots, n - 1\} \) we define \( L^n_0 = 0 \) and

\[
L^n_{j+1} = L^n_j + \frac{1}{n} \delta_{X_j}.
\]

Thus

\[
L^n_j = \frac{1}{n} \sum_{k=0}^{j-1} \delta_{X_k},
\]

and \( L^n_j \) equals the empirical measure \( L^n \). \( L^n_j \) takes values in \( M_{j/n}(\mathcal{S}) \). We refer to the random measures \( \{L^n_j\} \) as the empirical subprobability measures of \( \{X_i\} \).

It is also useful to have quantities analogous to \( W^n \) which are defined for each \( i \in \{0, 1, \ldots, n\} \) and each measure \( \mu \in M_{i/n}(\mathcal{S}) \) by conditioning on the event \( \{L^n_i = \mu\} \). Accordingly, we introduce

\[
W^n(i, \mu) \doteq -\frac{1}{n} \log E_{i,\mu} \{ \exp[-n h(L^n)] \},
\]

where \( E_{i,\mu} \) denotes expectation conditioned on \( L^n_i = \mu \) and \( h \) is the same bounded continuous function that appears in the definition of \( W^n \). \( W^n(0, 0) \) equals

\[
W^n \doteq -\frac{1}{n} \log E \{ \exp[-n h(L^n)] \}
\]

while \( W^n(n, \mu) \) equals \( h(\mu) \). Each \( W^n(i, \cdot) \) is a bounded measurable function on \( M_{i/n}(\mathcal{S}) \). For the remainder of this section the scaling parameter \( n \) is fixed.

In order to obtain a representation formula for \( W^n \), we first derive a recursive equation relating \( W^n(i, \cdot) \) and \( W^n(i+1, \cdot) \) and then interpret this equation as the dynamic programming equation of an associated stochastic control problem.

To carry out the first step, we look at the random measures \( \{L^n_i, i = 0, 1, \ldots, n\} \) in a perhaps novel way. Since the random variables \( X_i \) are i.i.d. with common distribution \( \rho \), it follows that these random measures form a Markov chain with respective state spaces \( \{M_{i/n}(\mathcal{S}), i = 0, 1, \ldots, n\} \). For \( i \in \{0, 1, \ldots, n - 1\} \) and \( A \) a Borel subset of \( M_{(i+1)/n}(\mathcal{S}) \), the transition probability function of the Markov chain is given by

\[
P\{L^n_{i+1} \in A | L^n_i = \mu\} = P\left\{ \mu + \frac{1}{n} \delta_{X_i} \in A \right\} = \int_{\mathcal{S}} 1_A \left( \mu + \frac{1}{n} \delta_y \right) \rho(dy).
\]

(2.2)
2.3. THE REPRESENTATION FORMULA

Taking \( i \in \{0, 1, \ldots, n - 1 \} \) and \( \mu \in \mathcal{M}_{i/n}(S) \) and using the Markov property, we obtain

\[
\exp[-n \, W^n(i, \mu)] = \mathcal{E}_{i, \mu} \{ \exp[-n \, h(L^n)] \} = \mathcal{E}_{i, \mu} \left\{ \mathcal{E}_{i+1, L^{i+1}_n} \{ \exp[-n \, h(L^n)] \} \right\} = \mathcal{E}_{i, \mu} \{ \exp[-n \, W^n(i + 1, L^{i+1}_n)] \} = \int_S \exp \left[-n \, W^n \left(i + 1, \mu + \frac{1}{n} \delta_y \right) \right] \rho(dy).
\]

Thus

\[
W^n(i, \mu) = -\frac{1}{n} \log \int_S \exp \left[-n \, W^n \left(i + 1, \mu + \frac{1}{n} \delta_y \right) \right] \rho(dy).
\]

We now apply part (a) of Proposition 1.4.2 to the space \( V \equiv S \), the function \( k(y) \equiv n \, W^n(i + 1, \mu + \frac{1}{n} \delta_y) \), and the measure \( \theta \equiv \rho \). This gives the variational formula

\[
W^n(i, \mu) = \inf_{\nu \in \mathcal{P}(S)} \left\{ \frac{1}{n} R(\nu \| \rho) + \int_S W^n \left(i + 1, \mu + \frac{1}{n} \delta_y \right) \nu(dy) \right\}
\]

or

\[
W^n(i, \mu) = \inf_{\nu \in \mathcal{P}(S)} \left\{ \frac{1}{n} R(\nu \| \rho) + \int_{\mathcal{M}(S)} W^n(i + 1, \theta) \, p(\mu, d\theta | \nu) \right\}, \tag{2.3}
\]

where for \( A \) a Borel subset of \( \mathcal{M}_{(i+1)/n}(S) \)

\[
p(\mu, A | \nu) \equiv \int_S 1_A \left( \mu + \frac{1}{n} \delta_y \right) \nu(dy).
\]

In addition, we have the terminal condition

\[
W^n(n, \mu) = h(\mu). \tag{2.4}
\]

In order to derive the representation formula for \( W^n \), we must determine the stochastic control problem for which (2.3) is the dynamic programming equation and (2.4) the terminal condition. For this purpose we return to Section 1.5, where a dynamic programming equation appears relating the minimal cost functions \( V(i, \cdot) \) and \( V(i + 1, \cdot) \) defined in formula (1.20). For \( z \in Z_i \) the dynamic programming equation has the form

\[
V(i, z) = \inf_{u \in \mathcal{U}} \left\{ c_i(z, u) + \int_{Z_{i+1}} V(i + 1, \xi) \, p_i(z, d\xi | u) \right\}.
\]

Comparing this formula with formula (2.3), we have the following identifications of the spaces \( Z_i \) and \( \mathcal{U} \) and the functions \( p_i(z, d\xi | u) \), \( c_i \), and \( f \).

- The state space \( Z_i \) equals \( \mathcal{M}_{i/n}(S) \).
- The control space \( \mathcal{U} \) equals \( \mathcal{P}(S) \). Let \( \nu_i \in \mathcal{P}(S) \) denote the control applied at time \( i \). We exhibit the dependence of \( \nu_i \) upon \( \mu \in \mathcal{M}_{i/n}(S) \) by writing \( \nu_i = \nu_i(\cdot | \mu) \). Since the controls \( \nu_i \) are to be of feedback type, we require that \( \nu_i(\cdot | \mu) \) is a stochastic kernel on \( S \) given \( \mathcal{M}_{i/n}(S) \).
• Given \( \mu \in \mathcal{M}_{i/n}(\mathcal{S}) \), \( A \) a Borel subset of \( \mathcal{M}_{(i+1)/n}(\mathcal{S}) \), and \( \nu \in \mathcal{P}(\mathcal{S}) \), the controlled transition probability function \( p_i(\mu, d\theta|\nu) \) is defined by

\[
p_i(\mu, A|\nu) \doteq \int_A 1_A \left( \mu + \frac{1}{n} \delta_y \right) \nu(dy).
\]

• Each of the running costs is defined by the same formula

\[
c_i(\nu) \doteq \frac{1}{n} R(\nu||\rho) \text{ for } \nu \in \mathcal{P}(\mathcal{S}).
\]

• The terminal cost \( f \) equals \( h \).

Given these identifications, it is straightforward to specify the stochastic control problem associated with Sanov’s Theorem. Following the prescription given in Section 1.5, we describe first the dynamics and then the cost structure. Afterwards we give the representation formula.

**Dynamics.** The control that is applied at time \( j \in \{0, 1, \ldots, n-1\} \) is a stochastic kernel on \( \mathcal{S} \) given \( \mathcal{M}_{j/n}(\mathcal{S}) \) which we denote by \( \nu^n_j = \nu^n_j(dy|\mu) \). A sequence of controls \( \{\nu^n_j, j = 0, 1, \ldots, n-1\} \) is called an **admissible control sequence**. The controlled process is a Markov chain \( \{\bar{L}^n_j, j = 0, 1, \ldots, n\} \) with nonstationary transition probabilities. We define \( \bar{L}^n_0 = L^n_0 = 0 \). The transition probability functions of this Markov chain are given by

\[
\bar{P}\{\bar{L}^n_{j+1} \in A|\bar{L}^n_j = \mu\} \doteq p(\mu, A|\nu^n_j(\cdot|\mu)) = \int_A 1_A \left( \mu + \frac{1}{n} \delta_y \right) \nu^n_j(dy|\mu),
\]

where \( \mu \in \mathcal{M}_{j/n}(\mathcal{S}) \) and \( A \) is a Borel subset of \( \mathcal{M}_{(j+1)/n}(\mathcal{S}) \). This has the same form as the transition probability functions of the empirical subprobability measures

\[
L^n_j \doteq \frac{1}{n} \sum_{k=0}^{j-1} \delta_{X_k},
\]

except that the stochastic kernel \( \nu^n_j(\cdot|\mu) \) replaces the underlying probability measure \( \rho \) [see formula (2.2)]. From this we see that the sequence \( \{\bar{L}^n_j\} \) can be realized as the empirical subprobability measures of a sequence of controlled random variables \( \{\bar{X}^n_j, j = 0, 1, \ldots, n-1\} \). We set \( \bar{L}^n_0 = L^n_0 = 0 \) and define \( \bar{X}^n_j \) and \( \bar{L}^n_j \) recursively in the following order: the conditional distribution of \( \bar{X}^n_j \) given \( \bar{L}^n_0, \bar{L}^n_1, \ldots, \bar{L}^n_j \) is

\[
\bar{P}\{\bar{X}^n_j \in dy|\bar{L}^n_0, \bar{L}^n_1, \ldots, \bar{L}^n_j\} \doteq \nu^n_j(dy|\bar{L}^n_j); \tag{2.5}
\]

then

\[
\bar{L}^n_{j+1} \doteq \bar{L}^n_j + \frac{1}{n} \delta_{\bar{X}^n_j}.
\]

Thus for each \( j \in \{0, 1, \ldots, n\} \)

\[
\bar{L}^n_j = \frac{1}{n} \sum_{k=0}^{j-1} \delta_{\bar{X}^n_k}.
\]
2.3. THE REPRESENTATION FORMULA

We write $\bar{L}^n$ instead of $\bar{L}_n^n$. These measures will be referred to as the controlled empirical subprobability measures. For all $n \in \mathcal{N}$ the random variables $\bar{X}_j^n$ and $\bar{L}_j^n$ are defined on a common probability space, and we denote probability and expectation on this space by $\bar{P}$ and $\bar{E}$.

We now turn to the definition of the minimal cost functions, in which each control $\nu^n_j$ will give rise to a corresponding term in the running cost.

**Cost structure.** We have seen that for $\nu \in \mathcal{P}(\mathcal{S})$ the running cost equals $\frac{1}{n} R(\nu \| \rho)$ and that the terminal cost equals $h$. Hence for $i \in \{0, 1, \ldots, n\}$ and $\mu \in \mathcal{M}_{i/n}(\mathcal{S})$ we define the minimal cost functions

$$V^n(i, \mu) = \inf_{\{\nu^n_j\}} \bar{E}_{i, \mu} \left\{ \frac{1}{n} \sum_{j=i}^{n-1} R(\nu^n_j(\cdot | \bar{L}^n_j) \| \rho(\cdot)) + h(\bar{L}^n_i) \right\}.$$ 

In this formula the infimum is taken over all admissible control sequences $\{\nu^n_j\}$, $\{\bar{L}^n_i\}$ is the sequence of controlled empirical subprobability measures that are associated with a particular admissible control sequence $\{\nu^n_j\}$, and $\bar{E}_{i, \mu}$ denotes expectation conditioned on $\bar{L}^n_i = \mu$. When $i = 0$ and $\mu = 0$ we write $V^n$ instead of $V^n(0, 0)$ and $\bar{E}$ instead of $\bar{E}_{0,0}$. Thus

$$V^n = \inf_{\{\nu^n_j\}} \bar{E} \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot | \bar{L}^n_j) \| \rho(\cdot)) + h(\bar{L}^n) \right\}. \quad (2.6)$$

Our discussion of the dynamics and the cost structure of the stochastic control problem associated with Sanov’s Theorem reveals a number of general features that will recur throughout the book. For each random variable appearing in the dynamical description of the original process—in this case, each of the i.i.d. random variables $X_j$—we obtain a controlled random variable in the dynamical description of the controlled process—in this case $\bar{X}_j^n$. In this sense, the dynamics of the controlled process always have the same form as the dynamics of the original process. The conditional distributions of the controlled random variables are specified by the controls—in this case $\nu^n_j$. There exists a running cost at each time $j$ that is proportional to the relative entropy of the control at time $j$ with respect to the distribution of the original uncontrolled random variable at time $j$. The running cost measures the degree by which we have perturbed the dynamics from their original form. The sum of the running costs over all times $j$ gives the total running cost. Finally, the terminal cost equals the function $h$. We will see a similar pattern in the next chapter, when we consider Mogulskii’s Theorem, and in numerous other examples in Chapter 4.

Part (b) of Proposition 1.4.2 specifies the probability measure $\tilde{\nu}^n_i = \tilde{\nu}^n_i(\cdot | \mu)$ on $\mathcal{S}$ at which the infimum in (2.3) is uniquely attained. In the proof of the next theorem, we will need to know that $\tilde{\nu}^n_i(\cdot | \mu)$ is a stochastic kernel on $\mathcal{S}$ given $\mathcal{M}_{i/n}(\mathcal{S})$. This is a consequence of the explicit formula for $\tilde{\nu}^n_i(\cdot | \mu)$ given in the next lemma, which clearly implies that for every Borel subset $A$ of $\mathcal{S}$ the function mapping $\mu \in \mathcal{M}_{i/n}(\mathcal{S}) \mapsto \tilde{\nu}^n_i(A | \mu) \in [0, 1]$ is measurable.
**Lemma 2.3.1.** For $i \in \{0, 1, \ldots, n - 1\}$ and $\mu \in \mathcal{M}_{i/n}(\mathcal{S})$ the infimum in equation (2.3) is uniquely attained at the measure $\tilde{\nu}^n_i(dy) = \tilde{\nu}^n_i(dy|\mu)$ that assigns to a Borel subset $A$ of $\mathcal{S}$ the measure

$$
\tilde{\nu}^n_i(A|\mu) = \frac{1}{\int_{\mathcal{S}} \exp \left[ -n W^n \left( i + 1, \mu + \frac{1}{n} \delta_y \right) \right] \rho(dy)} \int_A \exp \left[ -n W^n \left( i + 1, \mu + \frac{1}{n} \delta_y \right) \right] \rho(dy).
$$

For each $i$ $\tilde{\nu}^n_i(dy|\mu)$ is a stochastic kernel on $\mathcal{S}$ given $\mathcal{M}_{i/n}(\mathcal{S})$.

We have now completed all the steps needed to derive a representation formula for the quantity $W^n$ in Sanov’s Theorem. This formula will be stated in the next theorem. The continuity of the function $h$ is needed in order to prove the Laplace principle, not in order to prove the representation formula, for which a bounded measurable function $h$ suffices.

**Theorem 2.3.2.** Let $h$ be a bounded measurable function mapping $\mathcal{P}(\mathcal{S})$ into $\mathbb{R}$. Then for all $n \in \mathbb{N}$

$$
W^n = -\frac{1}{n} \log E\{\exp[-n h(L^n)]\}
$$

equals the minimal cost function $V^n$ defined in equation (2.6).

**Proof.** As we have seen, the quantities $W^n(i, \mu)$ satisfy

$$
W^n(i, \mu) = \inf_{\nu \in \mathcal{P}(\mathcal{S})} \left\{ \frac{1}{n} R(\nu; \rho) + \int_{\mathcal{S}} W^n \left( i + 1, \mu + \frac{1}{n} \delta_y \right) \nu(dy) \right\}, \tag{2.7}
$$

which we have interpreted as a dynamic programming equation. In addition we have the terminal condition

$$
W^n(n, \mu) = h(\mu),
$$

which is also satisfied by the minimal cost function $V^n(n, \mu)$. The dynamic programming equation (2.7) and the terminal condition have a unique solution. Hence we will be able to conclude that each quantity $W^n(i, \mu)$ equals the corresponding minimal cost function $V^n(i, \mu)$, and thus that $W^n = V^n$, once it is confirmed that the minimal cost functions $V^n(i, \mu)$ satisfy the dynamic programming equation (2.7). This, in turn, will follow from Theorem 1.5.2 if we can verify the Attainment Condition. The latter requires that there exists a stochastic kernel $\nu = \tilde{\nu}^n_i(\cdot|\mu)$ at which the infimum in the dynamic programming equation is attained. But this has already been checked in Lemma 2.3.1, which in fact identifies the unique optimal control. The proof of the representation formula is complete.

We will use the representation formula to prove the Laplace principle by showing that for all bounded continuous functions $h$ mapping $\mathcal{P}(\mathcal{S})$ into $\mathbb{R}$

$$
limit_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(L^n)]\} = - \inf_{\nu \in \mathcal{P}(\mathcal{S})} \{R(\nu; \rho) + h(\nu)\}.
$$
2.4 Proof of the Laplace Principle Lower Bound

In order to prove (2.8), we will use the representation formula

\[ W^n = V^n = \inf_{\nu_j} \mathbb{E} \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R\left( \nu_j^n \cdot \| L^n_j \| \rho(\cdot) \right) + h(L^n) \right\}, \]

the infimum being taken over all admissible control sequences. We prove this by showing the upper limit

\[ \limsup_{n \to \infty} W^n \leq \inf_{\nu \in \mathcal{P}(S)} \{R(\nu \| \rho) + h(\nu)\}. \] (2.10)

Let \( \gamma \) be any probability measure on \( S \) and for each \( j \in \{0, 1, \ldots, n - 1\} \) consider the control \( \nu_j^n = \gamma \), which does not depend on the controlled measure \( L^n_j \). The representation formula yields the upper bound

\[ W^n = V^n \leq \mathbb{E}\left\{ R(\gamma \| \rho) + h(L^n) \right\}. \] (2.11)

With this choice of control, \( \bar{L}^n \) is now the empirical measure of i.i.d. random variables having the common distribution \( \gamma \). As we saw at the beginning of Section 2.2, we have the probability-1 weak convergence \( L^n \Longrightarrow \gamma \). Since \( h \) is a bounded continuous function mapping \( \mathcal{P}(S) \) into \( \mathbb{R} \), the Lebesgue Dominated Convergence Theorem yields

\[ \lim_{n \to \infty} \mathbb{E}\{h(L^n)\} = h(\gamma), \]

and from (2.11) we conclude that

\[ \limsup_{n \to \infty} W^n \leq R(\gamma \| \rho) + h(\gamma). \]

Since \( \gamma \in \mathcal{P}(S) \) is arbitrary, the upper limit (2.10), and thus the Laplace principle lower bound, follow. ■

We now turn to the proof of the Laplace principle upper bound (2.9), which is greatly facilitated by a judicious use of Jensen’s Inequality.
2.5 Proof of the Laplace Principle Upper Bound

We prove the Laplace upper bound by showing the lower limit

$$\liminf_{n \to \infty} W^n \geq \inf_{\nu \in \mathcal{P}(\mathcal{S})} \{ R(\nu \| \rho) + h(\nu) \}. \tag{2.12}$$

This will be carried out by applying Jensen’s Inequality to the sum of the relative entropies appearing in the representation formula obtained in Theorem 2.3.2, which states that

$$W^n = V^n \doteq \inf_{\{\nu_j^n\}} \mathbb{E} \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R \left( \nu_j^n(\cdot) \| \rho(\cdot) \right) + h(\bar{L}_n) \right\}. \tag{2.13}$$

In order to simplify the presentation, we will prove (2.12) only in the case $\mathcal{S}$ is compact. The case of a noncompact space $\mathcal{S}$ is subsumed under our treatment of the empirical measures of a Markov chain, which will be carried out in Chapter 8 with full details.

Let $\{Y_n, n \in \mathbb{N}\}$ and $Y$ be random variables that are defined on a common probability space $(\bar{\Omega}, \mathcal{F}, \bar{P})$ and take values in a Polish space $\mathcal{X}$. Expectation with respect to $\bar{P}$ is denoted by $\mathbb{E}$. We say that the sequence $\{Y_n, n \in \mathbb{N}\}$ converges in distribution to $Y$, and write $Y_n \xrightarrow{D} Y$, if the sequence of distributions of $\{Y_n, n \in \mathbb{N}\}$ converges weakly to the distribution of $Y$. In other words, $Y_n \xrightarrow{D} Y$ if for any bounded continuous function $g$ mapping $\mathcal{X}$ into $\mathbb{R}$

$$\lim_{n \to \infty} \mathbb{E} \{ g(Y_n) \} = \mathbb{E} \{ g(Y) \}.$$ 

In this book we will make numerous applications of the Skorohod Representation Theorem [Theorem A.3.9] when dealing with random variables that converge in distribution. According to this theorem, if $Y_n \xrightarrow{D} Y$, then there exists a probability space on which are defined $\mathcal{X}$-valued random variables $\{\hat{Y}_n, n \in \mathbb{N}\}$ and $\hat{Y}$ with the following properties: each $\hat{Y}_n$ has the same distribution as $Y_n$, $\hat{Y}$ has the same distribution as $Y$, and $\hat{Y}_n$ converges to $\hat{Y}$ w.p.1. In applying the Skorohod Representation Theorem our convention will be to retain the original labels for the random variables—in the present case, $Y_n$ and $Y$ instead of $\hat{Y}_n$ and $\hat{Y}$—and to retain $(\bar{\Omega}, \mathcal{F}, \bar{P})$ and $\mathbb{E}$ as the new probability space and the expectation operator on the new space. We will often appeal to the theorem using the following terminology. If $Y_n \xrightarrow{D} Y$, then there exists a probability space such that the convergence occurs w.p.1 on this space. The Skorohod Representation Theorem is an extremely useful device if one is interested, as we will often be, in calculating limits of expectations of functions of $Y_n$.

We now turn to the proof of (2.12). It suffices to show that every subsequence of $\{W^n, n \in \mathbb{N}\}$ has a subsequence such that

$$\liminf_{n \to \infty} W^n \geq \inf_{\nu \in \mathcal{P}(\mathcal{S})} \{ R(\nu \| \rho) + h(\nu) \}. \tag{2.14}$$

We will work with a fixed subsequence of $\{W^n\}$ for the remainder of the proof. Since $R(\cdot \| \rho)$ is convex, Jensen’s Inequality applied to the sum in equation (2.13) yields the lower bound

$$W^n \geq \inf_{\{\nu_j^n\}} \mathbb{E} \left\{ R \left( \frac{1}{n} \sum_{j=0}^{n-1} \nu_j^n(\cdot) \| \rho(\cdot) \right) + h(\bar{L}_n) \right\}. \tag{2.15}$$
2.5. PROOF OF UPPER BOUND

Now let $\varepsilon > 0$ be given. For each $n \in \mathbb{N}$ there exists an admissible control sequence $\{\nu^n_j\}$ and a corresponding quantity

$$\alpha^n(dy) := \frac{1}{n} \sum_{j=0}^{n-1} \nu^n_j(dy|\bar{L}^n_j)$$

(2.16)

such that

$$W^n + \varepsilon \geq \mathbb{E}\{R(\alpha^n|\rho) + h(\bar{L}^n)\}.$$  

(2.17)

Since each quantity $\nu^n_j(\cdot|\bar{L}^n_j)$ defines a stochastic kernel on $\mathcal{S}$ given $\Omega$, $\alpha^n$ also defines such a stochastic kernel and thus is a random variable on $\Omega$ [Theorem A.5.2]. We call $\alpha^n$ an admissible control measure. For each $\omega (\alpha^n, \bar{L}^n)$ takes values in $\mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S})$. Since $\mathcal{S}$ is assumed to be compact, Prohorov’s Theorem implies that $\mathcal{P}(\mathcal{S})$ and thus $\mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S})$ are compact [Corollary A.3.16]. Thus, given any subsequence of $\{(\alpha^n, \bar{L}^n), n \in \mathbb{N}\}$, there exists a subsequence that converges in distribution to a limit $(\alpha, \bar{L})$ which takes values in $\mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S})$. Part (b) of the following lemma, which is the heart of the proof of the lower limit (2.12), states that $\alpha$ equals $\bar{L}$ w.p.1. Lemma 2.5.1 is a special case of a general result that will be proved in Lemma 8.2.7.

Before stating the lemma, we make two observations. First, a crucial simplifying step in the proof of the lower limit (2.12) is the existence of a convergent subsequence of $\{(\alpha^n, \bar{L}^n)\}$. It is a consequence of the assumption that $\mathcal{S}$ is compact. When we weaken this assumption in the treatment of the empirical measures of a Markov chain in Chapter 8, the proof of the existence of a corresponding convergent subsequence will require some work. Second, while the application of Jensen’s Inequality to derive (2.15) is a useful trick that simplifies the analysis, we must be careful in applying it elsewhere. Indeed, as we will point out in the last section of Chapter 3, applying it analogously in the proof of the Laplace principle upper bound in Mogulskii’s Theorem would yield an incorrect answer. The a posteriori justification for using Jensen’s Inequality in the present context, as elsewhere in this book, is the equality of the upper and lower bounds appearing in the Laplace principle. However, one can usually anticipate when its use can be justified.

**Lemma 2.5.1.** We assume that $\mathcal{S}$ is compact. The following conclusions hold.

(a) For any bounded measurable function $g$ mapping $\mathcal{S}$ into $\mathbb{R}$ and any $\varepsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}\left\{ \left| \int_{\mathcal{S}} g d\bar{L}^n - \int_{\mathcal{S}} g d\alpha^n \right| \geq \varepsilon \right\} = 0.$$  

(b) Given any subsequence of $\{(\alpha^n, \bar{L}^n), n \in \mathbb{N}\}$ there exists a subsequence and a random variable $\bar{L}$ such that $\bar{L}$ takes values in $\mathcal{P}(\mathcal{S})$ and the subsequence converges in distribution to $(\alpha, \bar{L})$.

**Proof.** (a) For each $j \in \{0, 1, \ldots, n - 1\}$ we denote by $\mathcal{F}^n_j$ the $\sigma$-field generated by $\bar{L}^n_0, \bar{L}^n_1, \ldots, \bar{L}^n_j$. Since $\nu^n_j(dy|\mu)$ is a stochastic kernel on $\mathcal{S}$ given $\mathcal{M}_{j,n}(\mathcal{S}), \ (2.5)$ implies that $\nu^n_j(\cdot|\bar{L}^n_j)$ is a regular conditional distribution for $X^n_j$ given $\mathcal{F}^n_j$. Hence for each bounded measurable function $g$ mapping $\mathcal{S}$ into $\mathbb{R}$

$$\mathbb{E}\left\{ g(X^n_j) - \int_{\mathcal{S}} g(y) \nu^n_j(dy|\bar{L}^n_j) \middle| \mathcal{F}^n_j \right\} = 0$$
\[ \mathbb{P}\text{-a.s. This implies that the sequence} \]
\[
\left\{ g(X_i^n) - \int_S g(y) \nu^n_j(dy | L^n_j), j = 0, 1, \ldots, n - 1 \right\}
\]
forms a martingale difference sequence with respect to \( \{ \mathcal{F}_j^n, j = 0, 1, \ldots, n - 1 \} \), and thus for \( 0 \leq i \neq \ell \leq n - 1 \)
\[
\mathbb{E} \left\{ \left( g(X_i^n) - \int_S g(y) \nu^n_j(dy | L^n_j) \right) \left( g(X_\ell^n) - \int_S g(y) \nu^n_\ell(dy | L^n_\ell) \right) \right\} = 0.
\]
Since
\[
\int_S g dL^n - \int_S g d\alpha^n = \frac{1}{n} \sum_{j=0}^{n-1} \left( g(X_j^n) - \int_S g(y) \nu^n_j(dy | L^n_j) \right),
\]
for any \( \varepsilon > 0 \) and all \( n \in \mathbb{N} \)
\[
\mathbb{P}\left\{ \left| \int_S g dL^n - \int_S g d\alpha^n \right| \geq \varepsilon \right\} \leq \frac{1}{\varepsilon^2} \mathbb{E} \left\{ \frac{1}{n^2} \sum_{j=0}^{n-1} \left( g(X_j^n) - \int_S g(y) \nu^n_j(dy | L^n_j) \right)^2 \right\} \leq \frac{4 \| g \|_\infty^2}{n \varepsilon^2}.
\]
The convergence in probability asserted in part (a) now follows from this display.

(b) According to Theorem A.6.1, \( \mathcal{S} \) admits an equivalent metric \( m(x, y) \) with the property that the space of bounded, uniformly continuous functions \( \mathcal{U}_b(\mathcal{X}, m) \) is separable. Let \( \Xi \) be a countable dense subset of \( \mathcal{U}_b(\mathcal{S}, m) \). Part (a) of the present lemma implies that for each \( g \in \Xi \) the sequence
\[
\left\{ \int_S g dL^n - \int_S g d\alpha^n, n \in \mathbb{N} \right\}
\]
converges to 0 in probability, and hence in distribution, with respect to \( \mathbb{P} [\text{Theorem A.3.7}] \). The compactness of \( \mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S}) \) implies that any subsequence of \( \{(\alpha^n, L^n), n \in \mathbb{N} \} \) has a subsubsequence that converges in distribution to a limit \( (\alpha, \bar{L}) \) which takes values in \( \mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S}) \). We now make our first application of the Skorohod Representation Theorem. This guarantees the existence of a probability space such that w.p.1 on this space the following three limits are valid for all \( g \in \Xi \):
\[
\lim_{n \to \infty} \left( \int_S g dL^n - \int_S g d\alpha^n \right) = 0,
\]
2.5. PROOF OF UPPER BOUND

\[ \lim_{n \to \infty} \int_S g d\tilde{L}^n = \int_S g d\tilde{L}, \quad \text{and} \quad \lim_{n \to \infty} \int_S g d\alpha^n = \int_S g d\alpha. \]

Combining these limits shows that w.p.1, for all \( g \in \Xi \) and hence all \( g \in \mathcal{U}_b(S, m) \),

\[ \int_S g d\alpha = \int_S g d\tilde{L}. \]

For probability measures \( \mu \) and \( \nu \) on \( S \) the equality \( \int_S g d\mu = \int_S g d\nu \) for all \( g \in \mathcal{U}_b(S, m) \) implies that \( \mu = \nu \) [Theorem A.2.2 (b)]. It follows that \( \alpha = \tilde{L} \) w.p.1. The proof of part (b) is complete. ■

It is not hard to motivate part (b) of the lemma by considering the case where all the stochastic kernels \( \nu^n_j \) equal a fixed probability measure \( \gamma \) on \( S \). Then for each \( n \in \mathbb{N} \) \( \alpha^n \) equals \( \gamma \), and so \( \alpha \) equals \( \gamma \). Furthermore, the random measures \( \tilde{L}^n \) are the empirical measures of the random variables \( \{\tilde{X}_j, j = 0, 1, \ldots, n - 1\} \), which in this case are i.i.d. with common distribution \( \gamma \). As we saw in Section 2.2, we have the probability-1 weak convergence \( \tilde{L}^n \Rightarrow \gamma \), and so in part (b) of Lemma 2.5.1 \( (\alpha^n, \tilde{L}^n) \overset{d}{\rightarrow} (\gamma, \gamma) \), as claimed.

A martingale argument is used in the proof of Lemma 2.5.1 in order to deal with the extra dependencies that are absent in the simple case just considered. A similar situation arises, for example, when one generalizes Kolmogorov’s law of large numbers to martingale difference sequences.

We return to the main argument, indexing by \( n \in \mathbb{N} \) the subsequence along which the convergence in distribution \( (\alpha^n, \tilde{L}^n) \overset{d}{\rightarrow} (\tilde{L}, \bar{L}) \) in part (b) of Lemma 2.5.1 is valid. According to equation (2.17)

\[ W^n + \varepsilon \geq \mathbb{E}\{R(\alpha^n \| \rho) + h(\tilde{L}^n)\}. \]

The Skorohod Representation Theorem allows us to assume that w.p.1 the subsequence \( (\alpha^n, \tilde{L}^n) \) converges to \( (\tilde{L}, \bar{L}) \) on an appropriate probability space. Our aim is now to study the limit as \( n \to \infty \) of the two terms on the right hand side of the last display. This is completely straightforward. Indeed, since \( R(\cdot \| \rho) \) is a nonnegative, lower semicontinuous function on \( \mathcal{P}(S) \) and \( h \) is a bounded continuous function on \( \mathcal{P}(S) \), the probability-1 convergence \( (\alpha^n, \tilde{L}^n) \rightarrow (\tilde{L}, \bar{L}) \) implies that

\[ \lim_{n \to \infty} \{R(\alpha^n \| \rho) + h(\tilde{L}^n)\} \geq R(\tilde{L} \| \rho) + h(\bar{L}). \]

Hence by Fatou’s Lemma

\[ \liminf_{n \to \infty} W^n + \varepsilon \geq \liminf_{n \to \infty} \mathbb{E}\{R(\alpha^n \| \rho) + h(\tilde{L}^n)\} \geq \mathbb{E}\{R(\tilde{L} \| \rho) + h(\bar{L})\} \geq \inf_{\nu \in \mathcal{P}(S)} \{R(\nu \| \rho) + h(\nu)\}. \]

Since \( \varepsilon > 0 \) is arbitrary, we have proved that every subsequence of the original sequence \( \{W^n, n \in \mathbb{N}\} \) has a subsubsequence satisfying the lower limit (2.14). An argument by contradiction applied to an arbitrary subsequence shows that the entire sequence \( \{W^n, n \in \mathbb{N}\} \) satisfies this same lower limit. This completes the proof of the Laplace principle upper bound in the case of a compact space \( S \). ■
It is worthwhile to identify two key issues that arise in the analysis of every Laplace principle. In each case we will need to understand first the conditions that guarantee compactness of controls and second the convergence properties of the joint distributions of controls and controlled processes. In the present example the compactness is automatic because $S$ is assumed to be compact. The convergence properties are the subject of Lemma 2.5.1.

With the proof of the Laplace principle upper bound, the proof of Sanov’s Theorem is done. In the next chapter we will prove the Laplace principle for a sequence of continuous-time random walk models constructed from i.i.d. random variables. We will carry this out by adapting to the new situation the procedure that has just been used to prove Sanov’s Theorem.
Chapter 3

A Second Example: Mogulskii’s Theorem

3.1 Introduction

In the previous chapter we introduced the basic ideas of the weak convergence approach in the context of proving Sanov’s Theorem. In the present chapter we continue our introduction of the weak convergence approach by analyzing a large deviation problem of an entirely different type. As before, the proof of the Laplace principle is based on a representation formula. The Laplace principle to be proved in the present chapter will be greatly generalized in Chapters 6 and 7.

Following [28], we call the large deviation theorem in this chapter Mogulskii’s Theorem. It is proved in the paper [72]. A more recent proof, which employs different methods, is given in Section 5.1 of [28] while a large deviation theorem for the same process with a different scaling is proved in [87]. Mogulskii’s Theorem is closely related to Schilder’s Theorem [Example 1.1.3], which is a large deviation theorem for Brownian motion, and results in [69], which analyzes a class of processes with stationary independent increments.

Let \( \{X^n, n \in \mathbb{N}\} \) be a sequence of random variables taking values in a Polish space \( \mathcal{X} \). As we recall, the proof of the Laplace principle entails evaluating the limit as \( n \to \infty \) of \( \frac{1}{n} \log E\{\exp[-n h(X^n)]\} \), where \( h \) is any bounded continuous function mapping \( \mathcal{X} \) into \( \mathbb{R} \). In general, we work with

\[
W^n = -\frac{1}{n} \log E\{\exp[-n h(X^n)]\}.
\]

An obvious issue that arises when one tries to prove the Laplace principle in a new setting is to determine the form of the rate function. In the previous chapter the rate function was given without motivation in the statement of Sanov’s Theorem. In the present chapter we will proceed differently. Namely, we will show how the weak convergence approach leads one naturally to guess the form of the rate function in Mogulskii’s Theorem. This will be carried out in Step 2 of the following systematic procedure for applying the weak convergence approach to any given large deviation problem. We will follow this procedure throughout the book.
1. Obtain a representation formula for $W^n$ in terms of the minimal cost function $V^n$
of an associated stochastic control problem.

2. Establish compactness and limit properties of the control measures and the con-
trolled processes. Then use these together with weak convergence methods to prove
the Laplace principle upper bound. Because of the first minus sign in the definition
of $W^n$, this entails evaluating the lower limit $\liminf_{n \to \infty} W^n = \liminf_{n \to \infty} V^n$. If
this lower limit is expressed in the form required by the Laplace principle—namely,
as an infimum over the space $\mathcal{X}$ of a sum $I + h$—then one is handed the candidate
$I$ for the rate function in the Laplace principle.

3. Verify that the function $I$ in Step 2 satisfies the properties of a rate function.

4. Prove the Laplace principle lower bound by evaluating the upper limit
\[
\limsup_{n \to \infty} W^n \leq \inf_{x \in \mathcal{X}} \{I(x) + h(x)\}.
\]

This is carried out as follows. Given $\varepsilon > 0$, choose a point $x^* \in \mathcal{X}$ such that
$I(x^*) + h(x^*)$ is within $\varepsilon$ of the infimum of $I + h$ over $\mathcal{X}$. In terms of $x^*$, then
construct a sequence of controls to be used in the representation formula for $W^n$
and prove the upper limit
\[
\limsup_{n \to \infty} W^n \leq I(x^*) + h(x^*) + \varepsilon.
\]

The right hand side does not exceed $\inf_{x \in \mathcal{X}} \{I(x) + h(x)\} + 2\varepsilon$, and since $\varepsilon > 0$ is
arbitrary, we obtain the Laplace principle lower bound.

In the next section we will use a dynamic programming argument to carry out Step
1 for the stochastic processes to be considered in Mogulskii’s Theorem. Step 2 will be
done in Section 3.3. With the candidate for the rate function $I$ in hand, we formulate
Mogulskii’s Theorem and carry out Step 4 in Section 3.4. Step 3 on the verification that $I$
is a rate function is omitted. That fact is proved in Proposition 6.2.4, which covers a much
more general situation. As a bonus, Mogulskii’s Theorem and the contraction principle
will yield Cramér’s Theorem for the sample means of i.i.d. random vectors (Section 3.5).

Before leaving Sanov’s Theorem and Mogulskii’s Theorem for the much more complica-
ted models of later chapters, we pause in the last section of this chapter to review the
methods of proof in these two elementary examples. As in the case of Sanov’s Theorem,
the derivation of the representation formula for $W^n$ in Step 1 of the proof of Mogulskii’s
Theorem can also be done by a purely measure theoretic argument. However, in con-
trast to the derivation in the case of Sanov’s Theorem, this measure theoretic argument
involves some complications that, in our opinion, render it a less satisfactory approach
than the derivation via dynamic programming. These complications arise in essentially
all the other large deviation problems that we will consider. These and related issues are
discussed in Section B.3.
3.2 The Representation Formula

We start by defining the random walk model for which we want to prove the Laplace principle. Given a probability measure \( \rho \) on \( \mathbb{R}^d \), let \( \{v_j, j \in \mathbb{N}_0\} \) be a sequence of i.i.d. random vectors that are defined on a probability space \((\Omega, \mathcal{F}, \mathbb{P})\), take values in \( \mathbb{R}^d \), and have the common distribution \( \rho \). For \( n \in \mathbb{N} \) and \( j \in \{0, 1, \ldots, n-1\} \), we define random vectors \( X^n_0 \triangleq 0 \) and

\[
X^n_j \triangleq \frac{1}{n} \sum_{k=0}^{j-1} v_k. 
\]

We are interested in the Laplace principle for the stochastic process \( \{X^n(t), t \in [0, 1]\} \) defined by

\[
X^n(t) \triangleq X^n_j + \left( t - \frac{j}{n} \right) v_j \quad \text{for} \ t \in [j/n, (j+1)/n], \ j = 0, 1, \ldots, n-1. 
\]

The random walk \( X^n \triangleq \{X^n(t), t \in [0, 1]\} \) is the piecewise linear interpolation of the discrete-time process \( \{X^n_j, j = 0, 1, \ldots, n\} \). The paths of \( X^n \) lie in \( C([0, 1] : \mathbb{R}^d) \), which is the space of continuous functions mapping \([0, 1]\) into \( \mathbb{R}^d \). The interval \([0, 1]\) has been chosen for notational convenience only, analogous results being obtainable for any interval of the form \([0, T], T \in (0, \infty)\). We denote by \( \| \cdot \| \) the Euclidean norm on \( \mathbb{R}^d \). \( C([0, 1] : \mathbb{R}^d) \) is a Banach space with respect to the supremum norm, for which \( \varphi \in C([0, 1] : \mathbb{R}^d) \) is defined by

\[
\| \varphi \|_\infty \triangleq \sup_{t \in [0, 1]} \| \varphi(t) \|.
\]

In order to simplify the presentation, we will assume that the distribution \( \rho \) of the i.i.d. random variables \( \{v_j, j \in \mathbb{N}_0\} \) is supported on a compact subset \( K \) of \( \mathbb{R}^d \). The case of noncompact support is covered by the results in Chapter 6.

Let \( h \) be any bounded continuous function mapping \( C([0, 1] : \mathbb{R}^d) \) into \( \mathbb{R} \). The first step in proving the large deviation principle is the derivation of a representation formula for

\[
W^n \triangleq -\frac{1}{n} \log E\{\exp[-n h(X^n)]\}.
\]

We will call \( \varphi \in C([0, 1] : \mathbb{R}^d) \) piecewise linear if it is linear on intervals of the form \([j/n, (j+1)/n], j = 0, 1, \ldots, n-1\). Since the sample paths of the random walk \( X^n \) are piecewise linear and satisfy \( X^n(0) = 0 \), the value of the function \( h \) appearing in the definition of \( W^n \) is irrelevant for any \( \varphi \in C([0, 1] : \mathbb{R}^d) \) that is not piecewise linear or does not satisfy \( \varphi(0) = 0 \). Furthermore, for any piecewise linear \( \varphi \) satisfying \( \varphi(0) = 0 \) \( h(\varphi) \) is determined by \( \varphi(1/n), \varphi(2/n), \ldots, \varphi(n/n) \). For any such \( \varphi \) we define a bounded measurable function \( \alpha_h \) mapping \( (\mathbb{R}^d)^n \) into \( \mathbb{R} \) by

\[
\alpha_h(\varphi(1/n), \varphi(2/n), \ldots, \varphi(n/n)) \triangleq h(\varphi). 
\]

In particular,

\[
\alpha_h(X^n(1/n), X^n(2/n), \ldots, X^n(n/n)) = \alpha_h(X^n_1, X^n_2, \ldots, X^n_n) = h(X^n).
\]
The last display directs our attention back to the discrete-time process \( \{X^n_i, i = 0, 1, \ldots, n\} \), which evolves according to \( X^n_0 = 0 \in \mathbb{R}^d \) and

\[ X^n_{i+1} = X^n_i + \frac{1}{n}v_i. \]

Since the random vectors \( v_i \) are i.i.d. with common distribution \( \rho \), the random vectors \( X^n_i \) form a Markov chain on \( \mathbb{R}^d \). For \( A \) a Borel subset of \( \mathbb{R}^d \), the transition probability function of this Markov chain is given by

\[ P\{X^n_{i+1} \in A | X^n_i = x\} = \int_{\mathbb{R}^d} 1_A \left( x + \frac{1}{n}y \right) \rho(dy). \quad (3.4) \]

In order to derive the representation formula for \( W^n \), we must introduce an appropriate sequence of functions for each starting time \( i \in \{0, 1, \ldots, n\} \). These functions will be defined so that an underlying Markov property can be exploited to obtain a recursive equation relating them. In general \( h(X^n) \) is a function not just of \( X^n \) but of the entire trajectory \( \{X^n(t), t \in [0, 1]\} \) or, equivalently, of the entire discrete-time process \( \{X^n_j, j = 0, 1, \ldots, n\} \). Because of this we must define the function at time \( i \) by conditioning on the entire past of the process up to time \( i \). With this in mind, let us proceed by defining for any starting time \( i \in \{0, 1, \ldots, n\} \) and any points \( x_1, x_2, \ldots, x_i \) in \( \mathbb{R}^d \)

\[ W^n(i, \{x_1, x_2, \ldots, x_i\}) = -\frac{1}{n} \log E\{\exp[-n h(X^n) || X^n_1 = x_1, X^n_2 = x_2, \ldots, X^n_i = x_i]\}. \]

\( W^n(0, \emptyset) \) equals

\[ W^n = -\frac{1}{n} \log E\{\exp[-n h(X^n)]\} \]

while \( W^n(n, \{x_1, x_2, \ldots, x_n\}) \) equals \( h(X^n) = \alpha_h(x_1, x_2, \ldots, x_n) \). For \( i \in \{0, 1, \ldots, n\} \) each \( W^n(i, \cdot) \) is a bounded measurable function on \( (\mathbb{R}^d)^i \).

As in the previous paragraph, we will soon be introducing other sequences of quantities indexed by a nonnegative integer \( i \). When \( i = 0 \), the definitions of these quantities will differ in small ways from the definitions when \( i \geq 1 \), and we will not always pause to point this out. In this context, when \( i = 0 \) we set \( (\mathbb{R}^d)^0 = \emptyset \), and we recall the convention of Section 1.4 that a stochastic kernel on \( \mathbb{R}^d \) given \( (\mathbb{R}^d)^0 = \emptyset \) is a probability measure on \( \mathbb{R}^d \).

We next derive an equation relating the functions \( W^n(i, \cdot) \) which will later be interpreted as the dynamic programming equation of an associated stochastic control problem. Taking \( i \in \{0, 1, \ldots, n - 1\} \), we use equation (3.3) and the Markov property of the discrete-time process \( \{X^n_i\} \) to write

\[
\exp[-n W^n(i, \{x_1, \ldots, x_i\})]
= E\{\exp[-n \alpha_h(X^n_1, \ldots, X^n_i) || X^n_1 = x_1, \ldots, X^n_i = x_i]\}
= E\left\{ E\{\exp[-n \alpha_h(X^n_1, \ldots, X^n_i) || X^n_1, \ldots, X^n_i, X^n_{i+1}] \} || X^n_1 = x_1, \ldots, X^n_i = x_i \right\}
= E\{\exp[-n W^n(i + 1, \{X^n_1, \ldots, X^n_i, X^n_{i+1}\}) || X^n_1 = x_1, \ldots, X^n_i = x_i]\}
= \int_{\mathbb{R}^d} \exp\left[-n W^n\left(i + 1, \{x_1, \ldots, x_i, x_i + \frac{1}{n}y\}\right)\right] \rho(dy).
\]
3.2. THE REPRESENTATION FORMULA

Thus

\[ W^n(i, \{x_1, \ldots, x_i\}) = -\frac{1}{n} \log \int_{\mathbb{R}^d} \exp \left[-n W^n(i + 1, \{x_1, \ldots, x_i, x_i + \frac{1}{n} y\})\right] \rho(dy). \]

We now apply part (a) of Proposition 1.4.2 to the space \( \mathcal{V} = \mathbb{R}^d \), the function \( k(y) = n \cdot W^n(i + 1, \{x_1, \ldots, x_i, x_i + \frac{1}{n} y\}) \), and the measure \( \theta = \rho \), obtaining the variational formula

\[ W^n(i, \{x_1, \ldots, x_i\}) = \inf_{\nu \in \mathcal{P}(\mathbb{R}^d)} \left\{ \frac{1}{n} I(\nu \| \rho) + \int_{\mathbb{R}^d} W^n(i + 1, \{x_1, \ldots, x_i, x_i + \frac{1}{n} y\}) \nu(dy) \right\}. \]  

(3.5)

We also have the terminal condition

\[ W^n(n, \{x_1, \ldots, x_n\}) = \alpha_h(x_1, \ldots, x_n). \]  

(3.6)

As in Section 1.5, we now interpret equations (3.5) and (3.6) as the dynamic programming equation and the terminal condition, respectively, of a particular stochastic control problem. Following the same procedure as in the last chapter, we can easily read off the components of the control problem from these equations. These components are the spaces \( \mathcal{Z} \), \( \mathcal{U} \), and \( \mathcal{P}(\mathbb{R}^d) \) and the functions \( p_i(z, d\xi^i|u) \), \( c_i \), and \( f \).

The state space \( \mathcal{Z} \) is the space on which \( W^n(i, \cdot) \) is defined and thus equals \( (\mathbb{R}^d)^i \). The control space \( \mathcal{U} \), which is the space over which the infimum in equation (3.5) is taken, equals \( \mathcal{P}(\mathbb{R}^d) \). For \( i \in \{0, 1, \ldots, n-1\} \) each control \( \nu^i \) is a stochastic kernel on \( \mathbb{R}^d \) given \((\mathbb{R}^d)^i \). Each of the running costs is defined by the same formula

\[ c_i(\nu) = \frac{1}{n} R(\nu \| \rho) \quad \text{for} \quad \nu \in \mathcal{P}(\mathbb{R}^d) \]

while the terminal cost \( f \) equals the function \( \alpha_h \) defined in formula (3.3).

The final component of the control problem is the sequence of controlled transition probability functions. Although there is no difficulty in determining these functions from equation (3.5), their somewhat complicated form seems to obscure matters rather than illuminate them. Instead, we will jump ahead and specify the dynamics of the control problem together with the cost structure. After that we will state the representation formula.

**Dynamics.** The discrete-time process \( \{X^n_j, j = 0, 1, \ldots, n\} \) evolves according to \( X^n_0 = 0 \in \mathbb{R}^d \) and

\[ X^n_{j+1} = X^n_j + \frac{1}{n} v_j. \]  

(3.7)

The random vectors \( v_j \) are i.i.d. with common distribution \( \rho \). Exactly as in Section 2.4, we obtain the dynamics of the associated controlled process by replacing each of the random vectors \( v_j \) by an appropriate controlled random vector. The conditional distributions of
the controlled random vectors will be specified by a sequence \( \{ \nu_j^n, j = 0, 1, \ldots, n - 1 \} \), where each quantity \( \nu_j^n = \nu_j^n(dy|x_1, x_2, \ldots, x_j) \) is a stochastic kernel on \( \mathbb{R}^d \) given \( (\mathbb{R}^d)^j \).

We call such a sequence an admissible control sequence. Each control \( \nu_j^n \) will give rise to a corresponding term in the running cost in the definition of the minimal cost function.

The controlled process is a nonstationary Markov chain \( \{ \tilde{Z}_j^n, j = 1, 2, \ldots, n \} \) with the respective state spaces \( \{(\mathbb{R}^d)^j, j = 1, 2, \ldots, n \} \). Suppose that \( \nu_j^n \) is the control applied at time \( j \) and that \( \tilde{X}_j^n, \tilde{X}_0^n, \ldots, \tilde{X}_j^n \) are random vectors taking values in \( \mathbb{R}^d \) such that \( (\tilde{X}_1^n, \tilde{X}_2^n, \ldots, \tilde{X}_j^n) \) gives the state \( \tilde{Z}_j^n \) at time \( j \). Then according to formula (3.5) \( \tilde{Z}_{j+1}^n \) equals \( (\tilde{X}_1^n, \tilde{X}_2^n, \ldots, \tilde{X}_j^n, \tilde{X}_{j+1}^n) \), where

\[
\tilde{X}_{j+1}^n = \tilde{X}_j^n + \frac{1}{n} \tilde{Y}_j^n, \quad \tilde{X}_0^n = X_0^n = 0 \in \mathbb{R}^d,
\]

(3.8)

and where \( \tilde{Y}_j^n \) has the conditional distribution \( \nu_j^n \). In other words, for \( j \in \{1, 2, \ldots, n - 1\} \) the conditional distribution of \( \tilde{Y}_j^n \) is given by

\[
P\{ \tilde{Y}_j^n \in dy | \tilde{X}_1^n, \tilde{X}_2^n, \ldots, \tilde{X}_j^n \} = \nu_j^n(dy | \tilde{X}_1^n, \tilde{X}_2^n, \ldots, \tilde{X}_j^n)
\]

(3.9)

while the distribution of \( \tilde{Y}_0^n \) is given by

\[
P\{ \tilde{Y}_0^n \in dy \} = \nu_0^n(dy).
\]

(3.10)

The random vectors \( \tilde{X}_j^n \) and \( \tilde{Y}_j^n \) are defined recursively in the following order: \( \tilde{X}_0^n, \tilde{Y}_0^n, \tilde{X}_1^n, \tilde{Y}_1^n, \ldots, \tilde{Y}_n^n, \tilde{X}_n^n \). For all \( n \in \mathbb{N} \) the controlled random vectors \( \tilde{X}_j^n \) and \( \tilde{Y}_j^n \) are defined on a common probability space \( (\Omega, \mathcal{F}, P) \), and expectation on this space is denoted by \( \tilde{E} \). The first equation in (3.8), which defines the controlled random vectors, has the same form as equation (3.7), which defines the original process.

Cost structure. The running cost equals \( \frac{1}{n} R(\nu || \rho) \) and the terminal cost equals \( \alpha_h \). Hence for \( i \in \{0, 1, \ldots, n\} \) and \( (x_1, x_2, \ldots, x_i) \in (\mathbb{R}^d)^i \) we define the minimal cost functions

\[
V^n(i, \{x_1, x_2, \ldots, x_i\}) = \inf_{\{\nu_j^n\}} \tilde{E}_{i, \{x_1, x_2, \ldots, x_i\}} \left\{ \frac{1}{n} \sum_{j=1}^{n-1} R(\nu_j^n(\cdot) || \rho(\cdot)) + \alpha_h(\tilde{X}_1^n, \tilde{X}_2^n, \ldots, \tilde{X}_n^n) \right\}.
\]

In this formula \( \nu_j^n(\cdot) = \nu_j^n(\cdot | \tilde{X}_1^n, \tilde{X}_2^n, \ldots, \tilde{X}_j^n) \), the infimum is taken over all admissible control sequences \( \{\nu_j^n\}, \{\tilde{X}_j^n\} \) is the sequence of controlled random vectors that are associated with a particular admissible control sequence \( \{\nu_j^n\} \), and \( \tilde{E}_{i, \{x_1, x_2, \ldots, x_i\}} \) denotes expectation conditioned on \( \tilde{X}_1^n = x_1, \tilde{X}_2^n = x_2, \ldots, \tilde{X}_i^n = x_i \). When \( i = 0 \) we write \( V^n \) instead of \( V^n(0, \emptyset) \) and \( \tilde{E} \) instead of \( \tilde{E}_{0, \emptyset} \). Thus

\[
V^n = \inf_{\{\nu_j^n\}} \tilde{E} \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu_j^n(\cdot) || \rho(\cdot)) + \alpha_h(\tilde{X}_1^n, \tilde{X}_2^n, \ldots, \tilde{X}_n^n) \right\}.
\]

The calculation of the asymptotic behavior of \( V^n \) depends on taking limits in \( n \) with respect to admissible control sequences \( \{\nu_j^n, j = 0, 1, \ldots, n - 1\} \). In order to carry this out,
it is convenient to place the admissible control sequences in a space that is independent of $n$. Given $n \in \mathbb{N}$, an admissible control sequence $\{\nu^*_j\}$, and $t \in [0, 1]$, we define the stochastic kernel

$$
\nu^n(dy|t) = \begin{cases} 
\nu^n_j(dy) & \text{if } t \in [j/n, (j+1)/n), j = 0, 1, \ldots, n-2 \\
\nu^n_{n-1}(dy) & \text{if } t \in [(n-1)/n, 1],
\end{cases}
$$

(3.11)

where $\nu^n_j(dy) = \nu^n_j(dy|X^n_1, X^n_2, \ldots, X^n_j)$. We then define a random probability measure $\nu^n$ on $\mathbb{R}^d \times [0, 1]$ by

$$
\nu^n(A \times B) = \int_B \nu^n(A|t) \, dt
$$

(3.12)

for Borel subsets $A$ of $\mathbb{R}^d$ and $B$ of $[0, 1]$. The term admissible control measure will be used to refer to $\nu^n$, and the definition of this measure will be summarized as $\nu^n(dy \times dt) = \nu^n(dy|t) \otimes dt$. As we will see in Lemma 3.3.2, the definition that we adopt here will allow us to relate the limits of the admissible control measures and the limits of the controlled processes.

The study of the asymptotic behavior of $V^n$ is also facilitated by introducing the piecewise linear process $\tilde{X}^n = \{\tilde{X}^n(t), t \in [0, 1]\}$ associated with the random vectors $\{\tilde{X}^n_j, j = 0, 1, \ldots, n\}$. This process is defined by

$$
\tilde{X}^n(t) = \tilde{X}^n_j + \left(1 - \frac{j}{n}\right) \tilde{Y}^n_j \quad \text{for } t \in [j/n, (j+1)/n], j = 0, 1, \ldots, n - 1
$$

and will be referred to as the controlled random walk. Equation (3.3) allows us to write $V^n$ in one of the following two equivalent forms:

$$
V^n = \inf_{\{\nu^n_j\}} \tilde{E} \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot)||\rho(\cdot)) + h(\tilde{X}^n) \right\}
$$

(3.13)

$$
= \inf_{\{\nu^n_j\}} \tilde{E} \left\{ \int_0^1 R(\nu^n(\cdot)|t)||\rho(\cdot)) dt + h(\tilde{X}^n) \right\}.
$$

Part (b) of Proposition 1.4.2 specifies for $i \in \{0, 1, \ldots, n - 1\}$ the probability measure $\tilde{\nu}^n_i$ at which the infimum in equation (3.5) is uniquely attained. For $i \in \{1, 2, \ldots, n - 1\}$ $\tilde{\nu}^n_i$ depends on $x_1, x_2, \ldots, x_i$, and we write $\tilde{\nu}^n_i(\cdot|x_1, x_2, \ldots, x_i)$. In the proof of the next theorem we will need to know that each $\tilde{\nu}^n_i(\cdot|x_1, x_2, \ldots, x_i)$ is a stochastic kernel on $\mathbb{R}^d$ given $(\mathbb{R}^d)^i$. This is a consequence of the explicit formula for $\tilde{\nu}^n_i$ given in the next lemma.

**Lemma 3.2.1.** For $i \in \{0, 1, \ldots, n - 1\}$ and $(x_1, \ldots, x_i) \in (\mathbb{R}^d)^i$ the infimum in equation (3.5) is uniquely attained at the measure $\tilde{\nu}^n_i(dy) = \tilde{\nu}^n_i(dy|x_1, \ldots, x_i)$ that assigns to a Borel subset $A$ of $\mathbb{R}^d$ the measure

$$
\tilde{\nu}^n_i(A|x_1, \ldots, x_i) = \int_A \exp \left[ -n W^\infty(i+1, \{x_1, \ldots, x_i, x_i + \frac{1}{n} \}) \right] \rho(dy)
$$

$$
\times \frac{1}{\int_{\mathbb{R}^d} \exp \left[ -n W^\infty(i+1, \{x_1, \ldots, x_i, x_i + \frac{1}{n} \}) \right] \rho(dy)}.
$$

For each $i \tilde{\nu}^n_i(dy|x_1, x_2, \ldots, x_i)$ is a stochastic kernel on $\mathbb{R}^d$ given $(\mathbb{R}^d)^i$.}
The representation formula for $W^n$ is now almost immediate and will be stated in the next theorem. This formula is a special case of a much more general representation formula that will be given in Theorem 4.3.1 in the next chapter. The continuity of the function $h$ is needed in order to prove the Laplace principle, not in order to prove the representation formula, for which a bounded measurable function $h$ suffices.

**Theorem 3.2.2.** Let $h$ be a bounded measurable function mapping $\mathcal{C}([0, 1] : \mathbb{R}^d)$ into $\mathbb{R}$. Then for all $n \in \mathbb{N}$ the quantity

$$W^n = -\frac{1}{n} \log E\{\exp[-nh(X^n)]\}$$

equals the minimal cost function $V^n$ given by either of the two formulas in (3.13).

**Proof.** The quantities $W^n(i, \{x_1, x_2, \ldots, x_i\})$ satisfy the dynamic programming equation (3.5), and we have the terminal condition

$$W^n(n, \{x_1, x_2, \ldots, x_n\}) = \alpha_n(x_1, x_2, \ldots, x_n),$$

which is also satisfied by the minimal cost function $V^n(n, \{x_1, x_2, \ldots, x_n\})$. Since the dynamic programming equation (3.5) and the terminal condition have a unique solution, we will be able to conclude that each quantity $W^n(i, \{x_1, x_2, \ldots, x_i\})$ equals the corresponding minimal cost function $V^n(i, \{x_1, x_2, \ldots, x_i\})$, and thus that $W^n = V^n$, once we confirm that the minimal cost functions $V^n(i, \{x_1, x_2, \ldots, x_i\})$ satisfy the same dynamic programming equation (3.5). This will be guaranteed by Theorem 1.5.2 after we verify the Attainment Condition. The latter requires that there exists a stochastic kernel $\nu = \tilde{\nu}_n^T$ at which the infimum in the dynamic programming equation is attained. Since this is the content of Lemma 3.2.1, the proof of the representation formula is complete. \(\blacksquare\)

We now turn to the Laplace principle for the random walks $\{X^n\}$. The representation formula will be used to prove that for all bounded continuous functions $h$ mapping $\mathbb{R}^d$ into $\mathbb{R}$

$$\lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-nh(X^n)]\} = -\inf_{\varphi \in \mathcal{C}([0, 1] ; \mathbb{R}^d)} \{I(\varphi) + h(\varphi)\},$$

where of course the rate function $I$ must be specified. Because of the first minus sign in the definition of $W^n$, the Laplace principle upper bound entails proving

$$\liminf_{n \to \infty} W^n \geq \inf_{\varphi \in \mathcal{C}([0, 1] ; \mathbb{R}^d)} \{I(\varphi) + h(\varphi)\}$$

and the Laplace principle lower bound entails proving

$$\limsup_{n \to \infty} W^n \leq \inf_{\varphi \in \mathcal{C}([0, 1] ; \mathbb{R}^d)} \{I(\varphi) + h(\varphi)\}.$$

In the next section the proof of the Laplace principle upper bound will enable us to identify the rate function in a straightforward way.
3.3 Proof of the Laplace Principle Upper Bound and
the Identification of the Rate Function

We recall the assumption that the underlying measure $\rho$ is supported on a compact subset
of $\mathbb{R}^d$. Before turning to the evaluation of the limit inferior of $W^n = V^n$, we explore
the implications of this assumption on $\rho$ with regard to compactness properties of the
admissible control measures and the associated controlled random walks.

Since $R(\nu\|\rho) = \infty$ for any probability measure $\nu$ on $\mathbb{R}^d$ that is not absolutely con-
tinuous with respect to $\rho$, it follows that $R(\nu\|\rho) = \infty$ if $\nu$ is not supported on $K$. Let us
apply this observation to the representation formula obtained in Theorem 3.2.2, which states that

$$
W^n = V^n = \inf_{\nu_j} \mathbb{E} \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu_j(\cdot)\|\rho(\cdot)) + h(\tilde{X}^n) \right\}.
$$

Without loss of generality we can restrict the infimum to admissible control sequences $\{\nu_j^n\}$
for which the running cost is finite. Each such control satisfies $\nu_j^n(K|\tilde{X}_1^n, \tilde{X}_2^n, \ldots, \tilde{X}_n^n) = 1
w.p.1$. The corresponding admissible control measures $\nu^n$ defined in equations (3.11)-(3.12)
satisfy $\nu^n(K \times [0,1]) = 1 w.p.1$, and so w.p.1 they take values in $\mathcal{P}(K \times [0,1])$. It fol-

low that w.p.1 the controlled random vectors $\{\tilde{Y}_j^n, j = 0, 1, \ldots, n\}$ defined in equations
(3.9) and (3.10) have support in $K$. We define $\Gamma$ to be the set of $\varphi \in C([0,1]: \mathbb{R}^d)$
that satisfy $\varphi(0) = 0$ and are Lipschitz continuous with constant $B \equiv \max\{\|x\| : x \in K\}.

With probability 1 $\tilde{X}^n$ takes values in $\Gamma$.

$\Gamma$ is obviously closed, and the Arzelà–Ascoli Theorem implies that it is compact [The-
orem A.6.2]. Furthermore, $\mathcal{P}(K \times [0,1])$ is compact by Prohorov’s Theorem [Theorem
A.3.15]. Thus the assumption that $\rho$ is supported on a compact set $K$ automatically
implies that any subsequence of $\{(\nu^n_j, \tilde{X}^n_j), n \in \mathbb{N}\}$ has a subsubsequence that con-
verges in distribution to a limit which takes values in $\mathcal{P}(K \times [0,1]) \times \Gamma$. This fact greatly sim-
plifies the evaluation of the limit inferior of $W^n$. In the general setting to be considered
in Chapter 6, the assumption that $\rho$ has compact support is replaced by a far weaker
condition.

We now turn to the evaluation of the limit inferior of $W^n$. We will identify a function
$I$ mapping $C([0,1]: \mathbb{R}^d)$ into $[0,\infty]$ and show that for every bounded continuous function
$h$ mapping $C([0,1]: \mathbb{R}^d)$ into $\mathbb{R}^d$

$$
\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{E} \{ \exp[-n h(X^n)] \} \leq -\inf_{\varphi \in C([0,1]: \mathbb{R}^d)} \{ I(\varphi) + h(\varphi) \}.
$$

It suffices to prove that every subsequence of $\{W^n\}$ has a subsubsequence satisfying

$$
\liminf_{n \to \infty} W^n \geq \inf_{\varphi \in C([0,1]: \mathbb{R}^d)} \{ I(\varphi) + h(\varphi) \}.
$$

We will work with a fixed subsequence of $\{W^n, n \in \mathbb{N}\}$ for the remainder of the proof.

Let $\varepsilon > 0$ be given. Theorem 3.2.2 guarantees that for each $n \in \mathbb{N}$ there exists an
admissible control measure $\nu^n$ such that $\nu^n \in \mathcal{P}(K \times [0,1]) w.p.1$ and

$$
W^n + \varepsilon = V^n + \varepsilon \geq \mathbb{E} \left\{ \int_0^1 R(\nu^n(\cdot|t)\|\rho(\cdot)) \, dt + h(\tilde{X}^n) \right\}.
$$

(3.14)
By compactness, given any subsequence of \( \{\nu^n, \bar{X}^n\} \) there exists a subsubsequence that converges in distribution to a limit \((\nu, \bar{X})\). Hence there exists a probability space, which we label \((\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})\), such that \(\nu\) maps \(\hat{\Omega}\) into \(\mathcal{P}(K \times [0,1])\) and \(\bar{X}\) maps \(\hat{\Omega}\) into \(\Gamma\). In particular, \(\nu\) is a stochastic kernel on \(K \times [0,1]\) given \(\hat{\Omega}\).

The next lemma gives a useful decomposition of the stochastic kernel \(\nu\). As equation (3.12) shows, the second marginal of each \(\nu^n\) is Lebesgue measure for all \(\omega\). The lemma merely uses the fact that w.p.1 any limit in distribution of \(\{\nu^n\}\) also has Lebesgue measure as its second marginal. In the proof of the lemma we apply the Skorohod Representation Theorem, which involves introducing a new probability space. Following our convention, we retain the notation \((\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})\) for this new space.

**Lemma 3.3.1.** Let \((\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})\) be a probability space. We assume that \(\nu = \nu(dy \times dt|\omega)\) is a stochastic kernel on \(K \times [0,1]\) given \(\hat{\Omega}\) which is the limit in distribution of a convergent subsequence of admissible control measures \(\nu^n, n \in \mathbb{N}\). There exists a stochastic kernel \(\nu(dy|\omega, t)\) on \(K\) given \([0,1] \times \hat{\Omega}\) such that \(\hat{P}\)-a.s. for \(\omega \in \hat{\Omega}\)

\[
\nu(A \times B|\omega) = \int_B \nu(A|t, \omega) \, dt
\]

for all Borel subsets \(A\) of \(K\) and \(B\) of \([0,1]\). In the sequel \(\omega\) will be suppressed in the notations for \(\nu(dy \times dt)\) and \(\nu(dy|t)\), and the last display will be summarized as \(\nu(dy \times dt) = \nu(dy|t) \otimes dt\).

**Proof.** For any Borel subset \(B\) of \([0,1]\) and \(\omega \in \hat{\Omega}\) the second marginals of \(\nu^n(dy \times dt|\omega)\) and of \(\nu(dy \times dt|\omega)\) are defined by

\[
(\nu^n)_2(B|\omega) = \nu^n(K \times B|\omega) \quad \text{and} \quad \nu_2(B|\omega) = \nu(K \times B|\omega).
\]

As we have just noted, each of the second marginals \((\nu^n)_2(\cdot|\omega)\) equals Lebesgue measure on \([0,1]\), which will be denoted by \(\lambda\). By the Skorohod Representation Theorem, since \(\nu^n \overset{D}{\rightharpoonup} \nu\), we can assume without loss of generality that w.p.1 \(\nu^n \rightharpoonup \nu\) on \((\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})\). For any bounded continuous function \(g\) mapping \([0,1]\) into \(\mathbb{R}\) we have w.p.1

\[
\lim_{n \to \infty} \int_0^1 g(t) \, dt = \lim_{n \to \infty} \int_0^1 g(t) (\nu^n)_2(dt|\omega) = \lim_{n \to \infty} \int_{\mathbb{R}^d \times [0,1]} g(t) \nu^n(dy \times dt|\omega) = \int_{\mathbb{R}^d \times [0,1]} g(t) \nu(dy \times dt|\omega) = \int_0^1 g(t) \nu_2(dt|\omega).
\]

This implies that w.p.1 \(\nu_2(dt|\omega)\) equals \(\lambda(dt)\) [Theorem A.2.2 (b)]. By the use of regular conditional distributions, there exists a stochastic kernel \(\nu(dy|t, \omega)\) on \(K\) given \([0,1] \times \hat{\Omega}\) such that w.p.1

\[
\nu(A \times B|\omega) = \int_B \nu(A|t, \omega) \nu_2(dt|\omega) = \int_B \nu(A|t, \omega) \, dt
\]

for all Borel subsets \(A\) of \(K\) and \(B\) of \([0,1]\) [Theorem A.5.6]. The proof is complete. \(\blacksquare\)

The next lemma relates the limit \(\bar{X}\) of the controlled random walks \(\{\bar{X}^n\}\) to the limit \(\nu\) of the admissible control measures. This lemma is the analogue of Lemma 2.5.1 in the previous chapter.
3.3. PROOF OF UPPER BOUND

Lemma 3.3.2. We assume that the underlying measure \( \rho \) is supported on a compact subset \( K \). The following conclusions hold.

(a) Given any subsequence of \( \{ (\nu^n, X^n), n \in \mathbb{N} \} \) there exists a subsequence, a probability space \((\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})\), a stochastic kernel \( \nu \) on \( K \times [0,1] \) given \( \tilde{\Omega} \), and a random variable \( \tilde{X} \) mapping \( \tilde{\Omega} \) into \( \mathcal{C}([0,1]; \mathbb{R}^d) \) such that the subsequence converges in distribution to \((\nu, \tilde{X})\). The stochastic kernel \( \nu = \nu(dy \times dt) \) has the decomposition given in Lemma 3.3.1.

(b) With probability 1, for every \( t \in [0,1] \) the limiting process \( \tilde{X} = \{ \tilde{X}(t), t \in [0,1] \} \) satisfies

\[
\tilde{X}(t) = \int_{K \times [0,t]} y \nu(dy \times ds) = \int_0^t \left( \int_K y \nu(dy|s) \right) ds,
\]

and \( \tilde{X}(t) \) is an absolutely continuous function of \( t \in [0,1] \). Therefore, a.s. with respect to Lebesgue measure for \( t \in [0,1] \) the derivative of \( \tilde{X}(t) \) is given by

\[
\dot{\tilde{X}}(t) = \int_K y \nu(dy|t).
\]

This lemma is a special case of a general result to be proved in Theorem 5.3.5. The lemma is easily motivated by considering the case where all the stochastic kernels \( \nu^n_j \) equal a fixed probability measure \( \gamma \) on \( K \). Then each of the control measures \( \nu^n \) equals the product measure \( \gamma \times \lambda \) on \( K \times [0,1] \), where \( \lambda \) denotes Lebesgue measure on \([0,1]\). In addition \( \tilde{X}_0^n = 0 \) and for each \( j \in \{1, 2, \ldots, n\} \)

\[
\tilde{X}_j^n = \frac{1}{n} \sum_{i=0}^{j-1} \tilde{Y}_i^n,
\]

where \( \{\tilde{Y}_i^n, i = 0, 1, \ldots, n-1\} \) are i.i.d. random vectors with common distribution \( \gamma \). The law of large numbers implies that as \( n \to \infty \)

\[
(\tilde{X}^n, \nu^n) \xrightarrow{D} \left( \left\{ t \int_K y \gamma(dy), t \in [0,1] \right\}, \gamma \times \lambda \right).
\]

This is consistent with Lemma 3.3.2.

We return to the proof of the Laplace principle upper bound, indexing by \( n \) the subsequence along which the convergence in distribution \((\nu^n, \tilde{X}^n) \xrightarrow{D} (\nu, \tilde{X})\) in part (a) of Lemma 3.3.2 is valid. As in the proof of the Laplace principle upper bound in Sanov’s Theorem, we now apply the Skorohod Representation Theorem. This guarantees the existence of a probability space such that the subsequence \((\nu^n, \tilde{X}^n)\) converges to \((\nu, \tilde{X})\) w.p.1 on this space. Expectation on this space is denoted by \( \tilde{E} \). Our goal is now to evaluate the limit inferior of \( W^n \) using equation (3.14), in which the running cost is given by \( \int_0^1 R(\nu^n(\cdot|t)\|\rho(\cdot)) \, dt \). Rather than take limits of the stochastic kernels \( \nu^n(\cdot|t) \) as \( n \to \infty \), it is much more convenient to rewrite the running cost as a relative entropy involving the admissible control measure \( \nu^n(dy \times dt) = \nu^n(dy|t) \otimes dt \). Let \( \lambda \) denote Lebesgue measure on \([0,1]\) and \( \rho \times \lambda \) product measure on \( K \times [0,1] \). Then according to part (f) of Lemma 1.4.3

\[
\int_0^1 R(\nu^n(\cdot|t)\|\rho(\cdot)) \, dt = R(\nu^n\|\rho \times \lambda).
\]
Rewriting the running cost in this way will ease the application of weak convergence theory in calculating the limit of the minimal cost functions.

The evaluation of the limit inferior of $W^n$ is carried out in the following string of inequalities, each line of which will be explained afterwards:

\[
\liminf_{n \to \infty} W^n + \varepsilon = \liminf_{n \to \infty} V^n + \varepsilon \\
\geq \liminf_{n \to \infty} \mathbb{E}\left\{ \int_0^1 R(\nu^n(\cdot)|t)\|\rho(\cdot))dt + h(X^n) \right\} \\
= \liminf_{n \to \infty} \mathbb{E}\left\{ R(\nu^n|\rho \times \lambda) + h(X^n) \right\} \\
\geq \mathbb{E}\left\{ R(\nu|\rho \times \lambda) + h(X) \right\} \\
= \mathbb{E}\left\{ \int_0^1 R(\nu(\cdot)|t)\|\rho(\cdot))dt + h(X) \right\}.
\]

The first two lines of this display are formula (3.14). As already noted, line three is implied by part (f) of Lemma 1.4.3. In order to derive line four, we have the limit

\[
\lim_{n \to \infty} \mathbb{E}\{h(X^n)\} = \mathbb{E}\{h(X)\},
\]

which is a consequence of the probability–1 convergence $X^n \to X$ and the boundedness and continuity of $h$, and we also have the lower limit

\[
\liminf_{n \to \infty} \mathbb{E}\{R(\nu^n|\rho \otimes \lambda)\} \geq \mathbb{E}\{R(\nu|\rho \otimes \lambda)\},
\]

which follows from the probability–1 convergence $\nu^n \Rightarrow \nu$, the nonnegativity and lower semicontinuity of $R(\cdot|\rho)$, and Fatou’s Lemma. Line five of (3.15) is a consequence of the decomposition of $\nu$ given in Lemma 3.3.1 and again part (f) of Lemma 1.4.3.

Let us now massage the last line of (3.15) so that it is written in the form required by the Laplace principle; namely, as an infimum over the space $\mathcal{C}([0,1]:\mathbb{R}^d)$ in which the random walks $X^n$ take values. According to part (b) of Lemma 3.3.2, $\hat{X}$ takes values in $\mathcal{C}([0,1]:\mathbb{R}^d)$ and for almost every $t \in [0,1]$ the derivative $\hat{X}(t)$ equals the mean of the first argument of the relative entropy; viz.,

\[
\hat{X}(t) = \int_K y \nu(dy)|t).
\]

Hence it makes sense to introduce for $\beta \in \mathbb{R}^d$

\[
M(\beta) = \inf \left\{ R(\nu|\rho) : \nu \in \mathcal{P}(\mathbb{R}^d), \int_{\mathbb{R}^d} y \nu(dy) = \beta \right\}.
\]

Since $R(\nu|\rho) = \infty$ unless $\nu$ is supported on the same set $K$ as $\rho$, it follows that

\[
M(\beta) = \inf \left\{ R(\nu|\rho) : \nu \in \mathcal{P}(K), \int_K y \nu(dy) = \beta \right\}.
\]

Many readers having experience with large deviation theory will know that $M$ has an alternate expression as a Legendre–Fenchel transform; namely, $M(\beta)$ equals

\[
L(\beta) = \sup_{\alpha \in \mathbb{R}^d} \left\{ \langle \alpha, \beta \rangle - \log \int_K \exp(\alpha, y) \rho(dy) \right\},
\]

where $\langle \cdot, \cdot \rangle$ is the inner product.

**CHAPTER 3. MOGULSKII'S THEOREM**
where \(\langle \cdot, \cdot \rangle\) denotes the Euclidean inner product on \(\mathbb{R}^d\). Properties of such Legendre-Fenchel transforms will play a key role in developing the generalizations of Mogulskii’s Theorem in Chapters 6 and 7.

In the next lemma we state the facts about \(L\) that are needed in the proof of Mogulskii’s Theorem.

**Lemma 3.3.3.** Let \(\rho \in \mathcal{P}(\mathbb{R}^d)\) be supported on a compact set \(K\). For \(\beta \in \mathbb{R}^d\) we define \(L(\beta)\) by formula (3.17). The following conclusions hold.

(a) \(L(\beta)\) is a nonnegative, convex, lower semicontinuous function of \(\beta \in \mathbb{R}^d\).

(b) For each \(\beta \in \mathbb{R}^d\)

\[
L(\beta) = \inf \left\{ R(\nu||\rho) : \nu \in \mathcal{P}(K), \int_K y \nu(dy) = \beta \right\},
\]

and the infimum is always attained. If \(L(\beta) < \infty\), then the infimum is uniquely attained at some measure \(\nu\).

(c) If \(\nu\) is a probability measure on \(K\), then \(\int_K \|y\| \nu(dy) < \infty\) and

\[
R(\nu||\rho) \geq L\left(\int_K y \nu(dy)\right).
\]

**Comments on the Proof.** All of these facts are proved in greater generality in Lemma 6.2.3. The convexity and lower semicontinuity of \(L\) are a consequence of the definition of this function while the nonnegativity follows immediately from the variational formula in part (b). The latter also yields part (c).

Since all that is needed to complete the proof of the Laplace principle upper bound in Mogulskii’s Theorem is part (c), let us give a quick and independent proof of that. According to the Donsker–Varadhan variational formula for the relative entropy [Lemma 1.4.3 (a)], for any probability measure \(\nu\) on \(K\) and any bounded continuous function \(g\) mapping \(K\) into \(\mathbb{R}\)

\[
R(\nu||\rho) \geq \int_K g(y) \nu(dy) - \log \int_K \exp[g(y)] \rho(dy).
\]

Given \(\alpha\) be any point in \(\mathbb{R}^d\), we evaluate this inequality for the bounded continuous function \(g(y) = \langle \alpha, y \rangle\), \(y \in K\), obtaining

\[
R(\nu||\rho) \geq \int_K \langle \alpha, y \rangle \nu(dy) - \log \int_K \exp[\langle \alpha, y \rangle] \rho(dy) = \langle \alpha, \beta \rangle - \log \int_K \exp[\langle \alpha, y \rangle] \rho(dy),
\]

where \(\beta = \int_K y \nu(dy)\). Since \(\alpha \in K\) is arbitrary, it follows that

\[
R(\nu||\rho) \geq \sup_{\alpha \in \mathbb{R}^d} \left\{ \langle \alpha, \beta \rangle - \log \int_K \exp[\langle \alpha, y \rangle] \rho(dy) \right\} = L(\beta) = L\left(\int_K y \nu(dy)\right).
\]

This proves part (c) of the lemma.

We now complete the proof of the Laplace principle upper bound. Let us denote by \(\mathcal{A}_0([0,1] : \mathbb{R}^d)\) the subset of \(\mathcal{C}([0,1] : \mathbb{R}^d)\) consisting of absolutely continuous functions
CHAPTER 3. MOGULSKII'S THEOREM

\( \varphi \) satisfying \( \varphi(0) = 0 \). We have from (3.15), part (c) of Lemma 3.3.3, and part (b) of Lemma 3.3.2

\[
\liminf_{n \to \infty} W^n + \varepsilon \geq \mathbb{E} \left\{ \int_0^1 \mathcal{R}(\nu(\cdot \mid t)\|\nu(\cdot)) \, dt + h(\bar{X}) \right\}
\]
\[
\geq \mathbb{E} \left\{ \int_0^1 L \left( \int_0^1 \nu(dy|t) \right) \, dt + h(\bar{X}) \right\}
\]
\[
= \mathbb{E} \left\{ \int_0^1 L(\hat{X}(t)) \, dt + h(\bar{X}) \right\}
\]
\[
\geq \inf_{\varphi \in \mathcal{A}_0([0,1]:\mathbb{R}^d)} \left\{ \int_0^1 L(\hat{\varphi}(t)) \, dt + h(\varphi) \right\}.
\]

This display leads naturally to the following definition of a rate function \( I \):

\[
I(\varphi) = \begin{cases} 
\int_0^1 L(\hat{\varphi}(t)) \, dt & \text{if } \varphi \in C([0,1]:\mathbb{R}^d) \text{ is absolutely continuous and } \varphi(0) = 0 \\
\infty & \text{for all other } \varphi \in C([0,1]:\mathbb{R}^d).
\end{cases}
\]

(3.18)

Since \( \varepsilon > 0 \) is arbitrary, we have proved that every subsequence of the original sequence \( \{W^n, n \in \mathbb{N}\} \) has a subsubsequence satisfying

\[
\liminf_{n \to \infty} W^n \geq \inf_{\varphi \in C([0,1]:\mathbb{R}^d)} \{ I(\varphi) + h(\varphi) \}.
\]

An argument by contradiction applied to an arbitrary subsequence shows that the entire sequence \( \{W^n, n \in \mathbb{N}\} \) satisfies this same lower limit, which has the required form of the Laplace principle upper bound. In the next section we complete the proof of Mogulskii’s Theorem by showing the Laplace principle lower bound.

3.4 Statement of Mogulskii’s Theorem and Completion of the Proof

Having guessed the form of the rate function, we can now state Mogulskii’s Theorem, which we do in the form of the Laplace principle. By Theorems 1.2.1 and 1.2.3 the Laplace principle is equivalent to the large deviation principle with the same rate function. The theorem that is traditionally known as Mogulskii’s Theorem uses an assumption that considerably weakens our assumption of compact support. A far-reaching generalization of Mogulskii’s Theorem is given in Theorem 6.3.3.

**Theorem 3.4.1 (Mogulskii).** Let \( \rho \in \mathcal{P}(\mathbb{R}^d) \) be supported on a compact set \( K \) and consider the random walks \( \{X^n, n \in \mathbb{N}\} \) defined in equations (3.1) and (3.2). Then the sequence \( \{X^n\} \) satisfies the Laplace principle on \( C([0,1]:\mathbb{R}^d) \) with rate function \( I \) defined in equation (3.18). In other words, for all bounded continuous functions \( h \) mapping \( C([0,1]:\mathbb{R}^d) \) into \( \mathbb{R} \)

\[
\lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}\{\exp[-n \, h(X^n)]\} = -\inf_{\varphi \in C([0,1]:\mathbb{R}^d)} \{ I(\varphi) + h(\varphi) \}.
\]
3.4. STATEMENT OF THEOREM AND COMPLETION OF PROOF

The proof of the Laplace principle upper bound in Mogulskii’s Theorem was sketched in the previous section. We now turn to the Laplace principle lower bound, omitting the proof that \( I \) is a rate function. That fact follows from Proposition 6.2.4, which covers a much more general situation. It is worth noting that the proof of the proposition is a deterministic version of the proof of the corresponding Laplace principle upper bound, and the same can be done here.

We now prove the Laplace principle lower bound, which is equivalent to

\[
\limsup_{n \to \infty} W_n \leq \inf_{\varphi \in C([0,1]:\mathbb{R}^d)} \{ I(\varphi) + h(\varphi) \}. \tag{3.19}
\]

Let \( \varepsilon > 0 \) be given. We choose \( \varphi^* \in C([0,1]:\mathbb{R}^d) \) such that

\[
I(\varphi^*) + h(\varphi^*) \leq \inf_{\varphi \in C([0,1]:\mathbb{R}^d)} \{ I(\varphi) + h(\varphi) \} + \varepsilon < \infty.
\]

It follows that \( \varphi^* \) is absolutely continuous and satisfies \( \varphi^*(0) = 0 \). The basic idea in proving (3.19) is to construct from \( \varphi^* \) a sequence of admissible control measures to be used in the representation formulas for \( W_n \). However, rather than work directly with \( \varphi^* \), we will replace it by a certain averaged function, a trick that will greatly simplify the proof.

For \( \kappa \in \mathbb{N} \) and \( t \in [0,1] \) we define

\[
\gamma_\kappa(t) = \begin{cases} 
\kappa \int_{i/k}^{(i+1)/k} \varphi^*(s) \, ds & \text{if } t \in [i/k, (i+1)/k), i \in \{0, 1, \ldots, \kappa - 2 \} \\
\kappa \int_{(\kappa-1)/k}^{1} \varphi^*(s) \, ds & \text{if } t \in [(\kappa - 1)/k, 1]
\end{cases}
\]

and

\[
\varphi_\kappa(t) = \int_0^t \gamma_\kappa(\tau) \, d\tau.
\]

Then

\[
\lim_{\kappa \to \infty} \sup_{t \in [0,1]} \| \varphi^*(t) - \varphi_\kappa(t) \| = 0,
\]

and since \( h \) is continuous, \( h(\varphi_\kappa) \to h(\varphi^*) \) as \( \kappa \to \infty \). The convexity of \( L \) and Jensen’s Inequality imply that

\[
\int_0^1 L(\varphi_\kappa(t)) \, dt = \frac{1}{\kappa} \sum_{i=0}^{\kappa-1} L\left( \kappa \int_{i/k}^{(i+1)/k} \varphi^*(s) \, ds \right) \leq \sum_{i=0}^{\kappa-1} \int_{i/k}^{(i+1)/k} L(\varphi^*(s)) \, ds = \int_0^1 L(\varphi^*(s)) \, ds.
\]

For the given \( \varepsilon > 0 \) it follows that we can fix a value of \( \kappa \in \mathbb{N} \) such that

\[
\int_0^1 L(\varphi_\kappa(t)) \, dt + h(\varphi_\kappa) \leq \int_0^1 L(\varphi^*(s)) \, ds + h(\varphi^*) + \varepsilon = I(\varphi^*) + h(\varphi^*) + \varepsilon 
\]

\[
\leq \inf_{\varphi \in C([0,1]:\mathbb{R}^d)} \{ I(\varphi) + h(\varphi) \} + 2\varepsilon. \tag{3.20}
\]
We now lift \( \varphi_\kappa \) to the realm of probability measures. For each \( i \in \{0, 1, \ldots, \kappa - 1\} \) and all \( t \in (i/\kappa, (i + 1)/\kappa) \), \( \varphi_\kappa(t) \) takes the constant value \( \gamma_\kappa(i/\kappa) \). For each such \( i \), part (b) of Lemma 3.3.3 guarantees the existence of \( \bar{\nu}_\kappa \in \mathcal{P}(K) \) such that

\[
R(\bar{\nu}_\kappa^n(\|1\|\rho)) = L(\gamma_\kappa(i/\kappa)) \quad \text{and} \quad \int_K y \, \bar{\nu}_\kappa^n(dy) = \gamma_\kappa(i/\kappa).
\]

For \( n \in \mathbb{N} \) we define a sequence \( \{\nu^n_n, j = 0, 1, \ldots, n - 1\} \) in \( \mathcal{P}(K) \) by

\[
\nu^n_j \doteq \begin{cases} 
\bar{\nu}_\kappa^n & \text{if } j/n \in [i/\kappa, (i + 1)/\kappa), i = 0, 1, \ldots, \kappa - 2 \\
\bar{\nu}_{\kappa - 1} & \text{if } j/n \in [(\kappa - 1)/\kappa, 1].
\end{cases}
\]

This sequence of probability measures defines an admissible control sequence. In terms of it we define the controlled random walk \( \bar{X}^n \) and the admissible control measure \( \nu^n \), which will be used in the representation formula for \( W^n \). Specifically, given Borel subsets \( A \) of \( K \) and \( B \) of \([0, 1]\), we define a stochastic kernel on \( K \) given \([0, 1]\) by

\[
\nu^n(A|t) \doteq \begin{cases} 
\nu^n_j(A) & \text{if } t \in [j/n, (j + 1)/n), j = 0, 1, \ldots, n - 2 \\
\nu^n_{n-1}(A) & \text{if } t \in [(n - 1)/n, 1]
\end{cases}
\]

and a probability measure \( \nu^n \) on \( K \times [0, 1] \) by

\[
\nu^n(A \times B) \doteq \int_B \nu^n(A|s) \, ds.
\]

The representation formula in Theorem 3.2.2 gives

\[
W^n = V^n \leq \int_0^1 R(\nu^n(\cdot|t)\|\rho(\cdot)) \, dt + \mathcal{E}\{h(\bar{X}^n)\}. \tag{3.21}
\]

In order to study the limits of the terms on the right hand side of this inequality, it is also convenient to define a stochastic kernel \( \bar{\nu}(dy|t) \) on \( K \times [0, 1] \) by

\[
\bar{\nu}(dy|t) \doteq \begin{cases} 
\bar{\nu}_\kappa^n(dy) & \text{if } t \in [i/\kappa, (i + 1)/\kappa), i = 0, 1, \ldots, \kappa - 2 \\
\bar{\nu}_{\kappa - 1}^n(dy) & \text{if } t \in [(\kappa - 1)/\kappa, 1].
\end{cases}
\]

For Borel subsets \( A \) of \( K \) and \( B \) of \([0, 1]\) we then define a probability measure \( \bar{\nu} \) on \( K \times [0, 1] \) by

\[
\bar{\nu}(A \times B) \doteq \int_B \bar{\nu}(A|t) \, dt.
\]

As we show in a moment, \( \bar{\nu} \) is the limit of \( \nu^n \) as \( n \to \infty \). The function \( \varphi_\kappa \) has a simple expression in terms of \( \varphi(\cdot|t) \). Indeed, except for finitely many values of \( t \in [0, 1] \)

\[
\int_{\mathbb{R}^d} y \, \bar{\nu}(dy|t) = \varphi_\kappa(t),
\]

and since \( \varphi_\kappa(0) = 0 \), it follows that for all \( t \in [0, 1] \)

\[
\varphi_\kappa(t) = \int_0^t \left( \int_K y \, \bar{\nu}(dy|s) \right) ds.
\]
3.4. STATEMENT OF THEOREM AND COMPLETION OF PROOF

We now use formula (3.21) to evaluate the limit superior of $W^n$. The main nuisance is the presence of two discretizations, the first in terms of $\kappa$ and the second in terms of $n$. When $n$ is multiple of $\kappa$, $\nu^n$ and $\tilde{\nu}$ are equal. When $n$ is not a multiple of $\kappa$, there are small errors that must be dealt with. Our first claim is that $\nu^n \rightarrow \tilde{\nu}$ as $n \rightarrow \infty$. Indeed, it follows from the definitions that $\nu^n(\cdot | t)$ and $\tilde{\nu}(\cdot | t)$ agree except for $t$ in a set of Lebesgue measure less than $\kappa/n$. Thus in total variation norm $\| \nu^n - \tilde{\nu} \|_v \leq 2\kappa/n$, which implies that $\nu^n \rightarrow \tilde{\nu}$.

According to part (a) of Lemma 3.3.2, every subsequence of $\{\nu^n, \tilde{\xi}^n, n \in \mathbb{N}\}$ has a subsequence that converges in distribution to a pair $(\nu, \tilde{\xi})$. We have just identified the limit $\nu = \tilde{\nu}$, and by part (b) of Lemma 3.3.2, w.p.1 we have for every $t \in [0, 1]$

$$\tilde{\xi}(t) = \int_{K \times [0, t]} y \tilde{\nu}(dy \times ds) = \int_0^t \left( \int_K y \tilde{\nu}(dy|s) \right) ds = \varphi_\kappa(t).$$

Since the limit $(\nu, \tilde{\xi})$ is independent of the subsequence chosen, it follows that the entire sequence $\{(\nu^n, \tilde{\xi}^n), n \in \mathbb{N}\}$ converges in distribution to $(\nu, \tilde{\xi}) = (\tilde{\nu}, \varphi_\kappa)$.

The limit of the relative entropy term on the right hand side of (3.21) is also easily determined. The claim is that

$$\lim_{n \rightarrow \infty} \int_0^1 R(\nu^n(\cdot | t)\| \rho(\cdot)) \ dt = \int_0^1 R(\tilde{\nu}(\cdot | t)\| \rho(\cdot)) \ dt.$$ 

Indeed, since $\nu^n(\cdot | t)$ and $\tilde{\nu}(\cdot | t)$ agree except for $t$ in a set of Lebesgue measure less than $\kappa/n$,

$$\left| \int_0^1 R(\nu^n(\cdot | t)\| \rho(\cdot)) \ dt - \int_0^1 R(\tilde{\nu}(\cdot | t)\| \rho(\cdot)) \ dt \right| < \frac{\kappa}{n} \left( \max_{i \in \{0, 1, \ldots, \kappa - 1\}} R(\tilde{\nu}^i\| \rho) \right).$$

This converges to 0 as $n \rightarrow \infty$ since each of the quantities $R(\tilde{\nu}^i\| \rho)$ is finite. The finiteness follows from (3.20) and

$$\frac{1}{\kappa} \sum_{i=0}^{\kappa-1} R(\tilde{\nu}^i\| \rho) + h(\varphi_\kappa) = \frac{1}{\kappa} \sum_{i=0}^{\kappa-1} L(\gamma_\kappa(i/\kappa)) + h(\varphi_\kappa)$$

$$= \int_0^1 L(\tilde{\varphi}_\kappa(t)) \ dt + h(\varphi_\kappa)$$

$$\leq \inf_{\varphi \in C([0,1]; \mathbb{R}^d)} \left\{ I(\varphi) + h(\varphi) \right\} + 2\varepsilon.$$

We now take limits in (3.21), obtaining

$$\limsup_{n \rightarrow \infty} W^n = \limsup_{n \rightarrow \infty} V^n$$

$$\leq \limsup_{n \rightarrow \infty} \left( \int_0^1 R(\nu^n(\cdot | t)\| \rho(\cdot)) \ dt + E\{h(\tilde{\xi}^n)\} \right)$$

$$= \int_0^1 R(\tilde{\nu}(\cdot | t)\| \rho(\cdot)) \ dt + h(\varphi_\kappa)$$

$$= \frac{1}{\kappa} \sum_{i=0}^{\kappa-1} R(\tilde{\nu}^i\| \rho) + h(\varphi_\kappa)$$

$$\leq \inf_{\varphi \in C([0,1]; \mathbb{R}^d)} \left\{ I(\varphi) + h(\varphi) \right\} + 2\varepsilon.$$
Since $\varepsilon > 0$ is arbitrary, we have proved the upper limit (3.19) and thus the Laplace principle lower bound. This completes our sketch of the proof of Mogulskii’s Theorem. ■

In the next section we show how to derive Cramér’s Theorem using the contraction principle.

### 3.5 Cramér’s Theorem

As a bonus of our work in this chapter, we obtain Cramér’s Theorem, which is the large deviation principle for the sequence of sample means

$$\frac{S_n}{n} \approx \frac{1}{n} \sum_{j=0}^{n-1} v_j$$

of the i.i.d. random vectors $v_j$. In order to carry this out, we define a function $f$ mapping $C([0, 1] : \mathbb{R}^d)$ into $\mathbb{R}^d$ by $f(\varphi) \equiv \varphi(1)$. Then $f$ is continuous and $f(X^n) = X^n(1) = S^n/n$. We recall the function $L$ defined for $\beta \in \mathbb{R}^d$ by

$$L(\beta) \equiv \sup_{\alpha \in \mathbb{R}^d} \left\{ \langle \alpha, \beta \rangle - \log \int_{K} \exp(\alpha, y) \rho(dy) \right\}.$$

Mogulskii’s Theorem and the contraction principle [Theorem 1.3.2] imply that the sequence $\{S^n/n, n \in \mathbb{N}\}$ satisfies the Laplace principle on $\mathbb{R}^d$ with rate function

$$J(\beta) \equiv \inf \{ I(\varphi) : \varphi \in f^{-1}(\beta) \} = \inf \left\{ \int_0^1 L(\dot{\varphi}(t)) dt : \varphi \in \mathcal{A}_0([0, 1] : \mathbb{R}^d), \varphi(1) = \beta \right\},$$

where $\mathcal{A}_0([0, 1] : \mathbb{R}^d)$ is the set of absolutely continuous functions $\varphi$ mapping $[0, 1]$ into $\mathbb{R}$ and satisfying $\varphi(0) = 0$. If for $\beta \in \mathbb{R}^d$ and $t \in [0, 1]$ we define $\varphi_{\beta}(t) \equiv \beta t$, then

$$J(\beta) \leq I(\varphi_{\beta}) = L(\beta).$$

On the other hand, by Jensen’s Inequality and the convexity of $L$, we obtain the opposite inequality

$$J(\beta) \geq \inf \left\{ L\left( \int_0^1 \dot{\varphi}(t) dt \right) : \varphi \in \mathcal{A}_0([0, 1] : \mathbb{R}^d), \varphi(1) = \beta \right\} = L(\beta).$$

Hence $J$ equals $L$. This yields Cramér’s Theorem in the case that the distribution $\rho$ of the random vectors $\{v_j, j \in \mathbb{N}_0\}$ satisfies the hypothesis of Theorem 3.4.1; namely, that $\rho$ is supported on a compact subset $K$ of $\mathbb{R}^d$. Applying the contraction principle to the generalization of Mogulskii’s Theorem given in Theorem 6.3.3 yields Cramér’s Theorem for a much broader class of measures $\rho$. The assumption of compact support is replaced there by a suitable boundedness condition, which is presented in the next theorem.
3.6. COMMENTS ON THE PROOFS

**Theorem 3.5.1 (Cramér).** Let \( \rho \) be a probability measure on \( \mathbb{R}^d \) such that for all \( \alpha \in \mathbb{R}^d \)

\[
\int_{\mathbb{R}^d} \exp \langle \alpha, y \rangle \rho(dy) < \infty.
\]

Then the sequence of sample means \( \{S^n/n, n \in \mathbb{N}\} \) satisfies the Laplace principle on \( \mathbb{R}^d \) with rate function

\[
L(\beta) = \sup_{\alpha \in \mathbb{R}^d} \left\{ \langle \alpha, \beta \rangle - \log \int_{\mathbb{R}^d} \exp \langle \alpha, y \rangle \rho(dy) \right\}.
\]

In other words, for all bounded continuous functions \( h \) mapping \( \mathbb{R}^d \) into \( \mathbb{R} \)

\[
\lim_{n \to \infty} \frac{1}{n} \log E\{\exp[-n h(S^n/n)]\} = - \inf_{\beta \in \mathbb{R}^d} \{L(\beta) + h(\beta)\}.
\]

In the next section we comment on several features of the proofs of Sanov’s Theorem and of Mogulskii’s Theorem which have a bearing on future developments in the book.

### 3.6 Comments on the Proofs

At this point we have exposed the weak convergence approach in the context of two elementary examples. Basic ingredients in these two examples, and indeed in all the examples that we will consider, are the convergence properties of the controls and controlled processes as \( n \to \infty \) (cf., Lemmas 2.5.1 and 3.3.2). As we have observed, in order to carry out the convergence analysis it is useful to introduce admissible control measures, which place the controls in a space that is independent of \( n \). However, it may appear that the definitions of the admissible control measures in the two chapters are somewhat ad hoc. In addition, Jensen’s Inequality is used in the proof of the upper bound in Sanov’s Theorem to average out the time parameter of the controls in equation (2.15), while in Mogulskii’s Theorem we were careful to preserve the time dependence of the controls in the definition of the control measures in equations (3.11) and (3.12). The purpose of the present section is to address these issues and to indicate how one can simplify the convergence analysis as much as possible in other problems.

We begin with the observation that the admissible controls appearing in the prelimit quantities \( V^n \) have a number of dependencies. For example, as is shown by Lemma 2.3.1, in the setting of Sanov’s Theorem the optimal control applied at time \( i \) depends both on \( i \) and on the controlled measure \( L^n_i \). In the case of Mogulskii’s Theorem there is a more complicated dependence. Here the optimal control applied at time \( i \) depends both on \( i \) and on the entire past of the controlled random walk up until time \( i \) [Lemma 3.2.1]. These dependencies are forced by the dynamic programming equation. As a consequence, one might consider it important to define the admissible control measures in these two problems so that these dependencies are preserved in the limit \( n \to \infty \). For example, in the case of Sanov’s Theorem this would suggest that the admissible control measure should be defined so that it explicitly records the dependencies mentioned above. Interestingly, the admissible control measure \( \alpha^n \) appearing in equation (2.17) in the proof of Sanov’s
CHAPTER 3. MOGULSKII’S THEOREM

Theorem ignores both of these dependencies, and in fact \( \alpha^n \) is a simple average of the
controls applied at times \( 0, 1, \ldots, n - 1 \). In contrast, in the case of Mogulskii’s Theorem
the dependence of the controls \( \nu^n \) on the time at which they are applied is preserved in
the definition of the admissible control measure \( \nu^n \). However, no other information is
kept. In particular, one cannot obtain from \( \nu^n \) the dependence of the controls on the past
of the random walk.

The issue of which dependencies need to be recorded in the admissible control measure
is intimately tied to the form of the rate function for the given problem. Put another
way, while it may be true that certain dependencies are required to obtain an exact
representation at the prelimit level, it is not always the case that optimizing or nearly
optimizing controls at the limit level require such dependencies. This is illuminated by
the procedure used to prove the Laplace principle lower bound. In both chapters we start
by considering a nearly minimizing element in a variational problem of the form \( \inf \{ I + h \} \).
We then use this element to manufacture a nearly optimizing control in the representation
formula for the prelimit problem. Clearly, any dependencies that these nearly optimizing
controls have must be reflected in a corresponding dependency in the admissible control
measure. Otherwise, we will not be able to obtain a class of limit control measures from
which the optimal control for the limit can be constructed. For example, in the case of
Mogulskii’s Theorem the controls are defined in terms of a nearly minimizing trajectory
\( \phi_n \), and the control applied at the discrete time \( j/n = t \) depends only on \( \phi_n(t) \) and not
on any other information. This is reflected in the form of the admissible control measure
\( \nu^n \) in Section 3.4. Given \( \nu^n \), one can obtain the control applied at a given time \( t \) by
examining the measures \( \nu^n(\cdot | t) \). In the case of Sanov’s Theorem, the controls used in the
proof of the lower bound are all equal to a particular fixed measure. None of the controls
depends either on the time at which it is applied or on the state of the controlled empirical
measure.

In problems to be studied later on, we will see that in some cases extensive information
must be incorporated into the admissible control measures. For example, in the analysis
of the empirical measures of a Markov chain in Chapter 8, the admissible control measure
will have to keep track of the dependence of the controls on the state of the controlled
Markov chain when they are applied, but not on the time or the state of the controlled
empirical measure when they are applied. In the case of the state-dependent random
disks to be treated in Chapter 6, the admissible control measure will have to record the
dependence of the controls both on the time and on the current state of the controlled
random walk, but not on the entire past of the controlled random walk. In Chapter
7, where we will consider a random walk model with discontinuous statistics, we must
keep track of a great deal of information: the time at which the control is applied, the
current state of the controlled random walk, as well as which side of the discontinuity the
controlled random walk is on. In every case the fact that such information is needed can
be seen by examining the form that the rate function takes. This is easily verified in the
context of the examples in Chapters 2 and 3.

While this discussion illuminates the dependencies that one should expect of the ad-
missible control measure for a given problem, it does not specify the actual form of the
control measure. This turns out to depend on two factors: first, whether or not Jensen’s
3.6. COMMENTS ON THE PROOFS

Inequality can be used to simplify the convergence analysis and second, the form of the representation that is most convenient for applying weak convergence methods. To illustrate the latter issue, we consider the proof of the Laplace principle upper bound in Mogulskii’s Theorem. The first representation formula that we derived, stated in Theorem 3.2.2, is not suited for the application of weak convergence methods. Indeed, in order to calculate limits using this formula we would need to know that there is a fixed subsequence of \( n \in \mathbb{N} \) such that for all \( t \in [0,1] \) the quantities \( \nu^n_j \) converge and have an identifiable limit whenever \( j/n \to t \). Such convergence need not hold. For the convergence analysis, it is much more convenient to rewrite the running cost in the form \( R(\nu^n || \rho \times \lambda) \) by using part (f) of Lemma 1.4.3 and then to extract a convergent subsequence. Such a representation allows us to exploit a weak convergence in the time variable, a situation that is not possible with the original representation. This form of the representation directly motivates the form of the admissible control measure given in (3.11) and (3.12).

We next discuss when Jensen’s Inequality can be used to simplify the convergence analysis in the proof of the Laplace principle. In order to understand when this is possible, recall that the representation formulas give exact representations; there are no error terms. If one uses Jensen’s Inequality to move an average with respect to some parameter from outside the relative entropy to the inside, then one obtains an inequality that is strict unless the quantity being averaged is actually independent of the parameter with respect to which the average is taken. If \( V^n \) is the minimal cost function appearing in the representation formula and if \( U^n \) is the function resulting from the application of Jensen’s Inequality, then we have \( V^n \geq U^n \). Since the representations are exact, the Laplace principle is determined by the asymptotic behavior of \( V^n \). Let us assume that as \( n \to \infty \) the limits of the quantities \( V^n \) exist for all bounded continuous functions \( h \) and that a Laplace principle is valid. What are the consequences?

If the functions \( U^n \) also possess limits, then the Laplace principle upper bound obtained from \( \lim_{n \to \infty} V^n \geq \lim_{n \to \infty} U^n \) is tight if and only if \( \lim_{n \to \infty} V^n = \lim_{n \to \infty} U^n \) for all bounded continuous functions \( h \). This means that any strict inequality that might have been introduced by applying Jensen’s Inequality must vanish in the limit \( n \to \infty \). Since Jensen’s Inequality is actually an equality only when there is independence of the parameter with respect to which the average is taken, one can interpret this as implying that in the limit \( n \to \infty \) the optimal controls must be asymptotically independent of the parameter with respect to which the average is taken. In other words, the set of elements over which the infimum is taken in the limiting variational problem should be independent of this parameter.

This parameter independence holds for the time variable and the state of the controlled empirical measure in the case of Sanov’s Theorem. In contrast, in Mogulskii’s Theorem one infimizes over paths with an explicit time dependence. Thus Jensen’s Inequality cannot be used there to average out the time variable. It is perhaps interesting to point out what would happen if one did apply Jensen’s Inequality in this case. One would still obtain a Laplace principle upper bound, but with the upper rate function

\[
I(\varphi) \triangleq \left\{ \begin{array}{ll}
L(\varphi(1)) & \text{if } \varphi \in \mathcal{C}([0,1]: \mathbb{R}^d) \text{ is absolutely continuous and } \varphi(0) = 0 \\
\infty & \text{for all other } \varphi \in \mathcal{C}([0,1]: \mathbb{R}^d).
\end{array} \right.
\]
This is clearly suboptimal, and it is of course related to the rate function through an application of Jensen’s Inequality.

Finally we must discuss the problem of what to do when one does not know the form of the rate function. There are two options. The first option is to use intuition and experience to guess what the crucial dependencies might be. For example, in a process-level version of Sanov’s Theorem, one could draw on the example of Mogulskii’s Theorem to conclude that time dependencies should not be averaged out. The alternative option is trial and error. If Jensen’s Inequality really simplifies the convergence analysis, then use it to average out a certain parameter, obtain an upper bound with some rate function, and consider the latter as a candidate for the true rate function. If this procedure works, then the upper bound is asymptotically tight, and the averaging is justified. If not, one must try again. This procedure often works well if it is first applied at a formal level.

The following remark discusses several features of the representation formula proved in Theorem 3.2.2. The reader may omit this remark without loss of continuity.

**Remark 3.6.1.** In the representation formula for Sanov’s Theorem given in Theorem 2.3.2, each of the controls $\nu^n_j$ depends only on the current value of the controlled process. In contrast, in the representation formula for Mogulskii’s Theorem given in Theorem 3.2.2, each of the controls $\nu^j_n$ depends on the entire past $(\bar{X}^n_1, \bar{X}^n_2, \ldots, \bar{X}^n_j)$ of the controlled process. In general, this cannot be simplified. However, for certain special cases of $h$ that are of practical importance the control problem can be rewritten so that it involves a running cost, an exit cost, and a terminal cost and so that the infimum in the definition of $V^n$ can be taken over controls $\nu^j_n$ that depend only on $\bar{X}^n_j$ and not on the entire past. In this case the optimally controlled process is a Markov process. Fix a Borel subset $A$ of $R^d$ and a real number $M$. An example of such an $h$ in Mogulskii’s Theorem is

$$h(\varphi) = \begin{cases} 
0 & \text{if } \varphi(t) \in A \text{ for all } t \in [0,1] \\
M & \text{otherwise.}
\end{cases}$$

The fact that for such special cases of $h$ the optimally controlled process is a Markov process has important implications with regard to computations. Indeed, the Markov property is used to construct relatively simple recursive algorithms for the approximation of $V^n$ [67]. Another indication of the relative simplicity of this case is that under broad conditions the limit of $V^n$ as $n \to \infty$ can be characterized in terms of the viscosity solution of a nonlinear partial differential equation defined on a finite dimensional space [54].

With the proof of Mogulskii’s Theorem and the comments just presented, we have finished our introduction of the weak convergence approach. In the next chapter we introduce some new models that generalize those already considered, and we derive for these models the corresponding representation formulas.
Chapter 4

Representation Formulas for Other Stochastic Processes

4.1 Introduction

It is convenient to divide the large deviation problems under consideration in this book into two classes: “random variable–level” problems and “process–level” problems. The first class includes the empirical measures of i.i.d. random variables, which have already been considered in Chapter 2 in the proof of Sanov’s Theorem. The second class includes the random walk models of Chapter 3 which were analyzed in the proof of Mogulskii’s Theorem. The division into two classes reflects the somewhat different role that a “time” variable plays in each of the two cases. It is possible to argue that this division into two classes is somewhat artificial. For example, the random walk model may be thought of as a random variable–level problem for the path space $C([0, 1] : \mathbb{R}^d)$. However, we find the division convenient because of the somewhat different forms of the stochastic control problems that arise. In addition, different convergence results are needed for the two classes.

The purpose of the present chapter is both to provide the representation formulas that will be needed in the rest of the book and to explore other representation formulas that arise in related but different settings. In Section 4.2 we treat the second example in the class of random variable–level problems, which is the empirical measures of Markov chains taking values in a Polish space. As the derivation of the representation formula is very similar to that of Chapter 2, this section will be brief. We will follow a similar format in Section 4.3, where the second example in the class of process–level problems is treated. This is a random walk model that generalizes the model analyzed in Mogulskii’s Theorem. The representation formulas obtained in Sections 4.2 and 4.3 will be applied in later chapters to prove the Laplace principle for the empirical measures of a Markov chain and for the random walk model. In the empirical measure case we are concerned with deviations away from the invariant measure of the underlying Markov chain. In the random walk case we are concerned with small noise asymptotics for trajectories of the stochastic processes.

This is an appropriate place to add a comment about different descriptions of Markov
processes and the resulting different representations. To begin, we observe that a Markov process can be described in various ways. A standard description is via the transition probability function for discrete–time processes or, analogously, via a martingale problem or a semigroup characterization for continuous–time processes. A second description is in terms of a dynamical system. For discrete–time processes this takes the form of a difference equation, where one specifies the distribution of the driving noises; for continuous–time processes one uses a stochastic differential equation. The various descriptions give rise to different dynamic programming equations and hence to different representations. It is important to find the representation most convenient for the large deviation problem at hand.

Luckily, there is a simple method for deciding which description is best. In the weak convergence approach the proofs of all the Laplace principle bounds are based on calculating law of large numbers limits for controlled processes. In addition, the description of the controlled process always parallels that of the original process. As a consequence, it is not at all surprising that the description which is best for a law of large numbers analysis of the original process will also be best for the weak convergence approach. For example, in order to prove the law of large numbers for ergodic phenomena as in the empirical measure problem, the description of the Markov process via the transition probability function is best. In contrast, to prove the functional law of large numbers in Mogulskii’s theorem and its generalizations, a dynamical systems description is needed.

The chapter ends with three sections dealing with topics that will not recur later in the book but are nevertheless interesting. All of these sections can be skipped. In Section 4.4 we will derive the representation formula for an extension of the random walk model having state–dependent noise. This extension encompasses many Markov–modulated models that arise in applications such as the analysis of digital data networks [83]. Although we will not consider the large deviation analysis of this class of processes, it may be carried out using the techniques introduced in this book.

So far, we have dealt with representation formulas for discrete–time processes involving bounded measurable functions $h$. In Section 4.5 we present extensions to the case where $h$ is only bounded from below or from above. The first is important because it allows one to obtain representation formulas for large deviation probabilities directly. The second is important because it allows one to study large deviation asymptotics of other interesting quantities such as functionals of escape times and quadratic functionals.

In most of this book the focus is on representation formulas that are associated with discrete–time processes. In order to introduce the last section of this chapter, we recall Breiman’s felicitous observation, expressed in the preface of [14], that probability theory has a right hand and a left hand. In the present book as in most theoretical texts in probability, the right hand, referring to the precise formulation of theorems and their rigorous proofs, predominates. In the last section of Chapter 4 we present a left–handed treatment of representation formulas for continuous–time processes. Rather than focus on the careful formulation of theorems, we attempt in this section to demonstrate the versatility of the ideas presented so far by seeing how they extend to this new situation. Namely, given a continuous–time Markov process $\{X(t), t \in [0, 1]\}$, we use the procedure of previous sections to derive a representation formula for a suitably sampled, discrete–time
version of this process and then formally take the limit as the sampling parameter tends to 0. The resulting limit should give a representation formula for the original process. We then specialize the formula so obtained to the two important cases of diffusion processes and jump Markov processes. Some remarks on proving the representation formulas for continuous-time processes are included, and references are given for proofs in particular cases.

4.2 The Representation Formula for the Empirical Measures of a Markov Chain

In this section we will exhibit the representation formula for a basic example in the class of random variable-level problems. The example of this section is simply the Markovian analogue of the i.i.d. case that was analyzed in Chapter 2. This representation will be used in Chapters 8 and 9 to prove the Laplace principle for the empirical measures under various assumptions on the transition probability function.

Let $\mathcal{S}$ be a Polish space. We consider a Markov chain $\{X_j, j \in \mathbb{N}_0\}$ that is defined on a probability space $(\Omega, \mathcal{F}, P)$, takes values in $\mathcal{S}$, and has stationary transition probabilities. The transition probability function of the Markov chain is denoted by $p(x, dy)$. For each $n \in \mathbb{N}$, $\omega \in \Omega$, and Borel subset $A$ of $\mathcal{S}$, we define the empirical measure, or the normalized occupation measure up to time $n - 1$, by

$$L^n(\omega, A) = L^n(A) = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{X_j(\omega)}(A).$$

Let $X_0 = x$ be the initial condition for the Markov chain $\{X_j, j \in \mathbb{N}_0\}$ and denote by $E_x$ expectation conditioned on $X_0 = x$. Our goal is to obtain a stochastic control representation for

$$W^n(x) = -\frac{1}{n} \log E_x \{\exp[-n \, h(L^n)]\},$$

where $h$ is a bounded measurable function mapping $\mathcal{P}(\mathcal{S})$ into $\mathbb{R}$. We condition on $x$ since in principle the asymptotic behavior may depend on this initial condition. When the Laplace principle for the empirical measures is proved in Chapters 8 and 9, we will require that in addition $h$ be continuous. However, the continuity is not needed in order to derive the representation formula.

The derivation of a representation formula for $W^n(x)$ in the present case of a Markov chain differs only in notation from the derivation in the analogous i.i.d. case given in Chapter 2. Because of this, some details will be omitted. We will describe the empirical measures in terms of a Markov process, derive the appropriate dynamic programming equation, and then read off the associated stochastic control problem. From this it will be easy to write down the representation formula.

$L^n$ takes values in $\mathcal{P}(\mathcal{S})$. As in Chapter 2, before the stochastic control problem is formulated, we must rewrite $L^n$ in terms of a Markov dynamical system. To this end,
we introduce a sequence of empirical subprobability measures by setting $L_0^n = 0$ and by defining for $j = 0, 1, \ldots, n - 1$

$$L^n_{j+1} = L^n_j + \frac{1}{n} \delta_{X_j}.$$ 

With this definition $L^n = L^n_0$. Let $\mathcal{M}(\mathcal{S})$ denote the space of subprobability measures on $\mathcal{S}$. With the weak topology, $\mathcal{P}(\mathcal{S})$ and $\mathcal{M}(\mathcal{S})$ are metrizable as Polish spaces [Section A.3], and in each case convergence with respect to the underlying metric is equivalent to the weak convergence of measures. $L^n_j$ takes values in $\mathcal{M}_{j/n}(\mathcal{S}) = \{\mu \in \mathcal{M}(\mathcal{S}) : \mu(\mathcal{S}) = j/n\}$.

In contrast to the situation in Chapter 2, in the present setting the quantities $\{L^n_j, j = 0, 1, \ldots, n\}$ do not, in general, form a Markov chain. Consequently, we consider the two-component Markov chain $\{(X_j, L^n_j), j = 0, 1, \ldots, n\}$. The state space of this chain at time $j$ is $\mathcal{S} \times \mathcal{M}_{j/n}(\mathcal{S})$ and its transition probability function is

$$P\{ (X_{i+1}, L^n_{i+1}) \in dy \times d\nu | (X_i, L^n_i) = (x, \mu) \} = p(x, dy) \cdot \delta_{\frac{\mu}{n}, \delta_x}(d\nu)$$

for $(x, \mu) \in \mathcal{S} \times \mathcal{M}_{i/n}(\mathcal{S})$.

Take $i \in \{0, 1, \ldots, n\}$ and let $(x, \mu)$ be any point in $\mathcal{S} \times \mathcal{M}_{i/n}(\mathcal{S})$. Again by analogy with Chapter 2, we introduce

$$W^n(i, x, \mu) = - \frac{1}{n} \log E_{i,x,\mu}[\exp[-n h(L^n)]],$$

where $E_{i,x,\mu}$ denotes expectation conditioned on $(X_i, L^n_i) = (x, \mu)$, and $h$ is the same bounded measurable function that appears in the definition of $W^n(x)$. Of course $W^n(0, x, 0) = W^n(x)$. By using the Markov property, we obtain

$$\exp[-n W^n(i, x, \mu)] = E_{i,x,\mu}[\exp[-n h(L^n)]] = E_{i,x,\mu}[E_{i+1,x_{i+1},L^n_{i+1}}[\exp[-n h(L^n)]]] = E_{i,x,\mu}[\exp[-n W^n(i+1, x_{i+1}, L^n_{i+1})]] = \int_{\mathcal{S}} \exp[-n W^n(i+1, y, \mu + \frac{1}{n} \delta_x)] p(x, dy).$$

Thus

$$W^n(i, x, \mu) = - \frac{1}{n} \log \int_{\mathcal{S}} \exp[-n W^n(i+1, y, \mu + \frac{1}{n} \delta_x)] p(x, dy),$$

and part (a) of Proposition 1.4.2 gives

$$W^n(i, x, \mu) = \inf_{\nu \in \mathcal{P}(\mathcal{S})} \left\{ \frac{1}{n} R(\nu(\cdot) || p(x, \cdot)) + \int_{\mathcal{S}} W^n(i+1, y, \mu + \frac{1}{n} \delta_x) \nu(dy) \right\}. \quad (4.2)$$

In addition, we have the terminal condition

$$W^n(n, x, \mu) = h(\mu).$$

Equation (4.2) is the dynamic programming equation from which the desired representation formula will be obtained. In order for the representation formula to be valid, we require that the Attainment Condition 1.5.1 be met, and this is taken care of by the following lemma. As always, the lemma is a consequence of part (b) of Proposition 1.4.2.
4.2. EMPIRICAL MEASURES OF A MARKOV CHAIN

Lemma 4.2.1. For \( i \in \{0, 1, \ldots, n - 1\} \), \( x \in \mathcal{S} \) and \( \mu \in \mathcal{M}_{i/n}(\mathcal{S}) \), the infimum in equation (4.2) is uniquely attained at the measure \( \nu^n_i(dy|x, \mu) = \nu^n_i(dy|x, \mu) \) that assigns to a Borel subset \( A \) of \( \mathcal{S} \) the measure

\[
\nu^n_i(A|x, \mu) = \frac{1}{\int_A \exp \left[ -n W^n \left( i + 1, y, \mu + \frac{1}{n} \delta_x \right) \right] p(x, dy) \cdot \frac{1}{\int_{\mathcal{S}} \exp \left[ -n W^n \left( i + 1, y, \mu + \frac{1}{n} \delta_x \right) \right] p(x, dy)}.
\]

For each \( i \), \( \nu^n_i(dy|x, \mu) \) is a stochastic kernel on \( \mathcal{S} \) given \( \mathcal{S} \times \mathcal{M}_{i/n}(\mathcal{S}) \).

We can now define the associated stochastic control problem in terms of its dynamics and cost structure. Since the only difference between the present situation and that of Chapter 2 is that the sequence \( \{X_j, j = 0, 1, \ldots, n - 1\} \) here forms a Markov chain rather than an i.i.d. sequence, it is not surprising that the two representation are very similar. The only significant difference is in the dependence of the control at time \( j \) upon the controlled process at time \( j \). This dependence is due to the fact that in the present situation we had to augment \( L^n_j \) by \( X_j \) in order to obtain a Markov process. The form of the dependence of the controls follows directly from the dynamic programming equation (4.2).

**Dynamics.** For \( n \in \mathbb{N} \) the control that is applied at time \( j \in \{0, 1, \ldots, n - 1\} \) is a stochastic kernel on \( \mathcal{S} \) given \( \mathcal{S} \times \mathcal{M}_{j/n}(\mathcal{S}) \) which is denoted by \( \nu^n_j = \nu^n_j(dy|x, \mu) \). A sequence of controls \( \{\nu^n_j, j = 0, 1, \ldots, n - 1\} \) is called an **admissible control sequence**. At time \( j \in \{0, 1, \ldots, n\} \) the controlled process takes values in \( \mathcal{S} \times \mathcal{M}_{j/n}(\mathcal{S}) \) and is denoted by \( (X^n_j, \bar{L}^n_j) \). We define \( X^n_0 = X^n_0 = x \), and for \( j \in \{0, 1, \ldots, n - 1\} \) we specify the conditional distribution of \( X^n_{j+1} \) by

\[
P \{X^n_{j+1} \in dy \mid (X^n_j, \bar{L}^n_j), i = 0, 1, \ldots, j \} = \nu^n_j(dy|X^n_j, \bar{L}^n_j).
\]

The measures \( \bar{L}^n_j \) evolve according to \( \bar{L}^n_0 = L^n_0 = 0 \) and

\[
\bar{L}^n_{j+1} = \bar{L}^n_j + \frac{1}{n} \delta_{X^n_j}.
\]

We write \( \bar{L}^n \) instead of \( \bar{L}^n_n \).

**Cost structure.** For \( n \in \mathbb{N} \) and \( x \in \mathcal{S} \) we define the minimal cost function

\[
V^n(x) = \inf_{\{\nu^n_j\}} \bar{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R \left( \nu^n_j(\cdot|X^n_j, \bar{L}^n_j) \| p(X^n_j, \cdot) \right) + h(\bar{L}^n) \right\}.
\]

(4.3)

The infimum is taken over all admissible control sequences \( \{\nu^n_j\} \), \( \bar{E}_x \) denotes expectation conditioned on \( (X^n_0, \bar{L}^n_0) = (x, 0) \in \mathcal{S} \times \mathcal{M}(\mathcal{S}) \), and \( \{(X^n_j, \bar{L}^n_j)\} \) is the controlled process that is associated with a particular admissible control sequence \( \{\nu^n_j\} \). We notice again
a general feature of this control problem. The sequence \( \{ L_j^* \} \) of empirical subprobability measures of the original Markov chain is replaced by a sequence \( \{ L^*_j \} \) of empirical subprobability measures of controlled random variables \( X^n_j \), the distributions of which are specified by an admissible control sequence \( \{ \nu^n_j \} \). Each control \( \nu^n_j \) gives rise to a corresponding term in the running cost in the minimal cost function.

We next state the representation theorem for the empirical measures of a Markov chain. The representation theorem is an immediate consequence of Theorem 1.5.2 and of Lemma 4.2.1, which shows that the Attainment Condition is satisfied.

**Theorem 4.2.2.** Let \( h \) be a bounded measurable function mapping \( P(S) \) into \( \mathbb{R} \). Then for all \( n \in \mathbb{N} \) and \( x \in S \)

\[
W^n(x) = -\frac{1}{n} \log E_x \{ \exp \{ -n h(L^n) \} \}
\]

equals the minimal cost function \( V^n(x) \) defined in formula (4.3).

The representation formula in this theorem will be used in Chapters 8 and 9 to prove the Laplace principle for the empirical measures of a Markov chain.

### 4.3 The Representation Formula for a Random Walk Model

In the previous section a representation formula was obtained in a setting that generalized that of Chapter 2, which dealt with Sanov’s Theorem. In this section a representation formula will be obtained in a setting that generalizes that of Chapter 3, which treated Mogulskii’s Theorem. The only difference between the present section and Chapter 3 is in the definition of the random walk model. In particular, the proof of the representation formula here is virtually identical to that in Chapter 3. Consequently, we will be content just to describe the model and state the representation formula.

The new random walk model is defined by a stochastic kernel \( \mu(dy|x) \) on \( \mathbb{R}^d \) given \( \mathbb{R}^d \). Let \( \{ v_j(x), j \in \mathbb{N}_0, x \in \mathbb{R}^d \} \) be a sequence of i.i.d. random vector fields on \( \mathbb{R}^d \) having for each \( x \in \mathbb{R}^d \) and all \( j \in \mathbb{N}_0 \) the common distribution

\[
P \{ v_j(x) \in dy \} = \mu(dy|x).
\]

Thus there exists a probability space \((\Omega, \mathcal{F}, P)\) such that each \( v_j(x) = v_j(x, \omega) \) is a measurable function mapping \( \mathbb{R}^d \times \Omega \) into \( \mathbb{R}^d \) and for all nonnegative integers \( j \neq k \) and all points \( x_1 \) and \( x_2 \) in \( \mathbb{R}^d \) the random vectors \( v_j(x_1) \) and \( v_k(x_2) \) are independent. Such a sequence of i.i.d. random vector fields is easily constructed. For example, when \( d = 1 \) define

\[
v_j(x) = (F_x)^{-1}(U_j),
\]

where \( (F_x)^{-1} \) is the standard generalized inverse of the nondecreasing function \( F_x(y) = \mu((-\infty, y]|x) \) and \( \{ U_j, j \in \mathbb{N}_0 \} \) is an i.i.d. sequence of uniformly distributed random
variables on $[0, 1]$. When $d \geq 2$, this construction can be extended via conditioning. For $x \in \mathbb{R}^d, n \in \mathbb{N}$, and $j \in \{0, 1, \ldots, n-1\}$, we define random vectors $X^n_0 \equiv x$ and

$$X^n_{j+1} \equiv X^n_j + \frac{1}{n} v_j(X^n_j). \quad (4.4)$$

Since the random vector fields $v_j(x)$ are i.i.d., the random vectors $\{X^n_j, j = 0, 1, \ldots, n\}$ form a Markov chain with stationary transition probabilities. In equation (4.4) large values of $n \in \mathbb{N}$ correspond to small noise.

We are interested in the Laplace principle for the stochastic process $X^n = \{X^n(t), t \in [0, 1]\}$ defined by

$$X^n(t) \equiv X^n_j + \left( t - \frac{j}{n} \right) v_j(X^n_j) \quad \text{for } t \in [j/n, (j+1)/n], \ j = 0, 1, \ldots, n-1. \quad (4.5)$$

$X^n$ is the piecewise linear interpolation of the discrete–time process $\{X^n_j\}$, and it takes values in $C([0, 1] : \mathbb{R}^d)$. As in Chapter 3, the interval $[0, 1]$ has been chosen for notational convenience only. Analogous results can be obtained for any interval of the form $[0, T], T \in (0, \infty)$.

The random walk $X^n$ reduces to the random walk studied in Mogulskii’s Theorem if the stochastic kernel $\mu(dy|x)$ is independent of $x$ and thus is a probability measure on $\mathbb{R}^d$. Since in the case of Mogulskii’s Theorem the random walk is translation invariant, we chose the initial condition $X^n(0) = 0$. In the present case, where in general $\mu(dy|x)$ varies with $x$, we allow the random walk to start at any point in $\mathbb{R}^d$. As we will see, when $\mu(dy|x)$ varies with $x$, the rate function takes a more complicated form than in Mogulskii’s Theorem.

Processes of the type defined by equation (4.5) arise in many recursive procedures that are common in statistics, adaptive control, and various parameter tracking algorithms. They also appear quite naturally as Euler-type approximations for a large class of continuous–time Markov processes. In fact, the typical procedure for proving a large deviation result for continuous–time Markov processes is first to prove the large deviation result for approximating processes of the form given by equation (4.5) and then to verify estimates that allow one to extend the result to the original process. An example of this procedure will be given in Chapter 10 of the book.

The representation formulas to be derived in this section do not require any additional assumptions on the smoothness of the function mapping $x \in \mathbb{R}^d \mapsto \mu(\cdot|x) \in \mathcal{P}(\mathbb{R}^d)$. However, the proofs of the Laplace principles do require additional smoothness assumptions. When the measures $\mu(\cdot|x)$ are continuous functions of $x$ in all of $\mathbb{R}^d$, we speak of a random walk with continuous statistics. The Laplace principle for such a process is given in Theorem 6.3.3. When the measures $\mu(\cdot|x)$ do not depend continuously on $x$ in all of $\mathbb{R}^d$, we speak of a random walk with discontinuous statistics. A Laplace principle for a particular such process is given in Theorem 7.2.3. While the continuous case is obviously the more common of the two, random walks with discontinuous statistics are also quite common, arising naturally in queueing, communication, and control applications [41].

Let $h$ be any bounded measurable function mapping $\mathcal{C}([0, 1] : \mathbb{R}^d)$ into $\mathbb{R}$. Our goal
is to obtain a stochastic control representation formula for

\[ W^n(x) \doteq -\frac{1}{n} \log E_x \{ \exp [-n h(X^n)] \}, \]

where \( E_x \) denotes expectation conditioned on \( X^n_0 = x \) and \( X^n = \{ X^n(t), t \in [0, 1] \} \) is the random walk defined in equation (4.5). We emphasize that as in Chapter 3, in general \( h(X^n) \) can be an arbitrary bounded measurable function of the entire trajectory \( X^n \) or, equivalently, of the Markov chain \( \{ X^n_j \} \). In fact, the only difference between the model that is presently under consideration and the model in Chapter 3 is the additional state dependence here; we now use the measure \( \mu(\cdot|X^n) \) to define the distribution of \( n(X^n_{j+1} - X^n_j) \) rather than the single measure \( \rho \). Despite this additional generality, the sequence \( \{ X^n_j \} \) is still a Markov chain, and this means that the form of the representation formula and the proof of its validity are almost exactly the same here as they were in Chapter 3. As a consequence, we now just state the representation formula without deriving it. We first specify the associated stochastic control problem, the form of which should come as no surprise. From our experience in previous sections we expect that the controlled process is obtained from the original process described in (4.4) by replacing the “noise” \( \nu_j(X^n_j) \) by a controlled random vector \( \tilde{Y}^n_j \). The conditional distribution of the controlled random vector should be specified in terms of a control that gives rise to a corresponding term in the minimal cost function. This is indeed the case.

**Dynamics.** For \( n \in \mathbb{N} \) the control that is applied at time \( j \in \{0, 1, \ldots, n - 1\} \) is a stochastic kernel \( \nu^n_j \) on \( \mathbb{R}^d \) given \( (\mathbb{R}^d)^{j+1} \). The dependence of \( \nu^n_j \) upon \( (x_0, x_1, \ldots, x_j) \in (\mathbb{R}^d)^{j+1} \) is indicated by writing \( \nu^n_j = \nu^n_j(dy|x_0, x_1, \ldots, x_j) \). In this case an **admissible control sequence** is a sequence \( \{ \nu^n_j, j = 0, 1, \ldots, n - 1 \} \) of such controls. The controlled process takes values in \( \mathbb{R}^d \) and is denoted by \( \{ X^n_j, j = 0, 1, \ldots, n \} \). The evolution of the controlled process is defined as follows. We set \( \tilde{X}^n_0 = X^n_0 = x \), and for \( j \in \{0, 1, \ldots, n - 1 \} \) we let \( \tilde{Y}^n_j \) be a random vector with the conditional distribution

\[ \tilde{P}_x \{ \tilde{Y}^n_j \in dy|\tilde{X}^n_0, \tilde{X}^n_1, \ldots, \tilde{X}^n_j \} = \nu^n_j(dy|\tilde{X}^n_0, \tilde{X}^n_1, \ldots, \tilde{X}^n_j). \]

We then recursively define

\[ \tilde{X}^n_{j+1} \doteq \tilde{X}^n_j + \frac{1}{n} \tilde{Y}^n_j. \]

Finally, the **controlled random walk** \( \tilde{X}^n = \{ \tilde{X}^n(t), t \in [0, 1] \} \) is defined by

\[ \tilde{X}^n(t) \doteq \tilde{X}^n + \left( t - \frac{j}{n} \right) \tilde{Y}^n_j \quad \text{for } t \in \left[ j/n, (j + 1)/n \right], \quad j = 0, 1, \ldots, n - 1. \]

For all \( n \in \mathbb{N} \) the controlled random variables \( \tilde{X}^n_j \) and \( \tilde{Y}^n_j \) are defined on a common probability space \( (\tilde{\Omega}, \mathcal{F}, \tilde{P}_x) \).

**Cost structure.** We define the minimal cost function

\[ V^n(x) \doteq \inf_{\{ \nu^n_j \}} \tilde{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot)\mu(\cdot|\tilde{X}^n_j)) + h(\tilde{X}^n) \right\}. \] (4.6)
4.4. MODEL WITH STATE–DEPENDENT NOISE

Here \( v^n_j(\cdot) = v^n_j(\cdot|\bar{X}_0^n, \ldots, \bar{X}_n^n) \), the infimum is taken over all admissible control sequences \( \{v^n_j\} \); \( \bar{E}_x \) denotes expectation with respect to \( P_x \), and \( \{\bar{X}_n^n\} \) is the sequence of controlled random vectors that are associated with a particular admissible control sequence \( \{v^n_j\} \).

We can now state the representation formula.

**Theorem 4.3.1.** Let \( h \) be a bounded measurable function mapping \( C([0,1]:\mathbb{R}^d) \) into \( \mathbb{R} \). Then for all \( n \in \mathbb{N} \) and \( x \in \mathbb{R}^d \)

\[
W^n(x) \triangleq -\frac{1}{n} \log E_x \{ \exp[-n h(X^n)] \}
\]

equals the minimal cost function \( V^n(x) \) defined in equation (4.6).

The representation formula in this theorem will be used in Chapters 6 and 7 to prove the Laplace principle for the random walk model. In the next section we consider a generalization of the random walk model that includes additional dependencies.

4.4 The Representation Formula for a Random Walk Model with State–Dependent Noise

In Chapters 5 through 10 of this book the representation formulas proved in the previous two sections will be applied to the study of large deviation phenomena. The two broad classes of problems on which we will focus are small noise asymptotics for trajectories of Markov processes and deviations from the invariant measure for empirical measures of a Markov chain. Beyond these specific problems, the methods that we will introduce in these later chapters are broadly applicable and can be extended in many directions. This section highlights an interesting and important example of other processes that one might consider. Namely, the representation formula will be derived for an extension of the random walk model just considered, an extension that encompasses a large number of the Markov–modulated processes arising in queueing theory [15] and in communication theory [7, 68] and also includes many recursive algorithms with state–dependent noise arising in the study of adaptive algorithms [65].

We first describe the extension of the model and then obtain the stochastic control representation formula. The derivation of the formula will be notionally complicated, but straightforward. In fact, except for the more complicated form of the state of the process, which is due to the presence of a “modulating” process, the derivation is essentially the same as that in Chapter 3.

**An Extension of the Random Walk Model.** Let \( S \) be a Polish space. The new random walk model is defined in terms of a stochastic kernel \( \mu(dy|x,\xi) \) on \( \mathbb{R}^d \) given \( \mathbb{R}^d \times S \). We work with a sequence \( \{v_j(x,\xi), j \in \mathbb{N}_0, x \in \mathbb{R}^d, \xi \in S\} \) of i.i.d. random vector fields on \( \mathbb{R}^d \) having for each \( x \in \mathbb{R}^d, \xi \in S \), and all \( j \in \mathbb{N}_0 \) the common distribution

\[
P\{v_j(x,\xi) \in dy\} = \mu(dy|x,\xi).
\]
Finally, for each $x \in \mathbb{R}^d$ let $p(\xi, d\zeta|x)$ be a stochastic kernel on $\mathcal{S}$ given $\mathcal{S} \times \mathbb{R}^d$. We think of this stochastic kernel as a collection of transition probability functions on $\mathcal{S}$ indexed by $x \in \mathbb{R}^d$. For each $n \in \mathbb{N}$, $x \in \mathbb{R}^d$, and $\xi \in \mathcal{S}$, a sequence of random variables $\{(X^n_i, \Xi^n_j), j = 0, 1, ..., n\}$ taking values in $\mathbb{R}^d \times \mathcal{S}$ is then defined by setting $X^n_0 \triangleq x, \Xi^n_0 \triangleq \xi$, and

$$X^n_{j+1} \triangleq X^n_j + \frac{1}{n} v_j(X^n_j, \Xi^n_{j+1}). \quad (4.7)$$

For $j \in \{0, 1, ..., n-1\}$ the conditional distribution of $\Xi^n_{j+1}$ is given by

$$P_{x, \xi}\{\Xi^n_{j+1} \in d\zeta|v_k(\cdot, \cdot), k \in \mathbb{N}_0; (X^n_i, \Xi^n_i), i = 0, 1, ..., j\} = p(\Xi^n_j, d\zeta|X^n_j).$$

In order to give a concrete interpretation of such a process, we consider an adaptive stochastic system in which the process $\{\Xi^n_j, j = 0, 1, ..., n\}$ represents the state of the system, the process $\{X^n_j, j = 0, 1, ..., n\}$ being a sequence of “control” parameters whose values influence the evolution of the state through the transition probability function $p(\xi, dy|x)$. At time $j$ these control parameters are adjusted, through equation (4.7), in an “on line” fashion as a function of the value $X^n_j$ and the current state of the system $\Xi^n_{j+1}$.

The dynamics of this adjustment process often include noise, and they are modeled here by the random vector fields $\{v_j(x, \xi)\}$. The design engineer can to some degree choose these dynamics, and she usually does so with the goal that once set into motion, the dynamics will steer the process $X^n$ towards a favorable operating point for the system. A large value of $n$ coincides with a desire to put relatively little importance on any single measurement $v_j(X^n_j, \Xi^n_{j+1})$, which is a common situation. For a large number of examples of such processes and a discussion of the important role that they play in communication problems, we refer the reader to [7, 68].

Suppose that the random vectors $\{X^n_j, j = 0, 1, ..., n\}$ are interpolated into a piecewise linear continuous–time process in the same way as in the previous section; i.e.,

$$X^n(t) \triangleq X^n_j + \left( t - \frac{j}{n} \right) v_j(X^n_j, \Xi^n_{j+1}) \quad \text{for } t \in [j/n, (j + 1)/n], \ j = 0, 1, ..., n - 1.$$ 

In order to study the large deviation behavior of the processes $X^n \triangleq \{X^n(t), t \in [0, 1]\}$, one can compute, in accordance with the Laplace principle, the asymptotic behavior of

$$W^n(x, \xi) \triangleq -\frac{1}{n} \log E_{x, \xi}\{\exp \left[ -n h(X^n) \right]\}. \quad (4.8)$$

In this formula $E_{x, \xi}$ denotes expectation with respect to $P_{x, \xi}$ and $h$ is any bounded continuous function mapping $\mathcal{C}([0, 1]: \mathbb{R}^d)$ into $\mathbb{R}$. The verification of the Laplace principle combines features of the process–level problem for random walks, carried out in Chapters 5 and 6, and features of the empirical measure problem, carried out in Chapter 8. A previous paper on these problems is [37]. Of course, one could also interpolate, in a similar way, the state variables $\{\Xi^n_j, j = 0, 1, ..., n\}$ into a continuous–time process $\Xi^n = \{\Xi^n(t), t \in [0, 1]\}$ taking values in $\mathcal{S}$ and study the large deviation behavior of the joint processes $\{(X^n, \Gamma^n), n \in \mathbb{N}\}$, where $\Gamma^n = \{\Gamma^n_t(dy), t \in [0, 1]\}$ and $\Gamma^n_t(dy) \triangleq \int_0^t \delta_{\Xi^n_s}(dy) \, ds$ is the normalized occupation process of $\Xi^n$. 

4.4. MODEL WITH STATE-DEPENDENT NOISE

We now specify the stochastic control problem whose minimal cost function gives a representation for the function $W^n(x, \xi)$ defined in equation (4.8).

The Dynamic Programming Equation. Since we are interested in a function of the entire trajectory $X^n$, we follow the example of Chapter 3. There is also an additional complication due to the fact that there are two sources of randomness in the model of this section: one due to the increments $v_j(X^n_j, \Xi^n_{j+1})$ and another due to the modulating process $\Xi^n_{j+1}$ that appears in these increments. While a perfectly good representation can be obtained by considering these noises together, a more suggestive representation can be found when we consider their effects separately. For this reason we introduce for $i \in \{0, 1, \ldots, n\}$ and points $x_0, \ldots, x_i$ in $\mathbb{R}^d$ and $\xi, \xi_{i+1}$ in $\mathcal{S}$

$$W^n(i, \{x_0, \ldots, x_i\}, \xi) = -\frac{1}{n} \log E_{x_0, \xi} \{ \exp[-n h(X^n)] | (\{X^n_0, \ldots, X^n_i\}, \Xi^n) = (\{x_0, \ldots, x_i\}, \xi) \}$$

and

$$U^n(i, \{x_0, \ldots, x_i\}, \xi_{i+1}) = -\frac{1}{n} \log E_{x_0, \xi} \{ \exp[-n h(X^n)] | (\{X^n_0, \ldots, X^n_i\}, \Xi^n_{i+1}) = (\{x_0, \ldots, x_i\}, \xi_{i+1}) \}.$$

$W^n(0, \{x\}, \xi)$ equals the function $W^n(x, \xi)$ defined in (4.8). By using the Markov property as in previous sections, we obtain the recursive formulas

$$W^n(i, \{x_0, \ldots, x_i\}, \xi) = -\frac{1}{n} \log \int_{\mathcal{S}} \exp[-n U^n(i, \{x_0, \ldots, x_i\}, \xi)] p(\xi, d\zeta|x_i)$$

and

$$U^n(i, \{x_0, \ldots, x_i\}, \xi_{i+1}) = -\frac{1}{n} \log \int_{\mathbb{R}^d} \exp\left[-n W^n \left( i + 1, \left\{ x_0, \ldots, x_i, x_i + \frac{1}{n}y \right\}, \xi_{i+1} \right) \right] \mu(dy|x_i, \xi_{i+1}).$$

In this way we have separated out the introduction of randomness according to type. This procedure is equivalent to introducing a new time variable that runs from 0 to $2n$, with even values corresponding to $W^n$ and transitions determined by $p(\xi, \cdot|x)$ and odd values corresponding to $U^n$ and transitions determined by $\mu(\cdot|x, \xi)$.

We now apply part (a) of Proposition 1.4.2 to each of these formulas, obtaining

$$W^n(i, \{x_0, \ldots, x_i\}, \xi) = \inf_{\nu_1 \in \mathcal{P}(\mathcal{S})} \left\{ \frac{1}{n} R(\nu_1(\cdot)) \| p(\xi_1, \cdot|x_i) \| + \int_{\mathcal{S}} U^n(i, \{x_0, \ldots, x_i\}, \zeta) \nu_1(d\zeta) \right\}$$

(4.9)

and

$$U^n(i, \{x_0, \ldots, x_i\}, \xi_{i+1}) = \inf_{\nu_2 \in \mathcal{P}(\mathbb{R}^d)} \left\{ \frac{1}{n} R(\nu_2(\cdot)) \| \mu(\cdot|x_i, \xi_{i+1}) \| + \int_{\mathbb{R}^d} W^n \left( i + 1, \left\{ x_0, \ldots, x_i, x_i + \frac{1}{n}y \right\}, \xi_{i+1} \right) \nu_2(dy) \right\}.$$

(4.10)
We also have the terminal condition
\[ W^n(n, \{x_0, \ldots, x_n\}, \xi_n) = h(x^n), \]
where \( x^n = \{x^n(t), t \in [0, 1]\} \) is the piecewise linear path defined by
\[ x^n(t) = x_j + (nt - j)(x_{j+1} - x_j) \text{ for } t \in [j/n, (j + 1)/n], j = 0, 1, \ldots, n - 1. \]

**The Stochastic Control Problem.** At this point the reader may wish to test his facility for reading the representation formula directly from the dynamic programming equations (4.9) and (4.10). From our experience in Chapters 2 and 3, we expect the form of the controlled process to parallel that of the original uncontrolled process. Hence, let us suppose that in the present situation the controlled process is a discrete-time process taking values in \( \mathbb{R}^d \times \mathcal{S} \) and denoted by \( \{(\tilde{X}^n_j, \Xi^n_j), j = 0, 1, \ldots, n\} \). Since we have identified two ways in which randomness enters the process, we expect that at each time \( j \) there will be two controls \( \nu^n_{1,j} \) and \( \nu^n_{2,j} \) giving the distributions of each of the controlled random variables that replace these noises. We also expect that there will be corresponding contributions to the running cost. The control \( \nu^n_{1,j} \), which corresponds to \( \nu_1 \) in (4.9), will be a stochastic kernel on \( \mathcal{S} \) given \( (\mathbb{R}^d)^{j+1} \times \mathcal{S} \). We denote it by \( \nu^n_{1,j}(dy|\tilde{X}^n_0, \ldots, \tilde{X}^n_j, \Xi^n_j) \). Similarly, the control \( \nu^n_{2,j} \), which corresponds to \( \nu_2 \) in (4.10), will be a stochastic kernel on \( \mathbb{R}^d \) given \( (\mathbb{R}^d)^{j+1} \times \mathcal{S} \). We denote it by \( \nu^n_{2,j}(dy|\tilde{X}^n_0, \ldots, \tilde{X}^n_j, \Xi^n_{j+1}) \).

A sequence of controls \( \{(\nu^n_{1,j}, \nu^n_{2,j}), j = 0, 1, \ldots, n - 1\} \) is called an **admissible control sequence**.

We now specify the evolution of the controlled process. We set \( \Xi^n_0 = \Xi^n_0 = \xi \) and following equation (4.9), define the conditional distribution of \( \Xi^n_{j+1} \) by
\[ \tilde{P}\{\Xi^n_{j+1} \in d\zeta|\tilde{X}^n_0, \ldots, \tilde{X}^n_j, \Xi^n_0, \ldots, \Xi^n_j\} \equiv \nu^n_{1,j}(d\zeta|\tilde{X}^n_0, \ldots, \tilde{X}^n_j, \Xi^n_j). \]
We also set \( \tilde{X}^n_0 = X^n_0 = x \). According to equation (4.10), we define
\[ \tilde{X}^n_{j+1} = \tilde{X}^n_j + \frac{1}{n}\tilde{Y}^n_j, \]
leaving \( \tilde{Y}^n_j \) have the conditional distribution
\[ \tilde{P}\{\tilde{Y}^n_j \in dy|\tilde{X}^n_0, \ldots, \tilde{X}^n_j, \Xi^n_0, \ldots, \Xi^n_j\} \equiv \nu^n_{2,j}(dy|\tilde{X}^n_0, \ldots, \tilde{X}^n_j, \Xi^n_{j+1}). \]

Finally, we define \( \tilde{X} = \{\tilde{X}^n(t), t \in [0, 1]\} \) to be the piecewise linear interpolation of the random vectors \( \{\tilde{X}^n_j, j = 0, 1, \ldots, n\} \) given by
\[ \tilde{X}^n(t) = \tilde{X}^n_j + \left( t - \frac{j}{n} \right) \tilde{Y}^n_j \text{ for } t \in [j/n, (j + 1)/n], j = 0, 1, \ldots, n - 1. \]

The definition of the stochastic control problem is complete once we specify the minimal cost function. This is defined by
\[
V^n(x, \xi) \equiv \inf_{\nu^n_{1,j}, \nu^n_{2,j}} E_x,\xi \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \left[ R\left( \nu^n_{1,j}(\cdot)|\tilde{X}^n_j, \Xi^n_j \right) \right] \right\}.
\]
4.5. EXTENSIONS TO UNBOUNDED FUNCTIONS

Here \( \nu^n_{i,j}(-) = \nu^n_{i,j}(-|\bar{X}^n_0, \ldots, \bar{X}^n_j, \bar{Z}^n_{j+1}) \) and \( \nu^n_{2,j}(-) = \nu^n_{2,j}(-|\bar{X}^n_0, \ldots, \bar{X}^n_n, \bar{Z}^n_{j+1}) \), the infimum is taken over all admissible control sequences \( \{\nu^n_{i,j}, \nu^n_{2,j}\} \), \( \tilde{E}_{x,\xi} \) denotes expectation conditioned on \( (\tilde{X}^n_0, \tilde{Z}^n_0) = (x, \xi) \), \( \{\tilde{X}^n_j, \tilde{Z}^n_j\} \) is the controlled process that is associated with a particular admissible control sequence \( \{\nu^n_{i,j}, \nu^n_{2,j}\} \), and \( h \) is the same function that appears in the definition (4.8) of \( \tilde{W}^n(x, \xi) \).

We can now state the representation formula. The proof is analogous to those in Chapters 2 and 3 and is omitted.

**Theorem 4.4.1.** Let \( h \) be a bounded measurable function mapping \( C([0,1] : \mathbb{R}^d) \) into \( \mathbb{R} \). Then for all \( n \in \mathbb{N} \), \( x \in \mathbb{R}^d \), and \( \xi \in \mathbb{R}^d \)

\[
\tilde{W}^n(x, \xi) \triangleq -\frac{1}{n} \log \tilde{E}_{x,\xi}\{\exp[-n h(X^n)]\}
\]

equals the minimal cost function \( V^n(x, \xi) \) defined in equation (4.11).

The discussion of the random walk model with state–dependent noise is now complete. In the next section we turn to another class of representation formulas that have potentially important applications but will not be used later in the book.

4.5 Extensions to Unbounded Functions

Let \( \{Z^n, n \in \mathbb{N}_0\} \) be a sequence of random variables whose large deviation properties are of interest. So far, when deriving representation formulas for quantities of the form

\[
\tilde{W}^n \triangleq -\frac{1}{n} \log \tilde{E}\{\exp[-n h(Z^n)]\},
\]

we have restricted attention to the case where \( h \) is bounded and measurable. According to Theorem 1.2.3, representation formulas for this class of functions are sufficient if one is interested in proving the Laplace principle. However, it is useful for some purposes to have representation formulas also when the measurable function \( h \) is only bounded from below or from above. A particularly interesting class of functions \( h \) that are bounded from below are those of the form \( h(z) \triangleq \infty 1_{A^c}(z) \), where \( A \) is a Borel subset of the state space of \( Z^n \). With the convention \( \exp(-\infty) \triangleq 0 \), we obtain for such functions \( h \) the probability of the event \( \{Z^n \in A\} \) since

\[
\tilde{E}\{\exp[-n h(Z^n)]\} = P\{Z^n \in A\}.
\]

An interesting class of functions \( h \) that are bounded from above are those of the form \( h(z) = -c\|z\|^2 \), where \( c > 0 \) and the state space of \( Z^n \) is a Banach space with norm \( \|\cdot\| \).

Our aim in the present section is rather modest. In Proposition 4.5.1 we show how to modify the proof of part (a) of Proposition 1.4.2 to allow functions that are unbounded either from below or from above. The assertion of the latter proposition can be regarded as the fundamental representation formula, and indeed all of the particular representations that we obtain for specific models are derived by essentially rewriting this fundamental
representation formula. This is done by exploiting structural properties of the measure that plays the role of \( \theta \). As a consequence, the techniques that we use to extend part (a) of Proposition 1.4.2 to Proposition 4.5.1 can be adapted to extend the representation formula for any particular model from the case of bounded functions \( h \) to the case where \( h \) is only bounded from below or from above. Although we will not provide the details, this is indeed the case. The paper \[13\] carries out this program for a particular class of processes.

As in previous sections, we restrict ourselves here to discrete-time processes. In the next section, representation formulas for a class of continuous-time processes involving bounded measurable functions \( h \) are treated, and the analysis can also be extended to functions \( h \) that are only bounded from below or from above.

**Proposition 4.5.1.** Let \( \mathcal{X} \) be a Polish space, \( k \) a measurable function mapping \( \mathcal{X} \) into \( \mathbb{R} \) which is either bounded from below or bounded from above, and \( \theta \) a probability measure on \( \mathcal{X} \). Then we have the variational formula

\[
- \log \int_{\mathcal{X}} e^{-k} d\theta = \inf_{\gamma \in \Delta(\mathcal{X})} \left\{ R(\gamma\|\theta) + \int_{\mathcal{X}} k d\gamma \right\}, \tag{4.12}
\]

where \( \Delta(\mathcal{X}) = \{ \gamma \in \mathcal{P}(\mathcal{X}) : R(\gamma\|\theta) < \infty \} \).

**Proof.** Since the infimum is restricted to probability measures \( \gamma \) satisfying \( R(\gamma\|\theta) < \infty \), the right hand side of equation (4.12) is well defined. We first prove the variational formula when \( k \) is bounded from below. Since for \( \alpha \in \mathbb{N} \) \( k \wedge \alpha \) is bounded and measurable, part (a) of Proposition 1.4.2 implies that

\[
- \log \int_{\mathcal{X}} e^{-(k \wedge \alpha)} d\theta = \inf_{\gamma \in \Delta(\mathcal{X})} \left\{ R(\gamma\|\theta) + \int_{\mathcal{X}} (k \wedge \alpha) d\gamma \right\}
\leq \inf_{\gamma \in \Delta(\mathcal{X})} \left\{ R(\gamma\|\theta) + \int_{\mathcal{X}} k d\gamma \right\}.
\]

Thus by the Lebesgue Dominated Convergence Theorem

\[
- \log \int_{\mathcal{X}} e^{-k} d\theta = \lim_{\alpha \to \infty} \left( - \log \int_{\mathcal{X}} e^{-(k \wedge \alpha)} d\theta \right)
\leq \inf_{\gamma \in \Delta(\mathcal{X})} \left\{ R(\gamma\|\theta) + \int_{\mathcal{X}} k d\gamma \right\}.
\]

In order to prove that

\[
\inf_{\gamma \in \Delta(\mathcal{X})} \left\{ R(\gamma\|\theta) + \int_{\mathcal{X}} k d\gamma \right\} \leq - \log \int_{\mathcal{X}} e^{-k} d\theta,
\]

we assume that \( - \log \int_{\mathcal{X}} e^{-k} d\theta < \infty \) since otherwise there is nothing to prove. Given \( \alpha \in \mathbb{N} \) and \( \varepsilon > 0 \) there exists a probability measure \( \gamma_\alpha \) on \( \mathcal{X} \) such that

\[
R(\gamma_\alpha\|\theta) + \int_{\mathcal{X}} (k \wedge \alpha) d\gamma_\alpha \leq \inf_{\gamma \in \Delta(\mathcal{X})} \left\{ R(\gamma\|\theta) + \int_{\mathcal{X}} (k \wedge \alpha) d\gamma \right\} + \varepsilon
\leq - \log \int_{\mathcal{X}} e^{-k} d\theta + \varepsilon
\leq - \log \int_{\mathcal{X}} e^{-k} d\theta + \varepsilon < \infty.
\]
4.5. EXTENSIONS TO UNBOUNDED FUNCTIONS

Since \( k \) is bounded from below, it follows that \( \sup_{\alpha \in \mathbb{N}} R(\gamma_\alpha\|\theta) < \infty \). This implies that the sequence \( \{\gamma_\alpha, \alpha \in \mathbb{N}\} \) is relatively compact with respect to the weak topology [Lemma 1.4.3 (c)]. Actually, more is true. According to another important property of the relative entropy, which we have not yet discussed, the bound \( \sup_{\alpha \in \mathbb{N}} R(\gamma_\alpha\|\theta) < \infty \) implies the compactness of \( \{\gamma_\alpha\} \) with respect to a stronger topology called the \( \tau \)-topology [Proposition 9.3.6]. It follows that there exists a probability measure \( \gamma \) on \( \mathcal{X} \) such that along some subsequence of \( \alpha \in \mathbb{N} \)

\[
\lim_{\alpha \to \infty} \int_{\mathcal{X}} \psi d\gamma_\alpha = \int_{\mathcal{X}} \psi d\gamma
\]

for all bounded measurable functions \( \psi \) mapping \( \mathcal{X} \) into \( \mathbb{R} \); in particular, \( \gamma_\alpha \Rightarrow \gamma \). Since \( R(\gamma\|\theta) \leq \liminf_{\alpha \to \infty} R(\gamma_\alpha\|\theta) < \infty \),

\[
- \log \int_{\mathcal{X}} e^{-k} d\theta + \varepsilon \\
\geq \liminf_{\alpha \to \infty} \left( R(\gamma_\alpha\|\theta) + \int_{\mathcal{X}} (k \land \alpha) d\gamma_\alpha \right) \\
\geq \liminf_{M \to \infty} \liminf_{\alpha \to \infty} \left( R(\gamma_\alpha\|\theta) + \int_{\mathcal{X}} (k \land \alpha \land M) d\gamma_\alpha \right) \\
\geq \liminf_{M \to \infty} \left( R(\gamma\|\theta) + \int_{\mathcal{X}} (k \land M) d\gamma \right) \\
\geq R(\gamma\|\theta) + \int_{\mathcal{X}} k d\gamma \\
\geq \inf_{\gamma \in \Delta(\mathcal{X})} \left\{ R(\gamma\|\theta) + \int_{\mathcal{X}} k d\gamma \right\}.
\]

Sending \( \varepsilon \to 0 \) completes the proof of the variational formula under the assumption that \( k \) is bounded from below.

We now prove the variational formula under the assumption that \( k \) is bounded from above. For \( \alpha \in \mathbb{N} \), \( k \lor (-\alpha) \) is bounded and measurable, and so by part (a) of Proposition 1.4.2

\[
- \log \int_{\mathcal{X}} e^{-[k \lor (-\alpha)]} d\theta = \inf_{\gamma \in \Delta(\mathcal{X})} \left\{ R(\gamma\|\theta) + \int_{\mathcal{X}} [k \lor (-\alpha)] d\gamma \right\} \\
\geq \inf_{\gamma \in \Delta(\mathcal{X})} \left\{ R(\gamma\|\theta) + \int_{\mathcal{X}} k d\gamma \right\}.
\]

The Monotone Convergence Theorem yields

\[
- \log \int_{\mathcal{X}} e^{-k} d\theta = \lim_{\alpha \to \infty} \left( - \log \int_{\mathcal{X}} e^{-[k \lor (-\alpha)]} d\theta \right) \\
\geq \inf_{\gamma \in \Delta(\mathcal{X})} \left\{ R(\gamma\|\theta) + \int_{\mathcal{X}} k d\gamma \right\}.
\]

Let \( \varepsilon > 0 \) be given. In order to prove that

\[
- \log \int_{\mathcal{X}} e^{-k} d\theta \leq \inf_{\gamma \in \Delta(\mathcal{X})} \left\{ R(\gamma\|\theta) + \int_{\mathcal{X}} k d\gamma \right\},
\]
we assume that the right hand side is less than $\infty$, for otherwise there is nothing to prove. We choose a probability measure $\tilde{\gamma}$ on $\mathcal{X}$ such that

$$R(\tilde{\gamma}||\theta) + \int_{\mathcal{X}} k \, d\tilde{\gamma} \leq \inf_{\gamma \in \Delta(\mathcal{X})} \left\{ R(\gamma||\theta) + \int_{\mathcal{X}} k \, d\gamma \right\} + \varepsilon < \infty.$$  

Then

$$- \log \int_{\mathcal{X}} e^{-[k \vee (-\alpha)]} \, d\theta = \inf_{\gamma \in \Delta(\mathcal{X})} \left\{ R(\gamma||\theta) + \int_{\mathcal{X}} [k \vee (-\alpha)] \, d\gamma \right\} \leq R(\tilde{\gamma}||\theta) + \int_{\mathcal{X}} [k \vee (-\alpha)] \, d\tilde{\gamma}.$$  

Since $R(\tilde{\gamma}||\theta) < \infty$, the Monotone Convergence Theorem yields

$$- \log \int_{\mathcal{X}} e^{-k} \, d\theta = \lim_{a \to \infty} \left( - \log \int_{\mathcal{X}} e^{-[k \vee (-\alpha)]} \, d\theta \right) \leq \lim_{a \to \infty} \left( R(\tilde{\gamma}||\theta) + \int_{\mathcal{X}} [k \vee (-\alpha)] \, d\tilde{\gamma} \right) = R(\tilde{\gamma}||\theta) + \int_{\mathcal{X}} k \, d\tilde{\gamma} \leq \inf_{\gamma \in \Delta(\mathcal{X})} \left\{ R(\gamma||\theta) + \int_{\mathcal{X}} k \, d\gamma \right\} + \varepsilon.$$  

Sending $\varepsilon \to 0$ completes the proof of the variational formula under the assumption that $k$ is bounded from above. $\blacksquare$

With this proposition the discussion of extensions to unbounded functions is done. In the next section we take up another interesting topic.

### 4.6 Representation Formulas for Continuous–Time Markov Processes

After an introductory subsection, we present two subsections in which representation formulas for general continuous–time Markov processes are formally derived and then are specialized to diffusion processes and jump Markov processes. We end with a subsection in which we remark on how to prove the representation formulas. The reader who is uninterested in representation formulas for continuous–time processes can skip this section since none of these results will be used elsewhere in the book.

#### 4.6.1 Introduction

Except in the present section, this book is concerned exclusively with representation formulas that are associated with discrete–time processes. This focus does not mean that one cannot derive representations for continuous–time processes. Consider, for example, a continuous–time Markov process such as a diffusion or a jump process. Since “samples”
of such a process taken at discrete–time instances form a discrete–time Markov process, it would seem possible to study large deviation phenomena for continuous–time Markov processes by investigating the large deviation properties of the sampled system when the sampling interval tends to 0 sufficiently fast as the large deviation parameter tends to its limit. For a detailed discussion of how such an approximation may be carried out for a particular class of processes, we refer the reader to Chapter 10. Because of this procedure, it may appear, at least from the viewpoint of large deviation theory, that not much is lost by restricting to representation formulas only for discrete–time processes.

Nonetheless, it is interesting and useful to know that the analogous representation formulas hold even in the continuous–time setting. A particular application for which such representation formulas would be especially well suited is in the derivation of higher-order asymptotics beyond the basic Laplace principle. In addition, for some problems involving discontinuous statistics the description and analysis of the sampled version of a continuous–time process could be rather cumbersome. An example of such a case is a diffusion process whose drift changes discontinuously as one crosses an interface \[63, 64\]. Finally, for purposes of computation the continuous–time representation formulas are often a more convenient starting point than the discrete–time approximations since in carrying out computations one often uses a discretization of time that differs from the discretization of time used in the large deviation analysis \[67\].

This discussion shows that there are indeed situations in which a representation formula in the continuous–time setting may be preferred. There are, however, more limitations on the availability of such representations, and the level of mathematical sophistication required to formulate and prove the representations can be substantially higher. This situation is analogous to other complications that arise in the study of continuous–time processes but are absent in the study of discrete–time processes. For example, to construct diffusion processes and analyze their uniqueness properties one must deal with concepts such as stochastic integrals, stochastic differential equations, and martingale problems. In contrast, in the discrete–time setting questions of existence and uniqueness are elementary. This contrast in difficulty carries over to the derivation of representation formulas. In the continuous–time setting we must deal, for example, with progressively measurable controls rather than with the simple feedback controls to which we have been able to restrict in the discrete–time setting \[61\].

In this section we will attempt to strike a reasonable balance. Instead of rigorously deriving the representation formulas, we will be content with a formal derivation, the main point being to indicate the basic form that each representation formula takes. We will start with a continuous–time Markov process, sample along a sequence of discrete times of the form \( j \Delta \) to obtain a discrete–time Markov process, write down the representation formula for this process, and then formally take the limit as \( \Delta \to 0 \). Then we will specialize to the two important cases of diffusions and jump Markov processes. It will come as no surprise that in such special cases the representation formula will take a much simpler form than in the general case. Although some of the representations for continuous–time processes can be derived directly, we wish to show the connection between the discrete–time and the continuous–time problems.

After the representation formula has been formally derived, there is still the issue of
proving it. Three approaches are discussed in Subsection 4.6.4. Derivations of stochastic control representations for exponential functions of continuous-time Markov processes are contained in a number of works including [13, 21, 51, 52, 54, 58, 82].

4.6.2 A Formal Derivation of Representation Formulas for Continuous-Time Markov Processes

The discussion in this section will be completely formal. Since these results will not be applied later on, we do not include any large deviation scaling parameter $n$.

Let $X = \{X(t), t \in [0, 1]\}$ be a continuous-time Markov process that has stationary transition probabilities and takes values in $\mathbb{R}^d$. We assume that the sample paths of $X$ are elements of $\mathcal{D}([0, 1] : \mathbb{R}^d)$, which is the space of functions that map $[0, 1]$ into $\mathbb{R}^d$, are right continuous, and have left hand limits. $\mathcal{D}([0, 1] : \mathbb{R}^d)$ is metrized by the Skorohod metric, with respect to which it is complete and separable. Details are given in Section A.6. In this subsection we address the problem of obtaining a representation formula for

$$W(x) \doteq - \log E_x \{ \exp[-h(X)] \},$$

where $E_x$ denotes expectation conditioned on $X(0) = x$ and where for simplicity we take $h$ to be a bounded continuous function mapping $\mathcal{D}([0, 1] : \mathbb{R}^d)$ into $\mathbb{R}$.

We begin by defining the sampled Markov process constructed from $X$. Take $\Delta > 0$ such that $1/\Delta$ is an integer. For $j \in \{0, 1, \ldots, 1/\Delta\}$ we define $X_j^\Delta \doteq X(j\Delta)$. The sequence \(\{X_j^\Delta, j = 0, 1, \ldots, 1/\Delta\}\) forms a Markov chain with stationary transition probabilities, and its transition probability function is

$$\gamma^\Delta(x, dy) \doteq P_x \{ X(\Delta) \in dy \}.$$

We also define a piecewise constant process $X^\Delta = \{X^\Delta(t), t \in [0, 1]\}$ by

$$X^\Delta(t) \doteq \begin{cases} X(j\Delta) & \text{for } t \in [j\Delta, j\Delta + \Delta), j = 0, 1, \ldots, 1/\Delta - 2 \\ X((1/\Delta - 1)\Delta) & \text{for } t \in [(1/\Delta - 1), 1]. \end{cases}$$

We work with this piecewise constant process $X^\Delta$, rather than with a piecewise linear process as in previous sections, because it is natural for the approximating process to take values in the same space $\mathcal{D}([0, 1] : \mathbb{R}^d)$ as the original process $X$. Finally, in lieu of $W(x)$ let us consider

$$W^\Delta(x) \doteq - \log E_x \{ \exp[-h(X^\Delta)] \}.$$

Since a representation formula for $W^\Delta(x)$ can be derived by following the same procedure used earlier in the book, we will merely record the formula without proof. We specify first the dynamics and then the cost structure of the associated stochastic control problem.

**Dynamics.** The control that is applied at time $j \in \{0, 1, \ldots, 1/\Delta - 1\}$ is a stochastic kernel $\nu^\Delta_j$ on $\mathbb{R}^d$ given $(\mathbb{R}^d)^{j+1}$. We denote the dependence of $\nu^\Delta_j$ upon a point $(x_0, x_1, \ldots, x_j) \in (\mathbb{R}^d)^{j+1}$ by writing $\nu^\Delta_j = \nu^\Delta_j(dy|x_0, x_1, \ldots, x_j)$. A sequence of controls
4.6. CONTINUOUS–TIME MARKOV PROCESSES

\( \{ \nu_j^\Delta, j = 0, 1, \ldots, 1/\Delta - 1 \} \) will be called an **admissible control sequence**. The controlled process takes values in \( \mathbb{R}^d \) and is denoted by \( \{ \tilde{X}_j^\Delta, j = 0, 1, \ldots, 1/\Delta \} \). The evolution of the controlled process is defined by setting \( \tilde{X}_0^\Delta \doteq X_0^\Delta = x \) and specifying \( \tilde{X}_{j+1}^\Delta \) to have the conditional distribution

\[
P_x \{ \tilde{X}_{j+1}^\Delta \in dy | \tilde{X}_0^\Delta, \tilde{X}_1^\Delta, \ldots, \tilde{X}_j^\Delta \} = \nu_j^\Delta(dy | \tilde{X}_0^\Delta, \tilde{X}_1^\Delta, \ldots, \tilde{X}_j^\Delta).
\]

We also define a piecewise constant controlled process by

\[
\tilde{X}^\Delta(t) \doteq \begin{cases} 
\tilde{X}(j\Delta) & \text{for } t \in [j\Delta, j\Delta + \Delta), j = 0, 1, \ldots, 1/\Delta - 2 \\
\tilde{X}((1/\Delta) - 1)\Delta) & \text{for } t \in [(1/\Delta - 1), 1].
\end{cases}
\]

**Cost structure.** The minimal cost function is defined by

\[
V^\Delta(x) \doteq \inf_{\{\nu_j^\Delta\}} \mathbb{E}_x \left\{ \sum_{j=0}^{(1/\Delta)-1} R(\nu_j^\Delta(\cdot) \| \gamma_j^\Delta(\tilde{X}_j^\Delta, \cdot)) + h(\tilde{X}^\Delta) \right\}.
\]

In this formula \( \nu_j^\Delta(\cdot) = \nu_j^\Delta(\cdot | \tilde{X}_0^\Delta, \tilde{X}_1^\Delta, \ldots, \tilde{X}_j^\Delta) \), the infimum is taken over all admissible control sequences \( \{\nu_j^\Delta\} \). \( \mathbb{E}_x \) denotes expectation conditioned on \( \tilde{X}_0^\Delta = x \), \( \{\tilde{X}_j^\Delta\} \) is the sequence of controlled random variables that are associated with a particular admissible control sequence \( \{\nu_j^\Delta\} \), \( \tilde{X}^\Delta = \{\tilde{X}^\Delta(t), t \in [0,1]\} \) is the corresponding piecewise constant controlled process, and \( h \) is the same function appearing in the definition of \( W^\Delta(x) \).

The representation formula states that for all \( \Delta > 0 \) such that \( 1/\Delta \) is an integer and for all \( x \in \mathbb{R}^d \)

\[
W^\Delta(x) \doteq - \log \mathbb{E}_x \{ \exp[-h(\tilde{X}^\Delta)] \}
\]

equals the minimal cost function \( V^\Delta(x) \). If one could calculate a limit for \( V^\Delta(x) \) as \( \Delta \to 0 \), then it would be reasonable to hope that this limit would provide a representation for \( W(x) \). Before attempting such a calculation, we first make a few formal approximations.

To begin, we introduce a basic concept. We will say that the Markov process \( X \) is associated with the generator \( \mathcal{L} \) if there exists a nonempty family \( \mathcal{G} \) of bounded continuous test functions such that

\[
\mathcal{L} f(x) = \lim_{\delta \downarrow 0} \frac{1}{\delta} : \mathbb{E}_x \{ f(X(\delta)) - f(x) \}
\]

for all \( f \in \mathcal{G} \) and \( x \in \mathbb{R}^d \). Of course, for this association to be useful one needs conditions that guarantee that knowing \( \mathcal{L} f(x) \) for all \( f \in \mathcal{G} \) and \( x \in \mathbb{R}^d \) characterizes a unique Markov process \( X \). However, for the purely formal purposes of this section such issues will be ignored. It is also useful to introduce an approximation to \( \mathcal{L} \) by defining for \( f \in \mathcal{G} \) and \( x \in \mathbb{R}^d \)

\[
\mathcal{L}^\Delta f(x) \doteq \frac{1}{\Delta} \mathbb{E}_x \{ f(X(\Delta)) - f(x) \} = \frac{1}{\Delta} \int_{\mathbb{R}^d} [f(y) - f(x)] \gamma^\Delta(x, dy).
\]
For $f \in \mathcal{G}$ we have the obvious limit

\[ \mathcal{L}f(x) = \lim_{\Delta \to 0} \mathcal{L}^\Delta f(x) = \lim_{\Delta \to 0} \frac{1}{\Delta} \int_{\mathbb{R}^d} [f(y) - f(x)] \gamma^\Delta(x, dy). \] (4.13)

Our hope in calculating the limit as $\Delta \to 0$ of the minimal cost functions $V^\Delta(x)$ is that we will obtain a limit function $V(x)$ that can be identified as the minimal cost function of a stochastic control problem. This limiting problem will be defined in terms of some limiting controlled continuous–time process. In order to describe the limiting problem, two basic questions must be answered. What are the dynamics of the limiting controlled process and what is the cost structure? We begin by addressing the first of these. We will find it convenient to describe the dynamics by identifying the generator that is associated with the limiting controlled process.

In each of the stochastic control problems that have been used so far to derive representation formulas in discrete time, the dynamics share a common feature. In each case, the control applied at time $j$ selects the distribution of a controlled random variable that replaces the randomness entering the description of the original process at time $j$. There is an obvious difficulty in adopting such a procedure in the continuous–time limit since the controlled distributions must scale in a $\Delta$–dependent fashion. As we will see, this can be avoided by choosing the Radon-Nikodym derivatives of the controlled distributions with respect to the original distributions instead. Exactly what is meant by this will be made clear momentarily.

The discretized process $\{X(j\Delta), j = 0, 1, \ldots, 1/\Delta\}$ is a Markov chain with stationary transition probabilities and transition probability function $\gamma^\Delta(x, dy) \equiv P_x\{X(\Delta) \in dy\}$. In the corresponding controlled process, this transition probability function is replaced at each time $j$ by a controlled distribution $\nu^\Delta = \nu^\Delta_j$ with the corresponding running cost $R(\nu^\Delta(\cdot)\|\gamma^\Delta(x, \cdot))$. The scaling being used requires that the interval $[0, 1]$ is divided into $1/\Delta$ subintervals each of length $\Delta$, and this subdivision corresponds to a running cost of $R(\nu^\Delta(\cdot)\|\gamma^\Delta(x, \cdot))/\Delta$ per unit time. Since we can always restrict our attention to controls for which the running cost is finite, we can assume without loss of generality that $\nu^\Delta(\cdot) \ll \gamma^\Delta(x, \cdot)$. In other words, we can always assume that the controlled distribution applied at any time $j$ has a Radon-Nikodym derivative with respect to $\gamma^\Delta(x, dy)$. Thus we can also define this Radon-Nikodym derivative to be the control. To simplify the notation, the time parameter will not be indicated.

This is in fact the procedure that we will follow although for technical reasons one must be more careful with the class of allowed controls than one might at first imagine. Let us consider a bounded continuous function $u$ mapping $\mathbb{R}^d$ into $\mathbb{R}$ with the additional property that $e^u \in \mathcal{G}$. In the remainder of this section we will calculate the generator of the controlled process and the running cost in the limit $\Delta \to 0$, given that a fixed control $u$ is applied and that the state of the controlled process is $x$. We will use $u$ to define the controlled distribution $\nu^\Delta(\cdot)$ through

\[ \nu^\Delta(A) \equiv \int_A e^{u(y)} \gamma^\Delta(x, dy) \cdot \frac{1}{\int_{\mathbb{R}^d} e^{u(y)} \gamma^\Delta(x, dy)} \]
for Borel subsets $A$ of $\mathbb{R}^d$. The Radon–Nikodym derivative is given by

$$
\frac{d\nu^\Delta(\cdot)}{d\gamma^\Delta(x, \cdot)}(y) = \frac{e^{u(y)}}{\int_{\mathbb{R}^d} e^{u(y)} \gamma^\Delta(x, dy)}.
$$

Because we assume only that $u$ is bounded, we need to include a normalization guaranteeing that $\nu^\Delta$ is a probability measure. This complicates the development slightly, but seems to be unavoidable.

For $f \in \mathcal{G}$ and $x \in \mathbb{R}^d$ a controlled discrete–time operator analogous to $\mathcal{L}^\Delta$ can be defined by

$$
\mathcal{L}^{u, \Delta} f(x) = \int_{\mathbb{R}^d} [f(y) - f(x)] e^{u(y)} \gamma^\Delta(x, dy) \cdot \frac{1}{\Delta} \frac{1}{\int_{\mathbb{R}^d} e^{u(y)} \gamma^\Delta(x, dy)}.
$$

If the formal limit exists, then we also define a limiting operator $\mathcal{L}^u$ for $f \in \mathcal{G}$ by setting

$$
\mathcal{L}^u f(x) = \lim_{\Delta \to 0} \mathcal{L}^{u, \Delta} f(x).
$$

If it exists, then $\mathcal{L}^u$ is the desired generator that is associated with the limiting controlled process. In order to determine the form of this limiting generator, an approximation will be made which we now explain.

Since for any $g \in \mathcal{G}$

$$
\int_{\mathbb{R}^d} g(y) \gamma^\Delta(x, dy) = g(x) + \Delta \mathcal{L} g(x) + o(\Delta),
$$

up to terms of order $\Delta$ the denominator of $\mathcal{L}^{u, \Delta} f(x)$ equals $\Delta \cdot \exp[u(x)]$. Therefore if the formal limit exists, then

$$
\mathcal{L}^u f(x) = \lim_{\Delta \to 0} e^{-u(x)} \frac{1}{\Delta} \int_{\mathbb{R}^d} [f(y) - f(x)] e^{u(y)} \gamma^\Delta(x, dy).
$$

Because of the product term $f(x) \exp[u(y)]$ appearing in this expression, it is not obvious either that the formal limit exists or if it does, that the formal limit is the generator of any process. However, if we rewrite this expression in the form

$$
e^{-u(x)} \frac{1}{\Delta} \int_{\mathbb{R}^d} [f(y) e^{u(y)} - f(x) e^{u(x)}] \gamma^\Delta(x, dy)$$

$$
- f(x) e^{-u(x)} \frac{1}{\Delta} \int_{\mathbb{R}^d} [e^{u(y)} - e^{u(x)}] \gamma^\Delta(x, dy),
$$

then using equation (4.13) we can recognize the formal limit as the generator

$$
\mathcal{L}^u f(x) = e^{-u(x)} [\mathcal{L}(fe^u)(x) - f(x) \mathcal{L}(e^u)(x)].
$$

This gives the form of the generator for the limiting controlled process. The function $u$ can be viewed as a control, the selection of which modifies the time evolution of the
process through the generator $\tilde{L}^u$. In the next subsection more explicit formulas for this generator will be given in two cases of interest.

We next determine the cost structure of the limiting stochastic control problem, the only issue being the form of the running cost. As it has already been defined, the running cost per unit time for the discrete-time stochastic control problem equals

$$\frac{1}{\Delta} R(\nu^\Delta(\cdot) \| \gamma^\Delta(x, \cdot)) = \frac{1}{\Delta} \int_{\mathbb{R}^d} \frac{d\nu^\Delta(\cdot)}{d\gamma^\Delta(x, \cdot)}(y) \left( \log \frac{d\nu^\Delta(\cdot)}{d\gamma^\Delta(x, \cdot)}(y) \right) \gamma^\Delta(x, dy).$$

By formula (4.14), up to terms of order $o(\Delta)$ the controlled distribution $\nu^\Delta(dy)$ is approximately equal to

$$e^{u(y)} \gamma^\Delta(x, dy) \quad \frac{e^{u(y)} \gamma^\Delta(x, dy)}{e^{u(x)} + \Delta \mathcal{L}(e^u)(x)}.$$

Since for real numbers $a, b$, and $c$

$$\frac{e^a}{e^b + \Delta c} = e^{a-b} \left( 1 - \Delta \frac{c}{e^b} + o(\Delta) \right)$$

and

$$\log \frac{e^a}{e^b + \Delta c} = a - b - \Delta \frac{c}{e^b} + o(\Delta),$$

it follows that up to terms of order $o(\Delta)$

$$\frac{d\nu^\Delta(\cdot)}{d\gamma^\Delta(x, \cdot)}(y) \approx e^{u(y)-u(x)} \left( 1 - \Delta \frac{\mathcal{L}(e^u)(x)}{e^{u(x)}} \right)$$

and

$$\log \frac{d\nu^\Delta(\cdot)}{d\gamma^\Delta(x, \cdot)}(y) \approx \log \frac{e^{u(y)}}{e^{u(x)} + \Delta \mathcal{L}(e^u)(x)} \approx u(y) - u(x) - \Delta \frac{\mathcal{L}(e^u)(x)}{e^{u(x)}}.$$

Substituting, we obtain up to terms of order $o(1)$ the approximation

$$\frac{1}{\Delta} R(\nu^\Delta(\cdot) \| \gamma^\Delta(x, \cdot)) \approx e^{-u(x)} \frac{1}{\Delta} \int_{\mathbb{R}^d} [u(y) - u(x)] e^{u(y)} \gamma^\Delta(x, dy)$$

$$- \frac{\mathcal{L}(e^u)(x)}{e^{u(x)}} \int_{\mathbb{R}^d} e^{u(y)-u(x)} \gamma^\Delta(x, dy)$$

$$- \frac{\mathcal{L}(e^u)(x)}{e^{u(x)}} \int_{\mathbb{R}^d} [u(y) - u(x)] e^{u(y)-u(x)} \gamma^\Delta(x, dy).$$

Using equation (4.15), we conclude that the limiting running cost per unit time is given by

$$\lim_{\Delta \to 0} \frac{1}{\Delta} R(\nu^\Delta(\cdot) \| \gamma^\Delta(x, \cdot)) = \tilde{L}^u u(x) - \frac{\mathcal{L}(e^u)(x)}{e^{u(x)}}.$$

In conclusion, the formal calculations given in this section show that if the control is $u$, then the generator of a limiting controlled process has for $f \in \mathcal{G}$ the form

$$\tilde{L}^u f(x) = e^{-u(x)} \left[ \mathcal{L}(fe^u)(x) - f(x) \mathcal{L}(e^u)(x) \right]$$

(4.16)
and the limiting running cost per unit time has the form
\[ \mathcal{L}^u u(x) - \frac{\mathcal{L}(e^u)(x)}{e^u(x)}. \]

In the next section we specialize these calculations to diffusion processes and jump Markov processes.

### 4.6.3 Examples of Continuous–Time Representation Formulas

In this section we consider representation formulas for the two important special cases of diffusion processes and jump–Markov processes. The discussion to follow relies on a number of results concerning existence and uniqueness which we will essentially ignore. Additional background material on control theory for continuous–time Markov processes can be found in a number of texts including [61, 77].

**Diffusion processes.** Consider the stochastic differential equation
\[ dX(t) = b(X(t)) \, dt + \sigma(X(t)) \, dw(t), \quad X(0) = x, \]
where the function \( b \), the matrix \( \sigma \), and the Wiener process \( \{w(t), t \in [0, 1]\} \) have compatible dimensions. Under suitable conditions this stochastic differential equation has a solution \( X = \{X(t), t \in [0, 1]\} \) that is unique in distribution. In the standard terminology, the solution is a diffusion process with drift vector \( b(x) \) and diffusion matrix \( a(x) \equiv \sigma(x) \sigma^T(x) \), where \( \sigma^T(x) \) denotes the transpose of the matrix \( \sigma(x) \). As is well known, the diffusion process is associated with the generator
\[ \mathcal{L}f(x) \equiv \langle f_x(x), b(x) \rangle + \frac{1}{2} \text{tr} [f_{xx}(x) \, a(x)], \]
where \( f \) is a twice continuously differentiable function, \( f_x \) and \( f_{xx} \) denote respectively the gradient and the Hessian of \( f \), and \( \text{tr} B \) denotes the trace of a square matrix \( B \).

Substituting into (4.16) this formula for the generator, we find after some calculation that the corresponding controlled generator takes the form
\[ \mathcal{L}^u f(x) = \mathcal{L}f(x) + \langle f_x(x), a(x) \, u_x(x) \rangle, \]
where \( u_x(x) \) denotes the gradient of \( u \) at the point \( x \). In other words, the controlled process is again a diffusion with diffusion matrix \( a(x) \). The effect of the control \( u \) is to alter the drift from the value \( b(x) \) that it has in the original generator to the new value \( b(x) + a(x) \, u_x(x) = b(x) + \sigma(x) \sigma^T(x) \, u_x(x) \). We also find that the running cost per unit time takes the form
\[ \frac{1}{2} \langle u_x(x), a(x) \, u_x(x) \rangle = \frac{1}{2} \| \sigma^T(x) \, u_x(x) \|^2. \]

According to these formulas, the dynamics of the controlled process and the corresponding running cost per unit time depend on \( u(x) \) only through \( \sigma^T(x) \, u_x(x) \). Therefore we can consider the latter quantity alone as defining the control, and in order to ease the notation
we relabel it as \( v(x) = \sigma^T(x) u_x(x) \). By analogy with the discrete-time random walk model of Section 4.3, we expect that the controls depend on the current time \( t \), on the past, and on the current state \( x \) of the controlled process. We now indicate how this will be formalized.

Generally in the context of continuous-time problems, one would like to consider controls that depend only on the time \( t \) and on the state of the controlled process at time \( t \). Such controls are called feedback controls. However, it is generally not the case that one can restrict attention to feedback controls when dealing with stochastic control problems in continuous time. In fact, one often has to work with what are called progressively measurable controls. These controls have the property that at each time \( t \) the control is measurable with respect to a certain \( \sigma \)-field, and the Wiener process that drives the controlled stochastic differential equation is a Wiener process with respect to this \( \sigma \)-field. Because of these properties the control does not anticipate the future increments of the Wiener process. The controls that are allowed by this definition are functions of \( t \) and \( \omega \) (and implicitly of the current state), and we will write such a control as \( v(t) = v(t, \omega) \).

For the precise definition of the class of progressively measurable controls, the reader can consult [54, 61]. This class is a natural continuous-time analogue of the class of admissible controls that were used for the random walk model.

Putting aside the technicalities regarding the class of admissible controls, we now turn to the representation formula for the class of diffusion processes under consideration. Let \((\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathcal{P}})\) be a probability space and \( \bar{\omega} = \{ \bar{x}(t), t \in [0, 1] \} \) a Wiener process on this space. Given an admissible control \( \{ v(t), t \in [0, 1] \} \), the controlled process \( \bar{X} = \{ \bar{X}(t), t \in [0, 1] \} \) is defined to be the solution of the stochastic differential equation

\[
d\bar{X}(t) = b(\bar{X}(t)) dt + \sigma(\bar{X}(t)) v(t) dt + \sigma(\bar{X}(t)) d\bar{\omega}(t), \quad \bar{X}(0) = x. \tag{4.17}
\]

Let \( h \) be a bounded continuous function mapping \( D([0, 1] : \mathbb{R}^d) \) into \( \mathbb{R} \) and define

\[
W(x) \doteq -\log \mathbb{E}_x \{ \exp[-h(X)] \}.
\]

Then the representation formula should have the form

\[
W(x) = \inf_v \bar{\mathbb{E}}_x \left\{ \int_0^1 \frac{1}{2}||v(t)||^2 dt + h(\bar{X}) \right\}, \tag{4.18}
\]

where the infimum is taken over the class of progressively measurable controls \( v \) and \( \bar{\mathbb{E}}_x \) denotes expectation with respect to \( \bar{\mathcal{P}}_x \). In the next subsection we will remark on different approaches that can be used to prove the representation formula. Using the representation formula as the starting point, one can directly analyze various large deviation problems related to the diffusion \( X \) such as small noise asymptotics [13] and the asymptotics of the empirical measure.

**Jump Markov processes.** For bounded continuous functions \( f \) mapping \( \mathbb{R}^d \) into \( \mathbb{R} \), we consider an operator of the form

\[
\mathcal{L} f(x) \doteq \int_{\mathbb{R}^d} [f(x + y) - f(x)] \mu_x(dy),
\]
where \( \{\mu_x(dy), x \in \mathbb{R}^d\} \) is a suitable family of measures on \( \mathbb{R}^d \) parametrized by \( x \in \mathbb{R}^d \).

When there exists a Markov process \( X = \{X(t), t \in [0,1]\} \) that is associated with this generator, \( \mu_x(dy) \) determines the jump intensity. Thus, the probability that \( X \) will jump by \( dy \) during a time interval of duration \( \Delta \), given that the process is currently at the state \( x \), is equal to \( \Delta \cdot \mu_x(dy) \) plus terms of order \( o(\Delta) \). Just as in the case of diffusions, one needs to impose suitable conditions on \( \mu_x(dy) \) in order to guarantee there exists a well defined Markov process that is associated with this generator.

Substituting into equation (4.16) this formula for the generator, we find that the corresponding controlled generator takes the form

\[
\mathcal{L}^u f(x) = \int_{\mathbb{R}^d} [f(x + y) - f(x)] e^{u(x+y) - u(x)} \mu_x(dy)
\]

and that the running cost per unit time takes the form

\[
\int_{\mathbb{R}^d} [(u(x+y) - u(x)) e^{u(x+y) - u(x)} - e^{u(x+y) - u(x)} + 1] \mu_x(dy).
\]

We can simplify these expressions slightly by redefining the control to be \( v(y) = e^{u(x+y) - u(x)} \).

The generator of the controlled process then becomes

\[
\mathcal{L}^u f(x) = \int_{\mathbb{R}^d} [f(x + y) - f(x)] v(y) \mu_x(dy).
\]

Thus the controlled process is again a jump Markov process, the jump intensity measure having been changed from \( \mu_x(dy) \) to \( v(y) \mu_x(dy) \). Making the same substitution in the running cost per unit time, we obtain

\[
\int_{\mathbb{R}^d} [v(y) \log v(y) - v(y) + 1] \mu_x(dy).
\]

In certain cases, an example of which is given in [41], one can restrict to feedback controls. However in general one must again extend the class of controls to include all suitably measurable functions \( v(y,t), y \in \mathbb{R}^d, t \in [0,1], \) of \( \omega \). This issue is discussed, for example, in [54].

Again putting aside the technicalities regarding the class of admissible controls, we now turn to the representation formula for the class of jump Markov processes under consideration. Given a control \( v(y,t) \), the controlled process \( \{\hat{X}(t), t \in [0,1]\} \) is a jump Markov process defined by the generator

\[
\int_{\mathbb{R}^d} [f(x + y) - f(x)] v(y,t) \mu_x(dy).
\]

Let \( h \) be a bounded measurable function mapping \( \mathcal{D}([0,1] : \mathbb{R}^d) \) and define

\[
W(x) \equiv - \log E_x \{\exp[-h(X)]\}.
\]

Then the representation formula for the original jump Markov process \( X \) should have the form

\[
W(x) = \inf_v \mathbb{E}_x \left\{ \int_0^1 \left( \int_{\mathbb{R}^d} [v(y,t) \log v(y,t) - v(y,t) + 1] \mu_{X(t)}(dy) \right) dt + h(\hat{X}) \right\},
\]
where the infimum is taken over all suitable controls \( v \). Using this representation formula as the starting point, one can analyze various large deviation phenomena of the jump Markov process \( X \).

The next subsection completes this chapter. In it we discuss how to prove the representation formulas for continuous-time processes which we have formally derived.

### 4.6.4 Remarks on the Proofs of the Representation Formulas

In this subsection we briefly mention three methods that can be applied to rigorize the formal development of Subsections 4.6.2 and 4.6.3. The first approach is based on the theory of viscosity solutions of nonlinear partial differential equations [20, 54]. For this theory to be applicable, we must restrict attention to functions \( h \) such that

\[
\exp[-W(x)] = E_x \{ \exp[-h(X)] \}
\]

can be represented as the value at time 0 of the solution of some linear partial differential equation. Under appropriate conditions, one can then derive a nonlinear partial differential equation that is satisfied in the viscosity sense by \( W(x) \) [54]. Again under appropriate conditions, one can further show that the minimal cost function appearing in the proposed representation formula for \( W(x) \) also satisfies this same nonlinear partial differential equation in the viscosity sense. If solutions to the nonlinear partial differential equation are unique in the viscosity sense, then the representation formula for \( W(x) \) is proved. For examples of specific cases where this is carried out and for further references to the literature, see Chapter VI of [54].

The second approach, which is based on weak convergence methods for stochastic control problems [65, 66], is much closer in spirit to the techniques used in this book. However, as in the case of the viscosity solution method, it does not seem to yield representations in the greatest possible generality. For simplicity, we assume that the function \( h \) is bounded and continuous. We also assume that there is a unique process that is associated with the generator \( \mathcal{L} \) and initial condition \( x \) in the weak sense, where “associated” means that \( X \) solves an appropriate martingale problem for \( \mathcal{L} \) [50, 86]. Given that such weak-sense uniqueness is valid, it is elementary to prove that

\[
W(x) = \lim_{\Delta \to 0} W^\Delta(x).
\]

Let \( V(x) \) denote the minimal cost function that we wish to show is equal to \( W(x) \). Then the representation formula \( W(x) = V(x) \) is established as soon as we prove that \( \lim_{\Delta \to 0} V^\Delta(x) = V(x) \). This type of weak convergence problem has been extensively studied in the literature, and the reader is referred to [67] for a detailed discussion of required conditions and techniques in many interesting cases.

The third approach appears to be best suited for continuous-time problems. It utilizes what might be called the direct substitution of Radon-Nikodym derivatives. Let us consider the special case of diffusion processes discussed in the last subsection. We suppose that the coefficients \( b \) and \( \sigma \) have the property that the stochastic differential
equation defining \( X = \{ X(t), t \in [0,1] \} \) admits a strong solution that is unique in the strong sense [61]. In this case there is a measurable function that maps the Wiener process \( w = \{ w(t), t \in [0,1] \} \) into \( X \), and consequently representation formulas for \( X \) become a special case of the representation formula for \( w \). As we discussed in the introduction to this chapter, some representation formulas are more convenient than others. While here a representation is possible either in terms of the distribution of \( X \) or that of \( w \), the more convenient representation is that based on the Wiener process.

Having reduced to the special case of a Wiener process, we will sketch for this process the proof of the representation formula (4.18) that was formally derived in the previous subsection. Full details of the derivation are provided in [13]. The starting point is the fundamental representation formula provided by Proposition 1.4.2. If \( h \) is any bounded measurable function mapping \( C([0,1] : \mathbb{R}^r) \) into \( \mathbb{R} \) and if \( w \) takes values in \( \mathbb{R}^r \), then

\[
- \log \mathbb{E}\{ \exp[-h(w)] \} = \inf_{\gamma \in \mathcal{P}(C([0,1] : \mathbb{R}^r))} \left\{ R(\gamma\|\theta) + \int_{C([0,1] : \mathbb{R}^r)} h \, d\gamma \right\}, \tag{4.19}
\]

where \( \theta \) denotes Wiener measure on \( C([0,1] : \mathbb{R}^r) \). For simplicity we assume that the probability space on which \( w \) is defined is the canonical space \( C([0,1] : \mathbb{R}^r) \).

We now exploit structural properties of Wiener measure to rewrite the representation. It is well known that under certain conditions probability measures \( \gamma \) that are absolutely continuous with respect to \( \theta \) can be represented via the Cameron–Martin–Girsanov Formula [61]. Specifically, such measures have a Radon–Nikodym derivative with respect to \( \theta \) of the form

\[
\frac{d\gamma}{d\theta} = \exp \left( \sum_{i=1}^r \int_0^1 v_i(s) \, dw_i(s) - \frac{1}{2} \sum_{i=1}^r \int_0^1 |v_i(s)|^2 \, ds \right),
\]

where \( v = \{(v_1(t), v_2(t), \ldots, v_r(t)), t \in [0,1]\} \) is measurable with respect to the filtration generated by \( w \) and, in fact, is progressively measurable. This assures that the integral is well defined. In addition, with respect to \( \gamma \), the measure induced on \( C([0,1] : \mathbb{R}^r) \) by \( \bar{w}(t) = w(t) - \int_0^t v(s) \, ds \) is Wiener measure. Under appropriate integrability conditions, this allows us to rewrite the relative entropy term in (4.19) in the form

\[
R(\gamma\|\theta) = \int_{C([0,1] : \mathbb{R}^r)} \left( \log \frac{d\gamma}{d\theta} \right) \, d\gamma
\]

\[
= \int_{C([0,1] : \mathbb{R}^r)} \left( \sum_{i=1}^r \int_0^1 v_i(s) \, dw_i(s) - \frac{1}{2} \sum_{i=1}^r \int_0^1 |v_i(s)|^2 \, ds \right) \, d\gamma
\]

\[
= E^\gamma \left\{ \sum_{i=1}^r \int_0^1 v_i(s) \, d\bar{w}_i(s) + \frac{1}{2} \sum_{i=1}^r \int_0^1 |v_i(s)|^2 \, ds \right\}
\]

\[
= E^\gamma \left\{ \frac{1}{2} \int_0^1 |v(s)|^2 \, ds \right\},
\]

where \( E^\gamma \) denotes expectation with respect to \( \gamma \).
Now if it were the case that the last two displays were valid for all probability measures \( \gamma \) that are absolutely continuous with respect to \( \theta \), then the fact that \( R(\gamma\|\theta) < \infty \) implies \( \gamma \ll \theta \) would make the derivation of the representation formula rather elementary. However, it turns out that this is not the case, and in fact the structure of the set of probability measures that are absolutely continuous with respect to Wiener measure is not so simple. Nevertheless, one can employ an approximation procedure which shows that when calculating the infimum in (4.19) one can restrict to a class \( \mathcal{A} \) of probability measures \( \gamma \) for which the last two displays hold [13]. It follows that

\[
- \log E\{\exp[-h(w)]\} = \inf_{\gamma \in \mathcal{A}} E \gamma \left\{ \frac{1}{2} \int_0^1 \|v(s)\|^2 ds + h(\bar{w}(.)) + \int_0^1 v(s) ds \right\}.
\]

This resembles formula (4.18), where consistently with the stochastic differential equation (4.17) the process \( \bar{w}(t) + \int_0^t v(s) ds \) plays the role of the controlled process \( \bar{X} \).

A few more steps will allow us to rewrite the last display precisely in the form of equation (4.18). Let \( \bar{w} \) be a Wiener process that takes values in \( \mathbb{R}^d \) and is defined on a probability space \( (\bar{\Omega}, \bar{\mathcal{F}}, \bar{P}) \) with expectation operator \( \bar{E} \). Given any control \( v \) and associated expectation operator \( E \gamma \), one can show that there exists a progressively measurable process \( \bar{v} = \{(\bar{v}_1(t), \bar{v}_2(t), \ldots, \bar{v}_r(t)), t \in [0, 1]\} \) such the \( \bar{P} \)-distribution of \( \bar{v} \) equals the \( \gamma \)-distribution of \( v \). Identifying the class \( \mathcal{A} \) of probability measures \( \gamma \) and the class of controls, we can rewrite the last display in the desired form

\[
- \log E\{\exp[-h(w)]\} = \inf_{\bar{v}} \bar{E} \left\{ \frac{1}{2} \int_0^1 \|\bar{v}(s)\|^2 ds + h(\bar{X}) \right\},
\]

where \( \bar{X} \) is the controlled process

\[
\bar{X}(t) = \bar{w}(t) + \int_0^t \bar{v}(s) ds, t \in [0, 1].
\]

This completes our sketch of the proof of the representation formula for a Wiener process.

Until now we have been concerned with deriving representation formulas and applying them to prove the Laplace principle in the relatively elementary cases of Sanov’s Theorem and Mogulskii’s Theorem. In the remainder of the book we will change our tack, using the representation formulas to prove the Laplace principle in much more sophisticated settings.
Chapter 5

Compactness and Limit Properties for the Random Walk Model

5.1 Introduction

One of the basic classes of processes with which we are concerned is the random walk model of Section 4.3. In the present chapter we prove the necessary compactness and limit properties needed to verify the Laplace principle for this model. As we will see, the form of the Laplace principle depends on the underlying statistics. Two interesting examples are given in Chapters 6 and 7.

Already in Chapter 3 we sketched the proof of the Laplace principle for the simplest case of the random walk model, which was defined in terms of a sequence of i.i.d. random vectors. There we made the simplifying assumption that the distribution $\rho$ of the i.i.d. random vectors has compact support. This translated immediately into a compactness property of any sequence of admissible control measures $\nu^n$ and associated controlled random walks $X^n$. As the reader may remember, the verification of one of the limits in the Laplace principle was based on a relationship, stated without proof in part (b) of Lemma 3.3.2, between the limits of convergent subsequences of these two quantities. The problem still remains to give a complete proof of the Laplace principle when the compactness assumption on $\rho$ is dropped. This will be carried out in the present chapter and in the next chapter in the context of the general random walk model of Section 4.3.

There are two main results in this chapter. Proposition 5.3.2 proves a compactness property for sequences of admissible controls. Theorem 5.3.5 is a limit theorem for sequences of admissible control measures and associated controlled random walks, proving a relationship between them that generalizes part (b) of Lemma 3.3.2. Although this theorem may look difficult, it is a detailed yet basically straightforward extension of Kolmogorov’s law of large numbers.

The material in the present chapter is needed in Chapters 6, 7, and 10. The reader who is interested in reaching as quickly as possible the statements and proofs of the Laplace principles in those chapters need only read the statement of Theorem 5.3.5. The analogous definitions and analysis of control measures and associated controlled processes for the empirical measures of a Markov chain are given in Chapter 8.
One of the hypotheses in Theorem 5.3.5 is Condition 5.3.1, which is a boundedness condition on the stochastic kernel \( \mu(dy|x) \) that defines the random walk model. In the last section of this chapter, we present a strictly weaker version of Condition 5.3.1 for which all the conclusions of Theorem 5.3.5 continue to hold. This strictly weaker version is useful in certain applications.

5.2 Definitions and a Representation Formula

We begin by recalling the random walk model of Section 4.3. It is defined by a stochastic kernel \( \mu(dy|x) \) on \( \mathbb{R}^d \) given \( \mathbb{R}^d \). For \( x \in \mathbb{R}^d, n \in \mathbb{N} \), and \( j \in \{0,1,\ldots,n-1\} \), we defined \( X^n_0 \doteq x \) and

\[
X^n_{j+1} \doteq X^n_j + \frac{1}{n} v_j(X^n_j).
\]

In this formula, \( \{v_j(\cdot), j \in \mathbb{N}_0\} \) denotes an sequence of i.i.d. random vector fields having the common distribution

\[
P\{v_j(x) \in dy\} \doteq \mu(dy|x).
\]

We are interested in the Laplace principle for the piecewise linear interpolation of the random vectors \( \{X^n_j, j = 0,1,\ldots,n\} \). This is the random walk \( X^n = \{X^n(t), t \in [0,1]\} \) defined by

\[
X^n(t) \doteq X^n_j + \left(t - \frac{j}{n}\right) v_j(X^n_j) \quad \text{for} \quad t \in [j/n,(j+1)/n], j = 0,1,\ldots,n-1.
\]

Let \( h \) be a bounded measurable function mapping \( C([0,1]:\mathbb{R}^d) \) into \( \mathbb{R} \). In Theorem 4.3.1 a representation formula was obtained for

\[
W^n(x) \doteq -\frac{1}{n} \log E_x\{\exp[-n h(X^n)]\}
\]

in terms of the minimal cost function of an associated stochastic optimal control problem. The aim of the present section is to recall this representation formula and to rewrite it in a somewhat simpler and more convenient form.

As in Section 4.3, the controlled process is a sequence of random vectors \( \{\hat{X}_j^n, j = 0,1,\ldots,n\} \), the dynamics of which are given by \( \hat{X}_0^n = X_0^n = x \) and

\[
\hat{X}_{j+1}^n = \hat{X}_j^n + \frac{1}{n} \hat{Y}_j^n
\]

for \( j \in \{0,1,\ldots,n-1\} \). The control that is applied at time \( j \in \{0,1,\ldots,n-1\} \) is a stochastic kernel \( \nu^n_j \) on \( \mathbb{R}^d \) given \( (\mathbb{R}^d)^{j+1} \). A sequence of controls \( \{\nu^n_j, j = 0,1,\ldots,n-1\} \) is called an admissible control sequence. The stochastic kernel \( \nu^n_j \) controls the conditional distribution of the increment \( \hat{Y}_j^n \) by

\[
\hat{P}_x\{\hat{Y}_j^n \in dy | \hat{X}_0^n, \hat{X}_1^n, \ldots, \hat{X}_j^n\} \doteq \nu^n_j(dy|\hat{X}_0^n, \hat{X}_1^n, \ldots, \hat{X}_j^n).
\]
5.2. REPRESENTATION FORMULA

For all $n \in \mathbb{N}$ the controlled random vectors $\hat{X}_j^n$ and $\bar{Y}_j^n$ are defined on a common probability space $(\Omega, \mathcal{F}, \bar{P}_x)$. In terms of these quantities, we define the minimal cost function

$$V^n(x) \doteq \inf_{\nu^*_j} \bar{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^*_j(\cdot)||\mu(\cdot|\hat{X}_j^n)) + h(\hat{X}_1^n) \right\}. \quad (5.2)$$

In this formula $\nu^*_j(\cdot) = \nu^*_j(\cdot|\hat{X}_0^n, \hat{X}_1^n, \ldots, \hat{X}_j^n)$, the infimum is taken over all admissible control sequences $\{\nu^*_j\}$, $\bar{E}_x$ denotes expectation with respect to $\bar{P}_x$, and $\{\hat{X}_j^n\}$ is the sequence of controlled random vectors that are associated with a particular admissible control sequence $\{\nu^*_j\}$.

The representation formula given in Theorem 4.3.1 states that for all $n \in \mathbb{N}$ and $x \in \mathbb{R}^d$ $W^n(x)$ equals the minimal cost function $V^n(x)$.

In order to rewrite the representation formula in a simpler form, we adapt to the present situation several definitions given in Section 3.2, which was concerned with the proof of Mogulskii’s Theorem. In order to take limits in $n$ along admissible control sequences $\{\nu^*_j\}$, it is convenient to place them in a space that is independent of $n$. For $n \in \mathbb{N}$ and $t \in [0, 1]$ we define the stochastic kernel

$$\nu^n(dy|t) \doteq \begin{cases} \nu^n(dy) & \text{if } t \in [j/n, (j+1)/n), \ j = 0, 1, \ldots, n-2 \\ \nu^n(dy) & \text{if } t \in [(n-1)/n, 1], \end{cases} \quad (5.3)$$

where $\nu^n(dy) = \nu^n(dy|\hat{X}_0^n, \hat{X}_1^n, \ldots, \hat{X}_n^n)$. A random probability measure $\nu^n$ on $\mathbb{R}^d \times [0, 1]$ is then defined by

$$\nu^n(A \times B) \doteq \int_B \nu^n(A|t) \, dt \quad (5.4)$$

for Borel subsets $A$ of $\mathbb{R}^d$ and $B$ of $[0, 1]$. The term admissible control measure will be used to refer to $\nu^n$, and the last display will be summarized as $\nu^n(dy \times \, dt) = \nu^n(dy|t) \otimes dt$.

For $n \in \mathbb{N}$ we also define stochastic processes $\check{X}^n = \{\check{X}^n(t), t \in [0, 1]\}$ by

$$\check{X}^n(t) \doteq \begin{cases} \hat{X}_j^n & \text{if } t \in [j/n, (j+1)/n), \ j = 0, 1, \ldots, n-2 \\ \hat{X}_{n-1}^n & \text{if } t \in [(n-1)/n, 1]. \end{cases} \quad (5.5)$$

$\check{X}^n$ is the piecewise constant interpolation of the random vectors $\{\hat{X}_j^n, j = 0, 1, \ldots, n\}$, to be contrasted with the piecewise linear interpolation $\bar{X}^n = \{\bar{X}^n(t), t \in [0, 1]\}$ defined by

$$\bar{X}^n(t) \doteq \check{X}_0^n + \left(t - \frac{j}{n}\right) \bar{Y}_j^n \quad \text{for } t \in [j/n, (j+1)/n], \ j = 0, 1, \ldots, n-1. \quad (5.6)$$

As we will see in Theorem 5.3.5, under suitable assumptions the asymptotic behavior of the two processes is the same as $n \to \infty$. The piecewise constant process $\check{X}^n$ is introduced in order to obtain the following formula for the minimal cost function $V^n(x)$:

$$V^n(x) = \inf_{\nu^*_j} \bar{E}_x \left\{ \int_0^1 R(\nu^*_j(\cdot)||\mu(\cdot|\check{X}_j^n(t))) \, dt + h(\check{X}_1^n) \right\}. \quad (5.7)$$

For easy reference we now restate the representation formula.
Corollary 5.2.1. Let $h$ be a bounded measurable function mapping $C([0,1] : \mathbb{R}^d)$ into $\mathbb{R}$. Then for all $n \in \mathbb{N}$ and $x \in \mathbb{R}^d$

$$W^n(x) \doteq -\frac{1}{n} \log E_x\{\exp[-n h(X^n)]\}$$

equals the minimal cost function $V^n$ given either in equation (5.2) or (5.7).

We will use Corollary 5.2.1 in the proofs of the process-level Laplace principles in Chapters 6 and 7 when we evaluate the asymptotic behavior of the minimal cost functions $V^n(x)$. The representation (5.3)–(5.4) of the admissible control measures $\nu^n$ has been chosen to facilitate the derivation of a relationship between limits of the controlled processes $\tilde{X}^n$ and $\tilde{X}^n$ and the limits of these measures. This relationship, which is needed in the proofs of both the upper bound and the lower bound in the Laplace principle, is given in the next section in part (b) of Theorem 5.3.5. In order to derive an analogous relationship in Chapter 8, where the Laplace principle for the empirical measures of a Markov chain will be proved, a different representation of the control measures will be needed. These and related issues are discussed at the end of Chapter 3.

5.3 Compactness and Limit Properties

In order to guarantee the existence of convergent subsequences of $\tilde{X}^n$, $\tilde{X}^n$, and $\nu^n$ and to relate them to each other, a uniform integrability property of $\{\nu^n\}$ will be needed. This property is stated in Proposition 5.3.2. If the supports of the distributions $\{\mu(\cdot|x), x \in \mathbb{R}^d\}$ that define the random walk model are assumed to lie in a fixed compact subset of $\mathbb{R}^d$, then this uniform integrability property is automatic [Remark 5.3.3], and the proof of the Laplace principle can be shortened considerably. The paper [4] proves the large deviation principle for the random walk model under this restrictive condition on the supports of $\{\mu(\cdot|x), x \in \mathbb{R}^d\}$ as well as under a number of other conditions.

The representation formula obtained in Section 4.3 and repeated in the last corollary holds in great generality. In contrast, the proof of the uniform integrability property of sequences of admissible control measures will require a condition on $\mu(dy|x)$ which we now state. For each $x$ and $\alpha$ in $\mathbb{R}^d$, we define the cumulant generating function

$$H(x, \alpha) \doteq \log \int_{\mathbb{R}^d} e^{\langle \alpha, y \rangle} \mu(dy|x).$$

Condition 5.3.1. In the random walk model, for each $\alpha \in \mathbb{R}^d$

$$\sup_{x \in \mathbb{R}^d} H(x, \alpha) < \infty.$$
more restrictive than necessary. For example, in the setting of Chapter 10 this condition rules out the case of small-noise diffusion processes with linear drift. In the last section of this chapter, we present a weaker version of Condition 5.3.1 that is satisfied by discrete-time analogues of such processes. In this weaker condition we consider a shifted version of $H(x, \alpha)$ for which boundedness conditions such as Condition 5.3.1 are more easily satisfied. As we will see in Example 8.2.3, an analogous condition is also important in treating the large deviations of the empirical measures of a Markov chain.

Before stating the first result of this section, we recall some basic definitions. A sequence of probability measures $\{\gamma_n, n \in \mathbb{N}\}$ on a Polish space $\mathcal{X}$ is said to be **tight** if for each $\varepsilon > 0$ there exists a compact subset $K$ of $\mathcal{X}$ such that

$$\inf_{n \in \mathbb{N}} \gamma_n(K) \geq 1 - \varepsilon.$$  

According to Prohorov's Theorem [Theorem A.3.15] $\{\gamma_n\}$ is tight if and only if it is relatively compact with respect to weak convergence. We must also define the analogous concept for a sequence of random measures such as the admissible control measures $\{\nu^n\}$. A sequence of random variables $\{Y_n, n \in \mathbb{N}\}$ taking values in $\mathcal{X}$ is said to be tight if the sequence of distributions of $\{Y_n\}$ on $\mathcal{X}$ is tight.

In the next proposition we establish the tightness of a sequence of admissible control measures $\{\nu^n\}$ and a uniform integrability property, assuming a bound on the sequence of running costs associated with this sequence of controls. This proposition will be used to prove the Laplace principles in the subsequent two chapters. The bound on the sequence of running costs, given in formula (5.8), will be automatically satisfied by the admissible control sequences that arise in the course of the proofs.

**Proposition 5.3.2.** In the random walk model we assume Condition 5.3.1. For each $n \in \mathbb{N}$ and $x \in \mathbb{R}^d$, consider any admissible control sequence $\{\nu^n_j, j = 0, 1, \ldots, n - 1\}$ such that

$$\sup_{n \in \mathbb{N}} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot)\|\mu(\cdot|\tilde{X}^n_j)) \right\} \equiv \Delta < \infty, \tag{5.8}$$

where $\nu^n_j(\cdot) = \nu^n_j(\cdot|\tilde{X}^n_0, \tilde{X}^n_1, \ldots, \tilde{X}^n_j)$. Let the sequence of admissible control measures $\{\nu^n, n \in \mathbb{N}\}$ be defined by equations (5.3) and (5.4). Then $\{\nu^n\}$ is tight and, in fact, has the uniform integrability property

$$\lim_{C \to \infty} \sup_{n \in \mathbb{N}} E_x \left\{ \int_{\{|y| \leq C\} \times [0, 1]} \|y\| \nu^n(dy \times dt) \right\} = 0. \tag{5.9}$$

**Remark 5.3.3.** Suppose that the supports of all the distributions $\{\mu(\cdot|x), x \in \mathbb{R}^d\}$ lie in a fixed compact subset $K$ of $\mathbb{R}^d$. If (5.8) holds, then it follows that w.p.1, for $n \in \mathbb{N}$ and $j \in \{0, 1, \ldots, n - 1\}$ the measures $\nu^n_j$ all have support in $K$. Thus the sequence of admissible control measures $\{\nu^n\}$ is uniformly integrable in the sense of (5.9).

In order to prove Proposition 5.3.2, we need a result from weak convergence theory which exploits a bootstrap procedure implicit in Prohorov’s Theorem. Let $\mathcal{X}$ be a Polish
CHAPTER 5. COMPACTNESS AND LIMIT PROPERTIES

space. A measurable function \( g \) mapping \( \mathcal{X} \) into \( \mathbb{R} \cup \{ \infty \} \) is called a **tightly function** on \( \mathcal{X} \) if \( \inf_{x \in \mathcal{X}} g(x) > -\infty \) and if for each \( M < \infty \) the level set \( \{ x \in \mathcal{X} : g(x) \leq M \} \) is a relatively compact subset of \( \mathcal{X} \). We now define a function \( G \) mapping \( \mathcal{P}(\mathcal{X}) \) into \( \mathbb{R} \cup \{ \infty \} \) by

\[
G(\theta) = \int_{\mathcal{X}} g \, d\theta.
\]

According to Theorem A.3.17, for each \( M < \infty \) the set \( \{ \theta \in \mathcal{P}(\mathcal{X}) : G(\theta) \leq M \} \) is tight and \( G \) is a tightness function on \( \mathcal{P}(\mathcal{X}) \).

**Proof of Proposition 5.3.2.** We first prove the tightness of \( \{ \nu^n \} \), using property (5.9). The latter will be proved in a moment. Property (5.9) implies that

\[
\sup_{n \in \mathbb{N}} \int_{\mathcal{P}(\mathbb{R}^d \times [0,1])} \left( \int_{\mathbb{R}^d \times [0,1]} \|y\| \gamma(dy \times dt) \right) \tilde{P}_x \{ \nu^n \in d\gamma \} < \infty. \tag{5.10}
\]

By Theorem A.3.17 applied to \( \mathcal{X} = \mathbb{R}^d \times [0,1] \) and the tightness function \( g(y, t) \equiv \|y\| \) for \( (y, t) \in \mathbb{R}^d \times [0,1] \),

\[
G(\gamma) = \int_{\mathbb{R}^d \times [0,1]} \|y\| \gamma(dy \times dt)
\]

is a tightness function on \( \mathcal{P}(\mathbb{R}^d \times [0,1]) \). Applying the theorem a second time to \( \mathcal{X} = \mathcal{P}(\mathbb{R}^d \times [0,1]) \) and \( g(\gamma) \equiv G(\gamma) \) for \( \gamma \in \mathcal{P}(\mathbb{R}^d \times [0,1]) \), we see that for each \( M < \infty \) the set

\[
\left\{ Q \in \mathcal{P}(\mathcal{P}(\mathbb{R}^d \times [0,1])) : \int_{\mathcal{P}(\mathbb{R}^d \times [0,1])} \left( \int_{\mathbb{R}^d \times [0,1]} \|y\| \gamma(dy \times dt) \right) Q(d\gamma) \leq M \right\}
\]

is tight. Because of (5.10) \( \{ \nu^n \} \) is tight.

We now prove the limit (5.9) by adapting the proof of part (d) of Lemma 1.4.3. For each initial condition \( \bar{X}^n_0 = x \in \mathbb{R}^d \) (5.8) implies that for each \( j \in \{0,1,\ldots,n-1\} \)

\[
\tilde{P}_x \left\{ \omega \in \Omega : R\left( \nu^n_j(\cdot | \bar{X}^n_0(\omega), \bar{X}^n_1(\omega), \ldots, \bar{X}^n_j(\omega)) \| \mu(\cdot | \bar{X}^n_j(\omega)) \right) < \infty \right\}
\]

has \( \tilde{P}_x \)-probability 1. For \( \omega \in B^n_j \)

\[
\nu^n_j(\cdot | \bar{X}^n_0(\omega), \bar{X}^n_1(\omega), \ldots, \bar{X}^n_j(\omega)) \ll \mu(\cdot | \bar{X}^n_j(\omega)),
\]

and Theorem A.5.7 guarantees that there exists a version of

\[
f^n_j(y) = f^n_j(\omega, y) \diveq \begin{cases} 
\frac{d
^n_j(\cdot | \bar{X}^n_0(\omega), \bar{X}^n_1(\omega), \ldots, \bar{X}^n_j(\omega))}{d\mu(\cdot | \bar{X}^n_j(\omega))}(y) & \text{if } \omega \in B^n_j, y \in \mathbb{R}^d \\
\infty & \text{if } \omega \in \tilde{\Omega} \setminus B^n_j, y \in \mathbb{R}^d
\end{cases}
\]

which is a nonnegative measurable function of \( (\omega, y) \in \tilde{\Omega} \times \mathbb{R}^d \). It follows that for each \( j \in \{0,1,\ldots,n-1\} \)

\[
E_x \left\{ \int_{\mathbb{R}^d} f^n_j(y) \log f^n_j(y) \, \mu(dy | \bar{X}^n_j) \right\} = E_x \left\{ \int_{\mathbb{R}^d} f^n_j(y) \log f^n_j(y) \, \mu(dy | \bar{X}^n_j) \right\}. \tag{5.11}
\]
5.3. COMPACTNESS AND LIMIT PROPERTIES

As noted in equation (1.18), for \( a \geq 0, b \geq 0, \) and \( \sigma \geq 1 \)
\[
abla \leq e^{\sigma a} + \frac{1}{\sigma} (b \log b - b + 1).
\]

Since \( b \log b - b + 1 \geq 0 \) for all \( b \geq 0, \) we find that for \( \sigma \geq 1 \) and \( C > 0 \)
\[
\sup_{n \in \mathbb{N}} \mathbb{E}_x \left\{ \int_{\{y \in \mathbb{R}^d : \|y\| > C\}} \| y \| \nu^n(dy \times ds) \right\}
= \sup_{n \in \mathbb{N}} \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \int_{\{y \in \mathbb{R}^d : \|y\| > C\}} \| y \| f^n_j(y) \mu(dy|\tilde{X}^n_{\tilde{j}}) \right\}
\leq \sup_{n \in \mathbb{N}} \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \int_{\mathbb{R}^d} \left( f^n_j(y) \log f^n_j(y) - f^n_j(y) + 1 \right) \mu(dy|\tilde{X}^n_{\tilde{j}}) \right\}.
\]

Condition 5.3.1 implies that
\[
\lim_{C \to \infty} \sup_{n \in \mathbb{N}} \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \int_{\{y \in \mathbb{R}^d : \|y\| > C\}} \nu^n(y) \mu(dy|\tilde{X}^n_{\tilde{j}}) \right\} = 0,
\]
and by equation (5.11)
\[
\frac{1}{\sigma} \sup_{n \in \mathbb{N}} \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \int_{\mathbb{R}^d} \left( f^n_j(y) \log f^n_j(y) - f^n_j(y) + 1 \right) \mu(dy|\tilde{X}^n_{\tilde{j}}) \right\}
= \frac{1}{\sigma} \sup_{n \in \mathbb{N}} \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \left( \int_{\mathbb{R}^d} f^n_j(y) \log f^n_j(y) \mu(dy|\tilde{X}^n_{\tilde{j}}) - \nu^n_{\tilde{j}}(\mathbb{R}^d) + \mu(\mathbb{R}^d|\tilde{X}^n_{\tilde{j}}) \right) \right\}
= \frac{1}{\sigma} \sup_{n \in \mathbb{N}} \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_{\tilde{j}}(\cdot)\|\mu(\cdot|\tilde{X}^n_{\tilde{j}})) \right\} = \frac{\Delta}{\sigma}.
\]

Since \( \sigma \) can be taken arbitrarily large, the proof of the lemma is complete. ■

In addition to the compactness of the admissible control measures as expressed in Proposition 5.3.2, we will need to understand the convergence properties of both the control measures and the associated controlled processes. Theorem 5.3.5 is a basic result along these lines, and it will be applied a number of times in the sequel. Before we turn to this, however, let us assume the bound (5.8), which by Proposition 5.3.2 guarantees that any subsequence of admissible control measures \( \{\nu^n\} \) has a subsequence converging in distribution to a random variable \( \nu \). There exists a probability space, also labeled
(\(\bar{\Omega}, \bar{\mathcal{F}}, \bar{P}_x\)), such that \(\nu\) is a stochastic kernel on \(\mathbb{R}^d \times [0,1]\) given \(\bar{\Omega}\). The next lemma gives a useful decomposition of \(\nu\) which will be applied in Theorem 5.3.5. Let us denote Lebesgue measure on [0,1] by \(\lambda\) and convergence in distribution by \(\xrightarrow{D}\). According to the definition (5.4) of the admissible control measure \(\nu^n\), for each \(\omega\) its second marginal equals \(\lambda\). The convergence in distribution \(\nu^n \xrightarrow{D} \nu\) implies that w.p.1 the second marginal of \(\nu\) also equals \(\lambda\). The decomposition of \(\nu\) given in the next lemma now follows from Theorem A.5.6. We omit the proof since it is carried out exactly like the proof of Lemma 3.3.1.

**Lemma 5.3.4.** Let \((\bar{\Omega}, \bar{\mathcal{F}}, \bar{P}_x)\) be a probability space. Assume that \(\nu = \nu(dy \times dt|\omega)\) is a stochastic kernel on \(\mathbb{R}^d \times [0,1]\) given \(\bar{\Omega}\) which is the limit in distribution of a convergent subsequence of admissible control measures \(\nu^n\). Then there exists a stochastic kernel \(\nu(dy|t, \omega)\) on \(\mathbb{R}^d\) given \([0,1] \times \bar{\Omega}\) such that \(\bar{P}_x\)-a.s. for \(\omega \in \bar{\Omega}\)

\[
\nu(A \times B|\omega) = \int_B \nu(A|t, \omega) \, dt
\]

for all Borel subsets \(A\) of \(\mathbb{R}^d\) and \(B\) of \([0,1]\). In the sequel \(\omega\) will be suppressed in the notation for \(\nu(dy \times dt)\) and \(\nu(dy|t)\), and the last display will be summarized as \(\nu(dy \times dt) = \nu(dy|t) \otimes dt\).

In the next theorem we turn to convergence properties of the controlled processes \(\bar{X}^n\) and \(\bar{X}^n\). The theorem is essentially a functional law of large numbers, and it is easily motivated by considering the case where all the stochastic kernels \(\nu^n_j\) equal a fixed probability measure on \(\mathbb{R}^d\). A similar motivation was carried out after the statement of Lemma 3.3.2 and will not be repeated here. The bound (5.12) on the sequence of running costs will automatically be satisfied in the proof of the Laplace principle. In the proof of Theorem 5.3.5 we apply the Skorohod Representation Theorem, which involves introducing a new probability space. Following our convention, we retain the notation \((\bar{\Omega}, \bar{\mathcal{F}}, \bar{P}_x)\) for this new space.

**Theorem 5.3.5.** In the random walk model we assume Condition 5.3.1. For each \(n \in \mathbb{N}\) and \(x \in \mathbb{R}^d\), consider any admissible control sequence \(\{\nu^n_j\}\) such that

\[
\sup_{n \in \mathbb{N}} \bar{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R\left(\nu^n_j(\cdot)\|\mu(\cdot | \bar{X}^n_j)\right) \right\} < \infty, \tag{5.12}
\]

where \(\nu^n_j(\cdot) = \nu^n_j(\cdot | \bar{X}^n_0, \bar{X}^n_1, \ldots, \bar{X}^n_j)\). In terms of these admissible control sequences we define the sequences of stochastic processes \(\{\bar{X}^n\}\) and \(\{\bar{X}^n\}\) by formulas (5.6) and (5.5), respectively, and the sequence of admissible control measures \(\{\nu^n\}\) by formulas (5.3) and (5.4). The following conclusions hold.

(a) Given any subsequence of \(\{(\nu^n, \bar{X}^n, \bar{X}^n), n \in \mathbb{N}\}\) there exists a subsequence, a probability space \((\bar{\Omega}, \bar{\mathcal{F}}, \bar{P}_x)\), a stochastic kernel \(\nu\) on \(\mathbb{R}^d \times [0,1]\) given \(\bar{\Omega}\), and a random variable \(\bar{X}\) mapping \(\bar{\Omega}\) into \(C([0,1] : \mathbb{R}^d)\) such that the subsequence converges in distribution to \((\nu, \bar{X}, \bar{X})\). The stochastic kernel \(\nu = \nu(dy \times dt)\) has the decomposition given in Lemma 5.3.4.
5.3. COMPACTNESS AND LIMIT PROPERTIES

(b) With probability 1, for every \( t \in [0, 1] \)

\[
X(t) = x + \int_{\mathbb{R}^d \times [0,t]} y \nu(dy \times ds) = x + \int_0^t \left( \int_{\mathbb{R}^d} y \nu(dy|s) \right) ds,
\]

and \( \bar{X}(t) \) is an absolutely continuous function of \( t \in [0, 1] \). Therefore, a.s. with respect to Lebesgue measure the derivative of \( \bar{X}(t) \) is given by

\[
\bar{X}(t) = \int_{\mathbb{R}^d} y \nu(dy|t).
\] (5.13)

Proof. For each \( n \in \mathbb{N} \) we define the process \( S^n \) \( = \{S^n(t), t \in [0, 1]\} \) by

\[
S^n(t) = x + \int_{\mathbb{R}^d \times [0,t]} y \nu^n(dy \times ds) = x + \int_0^t \left( \int_{\mathbb{R}^d} y \nu^n(dy|s) \right) ds.
\] (5.14)

The uniform integrability property of \( \{\nu^n, n \in \mathbb{N}\} \) expressed in Proposition 5.3.2 guarantees that \( S^n \) is well defined and takes values in \( C([0,1]: \mathbb{R}^d) \). The proof of Theorem 5.3.5 proceeds in three steps.

**Step 1.** Consider any subsequence of \( \{\nu^n, n \in \mathbb{N}\} \) satisfying the bound (5.12). Then there exist a subsequence, a probability space \( (\bar{\Omega}, \bar{\mathcal{F}}, \bar{P}_x) \), and a stochastic kernel \( \nu \) on \( \mathbb{R}^d \times [0, 1] \) given \( \bar{\Omega} \) with the properties that

\[
\nu^n \overset{\mathcal{D}}{\to} \nu \quad \text{and} \quad \mathbb{E}_x \left\{ \int_{\mathbb{R}^d \times [0,1]} \|y\| \nu(dy \times dt) \right\} < \infty.
\] (5.15)

\( \mathbb{E}_x \) denotes expectation with respect to \( \bar{P}_x \). For the remainder of the proof, we work with this convergent subsequence of \( \{\nu^n\} \).

**Step 2.** For \( t \in [0, 1] \) we define

\[
\bar{X}(t) = x + \int_{\mathbb{R}^d \times [0,t]} y \nu(dy \times dt),
\] (5.16)

which is well defined because of the bound in (5.15). Step 2 is to prove that \( (\nu^n, S^n) \overset{\mathcal{D}}{\to} (\nu, \bar{X}) \), where \( \bar{X} \) denotes the process \( \{\bar{X}(t), t \in [0, 1]\} \).

**Step 3.** For each \( \varepsilon > 0 \)

\[
\lim_{n \to \infty} \bar{P}_x \left\{ \sup_{t \in [0,1]} \|S^n(t) - \bar{X}^n(t)\| \geq \varepsilon \right\} = 0
\]

and

\[
\lim_{n \to \infty} \bar{P}_x \left\{ \sup_{t \in [0,1]} \|S^n(t) - \bar{X}^n(t)\| \geq \varepsilon \right\} = 0.
\]
Before verifying these three steps, let us check that Theorem 5.3.5 will follow from them. We first consider part (a). For each \( n \in \mathbb{N} \) \( S^n \) and the piecewise linearly interpolated process \( \bar{X}^n \) take values in \( C([0,1]:\mathbb{R}^d) \) while the piecewise constant process \( X^n \) takes values in \( D([0,1]:\mathbb{R}^d) \). \( C([0,1]:\mathbb{R}^d) \) is metrized by the uniform metric and \( D([0,1]:\mathbb{R}^d) \) by the Skorohod metric, which is denoted by \( \rho(\cdot, \cdot) \). With respect to these metrics the two spaces are Polish spaces [Theorem A.6.5 (a)]. Since by part (a) of Theorem A.6.5
\[
\rho(S^n, \bar{X}^n) \leq \sup_{t \in [0,1]} \|S^n(t) - \bar{X}^n(t)\|
\]
Steps 1, 2, and 3 and Theorem A.3.8 yield part (a). In particular, they imply that \( (\nu^n, \bar{X}^n, \bar{X}^n) \xrightarrow{D} (\nu, \bar{X}, \bar{X}) \) along the subsequence selected in Step 1.

Part (b) in the statement of the theorem gives a number of properties of \( \bar{X} \). The first display follows from the definition of \( \bar{X}(t) \) given in Step 2, the decomposition \( \nu(dy \times ds) = \nu(dy | s) \otimes ds \) given in Lemma 5.3.4, and the bound
\[
E_x \left\{ \int_{\mathbb{R}^d \times [0,1]} \|y\| \nu(dy \times dt) \right\} < \infty
\]
in Step 1. It follows that w.p. 1 \( \bar{X}(t) \) is an absolutely continuous function of \( t \in [0,1] \).

Equation (5.13) for \( \dot{X}(t) \) is an immediate consequence. We conclude that Theorem 5.3.5 follows from Steps 1, 2, and 3. We now verify these three steps.

**Proof of Step 1.** The convergence in distribution \( \nu^n \xrightarrow{D} \nu \) along a subsequence is a consequence of Proposition 5.3.2. For \( C \in (0, \infty) \) we define
\[
\psi^n(C) \doteq E_x \left\{ \int_{\{y \in \mathbb{R}^d: \|y\| > C\} \times [0,1]} \|y\| \nu^n(dy \times ds) \right\}
\]
and
\[
\psi(C) \doteq E_x \left\{ \int_{\{y \in \mathbb{R}^d: \|y\| > C\} \times [0,1]} \|y\| \nu(dy \times ds) \right\}.
\]
We prove the bound in equation (5.15) by showing that
\[
\lim_{C \to \infty} \psi(C) = 0.
\]
This limit is a consequence of the inequalities
\[
\sup_{n \in \mathbb{N}} \psi^n(C) \geq \liminf_{n \to \infty} \psi^n(C) \geq \psi(C) \tag{5.17}
\]
and the uniform integrability established in Proposition 5.3.2, which states that under condition (5.12)
\[
\lim_{C \to \infty} \sup_{n \in \mathbb{N}} \psi^n(C) = 0.
\]
The first inequality in (5.17) is elementary. So to complete the proof of the bound in (5.15), we must prove \( \liminf_{n \to \infty} \psi^n(C) \geq \psi(C) \).
5.3. COMPACTNESS AND LIMIT PROPERTIES

Since \( \nu^n \xrightarrow{\mathcal{D}} \nu \), the Skorohod Representation Theorem guarantees that \( \nu^n \xrightarrow{w.p.1} \nu \)\( \). Hence Fatou's Lemma yields the desired lower limit once we show that the function mapping
\[
\tau \in \mathcal{P}(\mathbb{R}^d \times [0, 1]) \mapsto \int_{\{y \in \mathbb{R}^d : \|y\| > C\} \times [0, 1]} \|y\| \, \tau(dy \times ds)
\]
is lower semicontinuous. Since the function mapping
\[
(y, s) \in \mathbb{R}^d \times [0, 1] \mapsto 1_{\{\zeta \in \mathbb{R}^d : \|\zeta\| > C\}}(y) \|y\|
\]
is nonnegative and lower semicontinuous, this follows directly from Theorem A.3.12. The proof of Step 1 is complete.

**Proof of Step 2.** Step 2 asserts that \( (\nu^n, S^n) \xrightarrow{\mathcal{D}} (\nu, X) \). To prove this, we first show that \( \{\nu^n, S^n\} \) is tight. Since \( \{\nu^n\} \) is tight [Proposition 5.3.2], it suffices to prove that \( \{S^n\} \) is tight. This is done in the following lemma.

**Lemma 5.3.6.** For each \( n \) and \( \delta \in (0, 1) \) we define \( S^n \) by formula (5.14) and the modulus of continuity of \( S^n \) by
\[
w^n(\delta) \triangleq \sup_{\{s, t \in [0, 1] : |s - t| \leq \delta\}} \|S^n(s) - S^n(t)\|
\]
The following conclusions hold.

(a) For each \( \epsilon > 0 \) and \( \eta > 0 \) there exists \( \delta \in (0, 1) \) such that for all \( n \)
\[
P_x\{w^n(\delta) \geq \epsilon\} \leq \eta.
\]

(b) \( \{S^n\} \) is tight.

**Proof.** (a) The uniform integrability of \( \{\nu^n\} \) as expressed in equation (5.9) allows \( C > 0 \) to be chosen so large that
\[
\sup_{n \in \mathbb{N}} E_x \left\{ \int_{\{y \in \mathbb{R}^d : \|y\| > C\} \times [0, 1]} \|y\| \, \nu^n(dy \times ds) \right\} \leq \eta \epsilon / 2. \tag{5.18}
\]
Defining \( \delta \triangleq (\epsilon / 2C) \land 1 \), we then substitute
\[
\|y\| = \|y\| 1_{\{\zeta \in \mathbb{R}^d : \|\zeta\| \leq C\}}(y) + \|y\| 1_{\{\zeta \in \mathbb{R}^d : \|\zeta\| > C\}}(y)
\]
into
\[
P_x\{w^n(\delta) \geq \epsilon\} \leq P_x \left\{ \sup_{\{s, t \in [0, 1] : |s - t| \leq \delta\}} \int_{\mathbb{R}^d \times [s \land t, s \lor t]} \|y\| \, \nu^n(dy \times ds) \geq \epsilon \right\}.
\]
Since \( C \delta \leq \epsilon / 2 \),
\[
\sup_{\{s, t \in [0, 1] : |s - t| \leq \delta\}} \int_{\{y \in \mathbb{R}^d : \|y\| \leq C\} \times [s \land t, s \lor t]} \|y\| \, \nu^n(dy \times ds) \leq C \delta \leq \epsilon / 2,
\]
and so (5.18) yields for all $n$

$$
P_x\{\omega^n(\delta) \geq \varepsilon\} \leq \frac{1}{\varepsilon/2} E_x\left\{ \int_{\{y \in \mathbb{R}^d : \|y\| > C\} \times [0,1]} \|y\| \nu^n(dy \times ds) \geq \varepsilon/2 \right\}
$$

$$
\leq \frac{1}{\varepsilon/2} E_x\left\{ \int_{\{y \in \mathbb{R}^d : \|y\| > C\} \times [0,1]} \|y\| \nu^n(dy \times ds) \right\}
$$

$$
\leq \eta.
$$

(b) Since $S^n(0) = x$ for all $n \in \mathbb{N}$, the tightness of $\{S^n\}$ is a consequence of part (a) and Theorem A.3.22. ■

The lemma implies that for any subsequence of $n \in \mathbb{N}$ there exists a subsequence, a stochastic kernel $\nu$, and a random variable $S$ taking values in $C([0,1] : \mathbb{R}^d)$ such that $(\nu^n, S^n) \xrightarrow{D} (\nu, S)$. We again invoke the Skorohod Representation Theorem, which allows us to assume that w.p.1 $(\nu^n, S^n) \rightarrow (\nu, S)$. In order to complete the proof of Step 2, it suffices to show that w.p.1 $S$ equals $\bar{X}$. We claim that for each fixed $t \in [0,1]$ $S(t) = \bar{X}(t)$ w.p.1. It will then follow that w.p.1, for all rational $t \in [0,1]$ $S(t) = \bar{X}(t)$, and since $S(\cdot)$ and $\bar{X}(\cdot)$ are continuous, we will be able to conclude that $S(\cdot) = \bar{X}(\cdot)$ w.p.1. The claim will now be proved.

We start by showing that for each $t \in [0,1]$ and each bounded continuous function $g$ mapping $\mathbb{R}^d$ into $\mathbb{R}$

$$
\lim_{n \to \infty} \int_{\mathbb{R}^d \times [0,1]} g(y) \nu^n(dy \times ds) = \int_{\mathbb{R}^d \times [0,1]} g(y) \nu(dy \times ds)
$$

w.p.1. Since the function mapping

$$(y, s) \in \mathbb{R}^d \times [0,1] \mapsto g(y) 1_{[0, t]}(s)$$

is discontinuous on $\mathbb{R}^d \times \{t\}$, the limit (5.19) is not immediate. In order to derive it, we recall from Lemma 5.3.4 that w.p.1, since the second marginal of $\nu$ equals Lebesgue measure $\lambda$ on $[0,1]$, we have $\nu(\mathbb{R}^d \times \{t\}) = \lambda(\{t\}) = 0$. We now appeal to a basic result in weak convergence theory. Let $\{\gamma_n, n \in \mathbb{N}\}$ be a sequence of probability measures on a Polish space $\mathcal{X}$ converging weakly to $\gamma$. If $f$ is a bounded measurable function on $\mathcal{X}$ whose points of discontinuity form a set of $\gamma$–measure zero, then by Theorem A.3.11

$$
\lim_{n \to \infty} \int_{\mathcal{X}} f d\gamma_n = \int_{\mathcal{X}} f d\gamma.
$$

This yields (5.19).

For fixed $t \in [0,1]$ we will prove that $S^n(t) \xrightarrow{D} \bar{X}(t)$, where $\xrightarrow{D}$ denotes convergence in probability. Since $S^n(t) \rightarrow S(t)$ w.p.1, this will imply that $S(t) = \bar{X}(t)$ w.p.1, as claimed. From the proof of Step 1, for $C \in (0, \infty)$ we recall that

$$
\psi^n(C) \leq E_x\left\{ \int_{\{y \in \mathbb{R}^d : \|y\| > C\} \times [0,1]} \|y\| \nu^n(dy \times ds) \right\}
$$
and 
\[
\psi(C) \doteq \mathbb{E}_x \left\{ \int_{y \in \mathbb{R}^d : \|y\| > C} \|y\| \nu(dy \times ds) \right\}
\]
are related by
\[
\psi(C) \leq \sup_{n \in \mathbb{N}} \psi^n(C).
\]

We define a bounded continuous function of \( y \in \mathbb{R}^d \) by
\[
\varphi_C(y) \doteq \begin{cases} 
\frac{y}{\|y\|} & \text{if } \|y\| \leq C \\
C & \text{if } \|y\| > C.
\end{cases}
\]

For any \( \varepsilon > 0 \) Chebyshev’s Inequality implies that
\[
\mathbb{P}_x \left\{ \|S^n(t) - \hat{X}(t)\| \geq \varepsilon \right\} 
\leq \frac{1}{\varepsilon} \mathbb{E}_x \left\{ \|\int_{\mathbb{R}^d \times [0, t]} y \nu^n(dy \times ds) - \int_{\mathbb{R}^d \times [0, t]} y \nu(dy \times ds)\| \right\} 
\leq \frac{1}{\varepsilon} \mathbb{E}_x \left\{ \|\int_{\mathbb{R}^d \times [0, t]} \varphi_C(y) \nu^n(dy \times ds) - \int_{\mathbb{R}^d \times [0, t]} \varphi_C(y) \nu(dy \times ds)\| \right\} 
+ \frac{1}{\varepsilon} \left( \sup_{n \in \mathbb{N}} \psi^n(C) + \psi(C) \right) 
\leq \frac{1}{\varepsilon} \mathbb{E}_x \left\{ \|\int_{\mathbb{R}^d \times [0, t]} \varphi_C(y) \nu^n(dy \times ds) - \int_{\mathbb{R}^d \times [0, t]} \varphi_C(y) \nu(dy \times ds)\| \right\} + \frac{2}{\varepsilon} \sup_{n \in \mathbb{N}} \psi^n(C).
\]

We now take \( n \to \infty \) followed by \( C \to \infty \). In the last line of this display, as \( n \to \infty \) the expectation converges to 0 by (5.19) and the Lebesgue Dominated Convergence Theorem, and
\[
\lim_{C \to \infty} \sup_{n \in \mathbb{N}} \psi^n(C) \to 0
\]
by Proposition 5.3.2. This finishes the proof that \( S^n(t) \xrightarrow{P} \hat{X}(t) \). The proof of Step 2 is complete.

**Proof of Step 3.** Step 3 asserts that
\[
\sup_{t \in [0, 1]} \|S^n(t) - \hat{X}(t)\| \xrightarrow{P} 0 \quad \text{and} \quad \sup_{t \in [0, 1]} \|S^n(t) - \hat{X}(t)\| \xrightarrow{P} 0.
\]

As defined in Lemma 5.3.6, for \( n \in \mathbb{N} \) and \( \delta > 0 \) \( w^n(\delta) \) equals
\[
\sup_{\{s, t \in [0, 1], |s - t| \leq \delta\}} \|S^n(s) - S^n(t)\|.
\]
Using the fact that $\tilde{X}^n$ and $\tilde{X}^n$ are respectively the piecewise linear interpolation and the piecewise constant interpolation of the random vectors $\{\tilde{X}^n_j\} = \{\tilde{X}^n(j/n)\}$, we have

$$
\sup_{t \in [0,1]} \left\| S^n(t) - \tilde{X}^n(t) \right\| = \max_{k \in \{0,1,\ldots,n-1\}} \sup_{t \in [k/n, (k+1)/n]} \left\| S^n(t) - \tilde{X}^n(t) \right\| \\
\leq w^n(1/n) + \max_{k \in \{0,1,\ldots,n\}} \left\| S^n(k/n) - \tilde{X}^n(k/n) \right\|
$$

and

$$
\sup_{t \in [0,1]} \left\| S^n(t) - \tilde{X}^n(t) \right\| = \max_{k \in \{0,1,\ldots,n-1\}} \sup_{t \in [k/n, (k+1)/n]} \left\| S^n(t) - \tilde{X}^n(t) \right\| \\
\leq w^n(1/n) + \max_{k \in \{0,1,\ldots,n\}} \left\| S^n(k/n) - \tilde{X}^n(k/n) \right\|.
$$

Part (a) of Lemma 5.3.6 implies that $w^n(1/n) \xrightarrow{P} 0$. Hence the next lemma completes the proof of Step 3.

**Lemma 5.3.7.** \(\max_{k \in \{0,1,\ldots,n\}} \left\| S^n(k/n) - \tilde{X}^n(k/n) \right\| \xrightarrow{P} 0.\)

**Proof.** This proof is made more complicated by the fact that the only bound on the tails of the admissible controls measures is the uniform integrability given in Proposition 5.3.2. This necessitates the use of a truncation argument, which is reminiscent of the argument used in the proof of the law of large numbers for sums of i.i.d. random variables having only a finite first moment.

For $n \in \mathbb{N}$ and $j \in \{0,1,\ldots,n\}$, let $\mathcal{F}_j^n$ denote the sigma field generated by $\{\tilde{X}_i^n, i = 0,1,\ldots,j\}$. Given $\theta > 0$, we define for $j \in \{0,1,\ldots,n-1\}$

$$
\xi_j^n = \begin{cases} 
\tilde{X}_{j+1}^n - \tilde{X}_j^n & \text{if } \|\tilde{X}_{j+1}^n - \tilde{X}_j^n\| < \theta \\
0 & \text{if } \|\tilde{X}_{j+1}^n - \tilde{X}_j^n\| \geq \theta.
\end{cases}
$$

Since $\tilde{X}_{j+1}^n - \tilde{X}_j^n = \tilde{Y}_j^n/n$, $\xi_j^n$ is a truncation of $\tilde{Y}_j^n/n$. The definition

$$
P_x \{ \tilde{Y}_j^n \in dy | \tilde{X}_0^n, \tilde{X}_1^n, \ldots, \tilde{X}_j^n \} \doteq \nu_j^n(dy|\tilde{X}_0^n, \tilde{X}_1^n, \ldots, \tilde{X}_j^n)
$$

implies that $\nu_j^n(dy|\tilde{X}_0^n, \tilde{X}_1^n, \ldots, \tilde{X}_j^n)$ is a regular conditional distribution for $\tilde{Y}_j^n$ given $\mathcal{F}_j^n$. Let $\gamma_j^n$ denote a regular conditional distribution for $n\xi_j^n$ given $\mathcal{F}_j^n$. Since $\tilde{X}^n(0) = S^n(0) = x$, for any $\varepsilon > 0$

$$
P_x \left\{ \max_{k \in \{0,1,\ldots,n\}} \| \tilde{X}^n(k/n) - S^n(k/n) \| \geq 3\varepsilon \right\}
$$

(5.20)

$$
= P_x \left\{ \max_{k \in \{0,1,\ldots,n-1\}} \| \tilde{X}^n((k+1)/n) - S^n((k+1)/n) \| \geq 3\varepsilon \right\}
$$

$$
= P_x \left\{ \max_{k \in \{0,1,\ldots,n-1\}} \| \tilde{X}_{k+1}^n - x - \int_{H^n} y \nu_j^n(dy \times ds) \| \geq 3\varepsilon \right\}
$$

$$
= P_x \left\{ \max_{k \in \{0,1,\ldots,n-1\}} \| \tilde{X}_{k+1}^n - x - \frac{1}{n} \sum_{j=0}^{k} \int_{H^n} y \nu_j^n(dy) \| \geq 3\varepsilon \right\}
$$
\[ \begin{align*}
&\leq \bar{P}_x \left\{ \max_{k \in \{0,1,\ldots,n-1\}} \left\| \sum_{j=0}^{k} \left( \xi_j^n - \frac{1}{n} \int_{\mathbb{R}^d} y \gamma_j^n(dy) \right) \right\| \geq \varepsilon \right\} \\
&+ \bar{P}_x \left\{ \max_{k \in \{0,1,\ldots,n-1\}} \left\| X_{k+1} - x - \sum_{j=0}^{k} \xi_j^n \right\| \geq \varepsilon \right\} \\
&+ \bar{P}_x \left\{ \max_{k \in \{0,1,\ldots,n-1\}} \left\| \frac{1}{n} \sum_{j=0}^{k} \left( \int_{\mathbb{R}^d} y \gamma_j^n(dy) - \int_{\mathbb{R}^d} y \nu_j^n(dy) \right) \right\| \geq \varepsilon \right\}.
\end{align*} \]

We now bound each of the three summands appearing on the right hand side of the last inequality.

The sequence
\[ \{ \xi_j^n - \frac{1}{n} \int_{\mathbb{R}^d} y \gamma_j^n(dy), j = 0,1,\ldots,n-1 \} \]
forms a vector-valued martingale difference sequence with respect to \( \{ \tilde{F}_j, j = 0,1,\ldots,n-1 \} \). Thus for \( 0 \leq i \neq \ell \leq n-1 \),
\[ E_x \left\{ \left( \xi_i^n - \frac{1}{n} \int_{\mathbb{R}^d} y \gamma_i^n(dy), \xi_\ell^n - \frac{1}{n} \int_{\mathbb{R}^d} y \gamma_\ell^n(dy) \right) \right\} = 0, \]
and for \( k \in \{0,1,\ldots,n-1\} \) the random variables
\[ \left\| \sum_{j=0}^{k} \left( \xi_j^n - \frac{1}{n} \int_{\mathbb{R}^d} y \gamma_j^n(dy) \right) \right\|^2 \]
form a submartingale. Using the submartingale inequality [50, Lem. 2.2.3] we have for any \( \varepsilon > 0 \)
\[ \bar{P}_x \left\{ \max_{k \in \{0,1,\ldots,n-1\}} \left\| \sum_{j=0}^{k} \left( \xi_j^n - \frac{1}{n} \int_{\mathbb{R}^d} y \gamma_j^n(dy) \right) \right\| \geq \varepsilon \right\} \]
\[ \leq \frac{1}{\varepsilon^2} \bar{E}_x \left\{ \left\| \sum_{j=0}^{n-1} \left( \xi_j^n - \frac{1}{n} \int_{\mathbb{R}^d} y \gamma_j^n(dy) \right) \right\|^2 \right\} \]
\[ = \frac{1}{\varepsilon^2} \sum_{j=0}^{n-1} \bar{E}_x \left\{ \left\| \xi_j^n - \frac{1}{n} \int_{\mathbb{R}^d} y \gamma_j^n(dy) \right\|^2 \right\} \]
\[ \leq \frac{1}{\varepsilon^2} \sum_{j=0}^{n-1} \bar{E}_x \{ \| \xi_j^n \|^2 \} \]
\[ \leq \frac{\theta}{\varepsilon^2} \sum_{j=0}^{n-1} \bar{E}_x \{ \| \xi_j^n \| \} \]
\[ \leq \frac{\theta}{\varepsilon^2} \sum_{j=0}^{n-1} \bar{E}_x \{ \| \tilde{X}_{j+1} - \tilde{X}_j^n \| \} \]
\[ = \frac{\theta}{\varepsilon^2} \sum_{j=0}^{n-1} \bar{E}_x \{ \| \tilde{Y}_j^n \| \} \]
\[
= \frac{\theta}{\varepsilon^2} E_x \left\{ \int_{\mathbb{R}^d} \|y\| \nu^n_y(dy) \right\}
\]
\[
= \frac{\theta}{\varepsilon^2} E_x \left\{ \int_{\mathbb{R}^d \times [0,1]} \|y\| \nu^n(dy \times ds) \right\}.
\]

By the uniform integrability in (5.9), there exists \( M < \infty \) such that for all \( n \)
\[
\bar{P}_x \left\{ \max_{k \in \{0, 1, \ldots, n-1\}} \left\| \sum_{j=0}^{k} \left( \xi_j^n - \frac{1}{n} \int_{\mathbb{R}^d} y \gamma^n_j(dy) \right) \right\| \geq \varepsilon \right\} \leq \frac{\theta M}{\varepsilon^2}. \tag{5.21}
\]

For each \( n \) we define
\[
c_{\theta}^{(n)} = \frac{1}{n} \sum_{j=0}^{n-1} \int_{\mathbb{R}^d : \|y\| \geq n\theta} \|y\| \nu^n(dy \times ds)
\]
and note that
\[
c_{\theta}^{(n)} = E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \left\| \gamma^n_j(dy) \right\| = E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \| \gamma^n_j \|_1 \right\} \right\}.
\]

Again, the uniform integrability in (5.9) implies that for each fixed \( \theta > 0 \)
\[
\lim_{n \to \infty} c_{\theta}^{(n)} = 0.
\]

In terms of \( c_{\theta}^{(n)} \), we have the following bound on the term appearing in the next to last line of (5.20):
\[
\bar{P}_x \left\{ \max_{k \in \{0, 1, \ldots, n-1\}} \left\| \tilde{X}_{k+1}^n - x - \sum_{j=0}^{k} \xi_j^n \right\| \geq \varepsilon \right\} \tag{5.22}
\]
\[
\leq \frac{1}{\varepsilon} E_x \left\{ \max_{k \in \{0, 1, \ldots, n-1\}} \left\| \tilde{X}_{k+1}^n - x - \sum_{j=0}^{k} \xi_j^n \right\| \right\}
\]
\[
\leq \frac{1}{\varepsilon} E_x \left\{ \sum_{j=0}^{n-1} \left\| \tilde{X}_{j+1}^n - \tilde{X}_j^n - \xi_j^n \right\| \right\}
\]
\[
= \frac{1}{\varepsilon} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \left\| \gamma^n_j \right\|_1 1_{\|\gamma^n_j\| \geq n\theta} \right\}
\]
\[
= \frac{c_{\theta}^{(n)}}{\varepsilon}.
\]

We also have the following bound on the term appearing in the last line of (5.20):
\[
\bar{P}_x \left\{ \max_{k \in \{0, 1, \ldots, n-1\}} \left\| \frac{1}{n} \sum_{j=0}^{k} \left( \int_{\mathbb{R}^d} y \gamma^n_j(dy) - \int_{\mathbb{R}^d} y \nu_j^n(dy) \right) \right\| \geq \varepsilon \right\}
\]
\[ \leq \frac{1}{\varepsilon} E_x \left\{ \max_{k \in \{0, 1, \ldots, n-1\}} \left\| \frac{1}{n} \sum_{j=0}^{n-1} \left( \int_{\mathbb{R}^d} y \gamma_j^n(dy) - \int_{\mathbb{R}^d} y \nu_j^n(dy) \right) \right\| \right\} \]
\[
\leq \frac{1}{\varepsilon} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \left\| \int_{\mathbb{R}^d} y \gamma_j^n(dy) - \int_{\mathbb{R}^d} y \nu_j^n(dy) \right\| \right\} \]
\[= \frac{1}{\varepsilon} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \left\| \hat{Y}_j^n \right\| \left[ 1 \left( \left\| \hat{Y}_j^n \right\| \geq \varepsilon \right) \right] \right\} \quad (5.23) \]

By substituting into (5.20) the bounds (5.21), (5.22), and (5.23), we obtain
\[ \hat{P}_x \left\{ \max_{k \in \{0, 1, \ldots, n\}} \| \hat{X}^n(k/n) - S^n(k/n) \| \geq 3\varepsilon \right\} \leq \frac{\theta M}{\varepsilon^2} + 2\varepsilon_n(n) \varepsilon. \]

Sending \( n \to \infty \) and then \( \theta \to 0 \) yields
\[ \lim_{n \to \infty} \hat{P}_x \left\{ \max_{k \in \{0, 1, \ldots, n\}} \| \hat{X}^n(k/n) - S^n(k/n) \| \geq 3\varepsilon \right\} = 0. \]

The proof of the lemma is complete. \( \blacksquare \)

This finishes the proof of Step 3 as well as the proof of Theorem 5.3.5.

The following result will be needed several times in the sequel. It was proved as part of Step 2 in the proof of Theorem 5.3.5.

**Proposition 5.3.8.** In the random walk model we assume Condition 5.3.1. For \( n \in \mathbb{N} \) and \( x \in \mathbb{R}^d \), consider any admissible control sequence \( \{\nu^n \} \) such that
\[ \sup_{n \in \mathbb{N}} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu_j^n(\cdot) \| \mu(\cdot \| \hat{X}^n_j)) \right\} < \infty, \]
where \( \nu^n(\cdot) = \nu^n(\cdot \| \hat{X}^n_0, \hat{X}^n_1, \ldots, \hat{X}^n_n) \), and define the sequence of admissible control measures \( \{\nu^n, n \in \mathbb{N}\} \) by formulas (5.3) and (5.4). Let \( (\Omega, \mathcal{F}, P_x) \) be a probability space and assume that \( \nu^n \overset{\mathcal{D}}{\to} \nu \), where \( \nu \) is a stochastic kernel on \( \mathbb{R}^d \times [0, 1] \) given \( \Omega \). Finally, define the stochastic processes \( S^n \triangleq \{S^n(t), t \in [0, 1]\} \) and \( \hat{X} \triangleq \{\hat{X}(t), t \in [0, 1]\} \) by
\[ S^n(t) \overset{d}{=} x + \int_{\mathbb{R}^d \times [0, t]} y \nu^n(dy \times ds) \quad \text{and} \quad \hat{X}(t) \overset{d}{=} x + \int_{\mathbb{R}^d \times [0, t]} y \nu(dy \times ds). \]

Then as processes taking values in \( \mathcal{C}([0, 1] : \mathbb{R}^d) \)
\( S^n \overset{\mathcal{D}}{\to} \hat{X}. \)

In the next section, we consider a strictly weaker version of Condition 5.3.1 which is useful in some applications. Without loss of continuity, the reader may skip this section and proceed directly to Chapter 6, where the Laplace principle is proved for the random walk model with continuous statistics.
5.4 A Weaker Version of Condition 5.3.1

The main result in this chapter is Theorem 5.3.5, which proves convergence properties both of appropriate sequences of admissible control measures and of the associated controlled processes. In this section, we present a strictly weaker version of Condition 5.3.1 for which all the conclusions of that theorem continue to hold. This weaker condition is Condition 5.4.1. Theorem 5.3.5 plays a key role in the proof of the Laplace principle for the random walk model to be proved in the next chapter [Theorem 6.3.3]. This Laplace principle is proved under a number of conditions including Condition 5.3.1. If the latter condition is replaced by Condition 5.4.1, then the Laplace principle remains true. Comments on the proof under this replacement of conditions are contained in part (b) of Remark 6.3.6. Similar remarks apply to the Laplace principle that will be proved in Chapter 10 [Theorem 10.2.6]. For example, in the setting of continuous-time processes, Condition 5.3.1 rules out diffusion processes with linear drifts. Such processes would be covered under the weaker hypotheses given in the continuous-time analogue of Condition 5.4.1.

We refer to Appendix E the proof that when Condition 5.3.1 is replaced by Condition 5.4.1, all the conclusions of Theorem 5.3.5 remain true.

In order to state the weaker version of Condition 5.3.1, some notation is needed. Given a measurable function $b$ mapping $\mathbb{R}^d$ into $\mathbb{R}^d$, we define for each $x \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}^d$ the shifted cumulant generating function

$$H_b(x, \alpha) = \log \int_{\mathbb{R}^d} \exp(\alpha, y - b(x)) \mu(dy|x).$$

A function $b$ mapping $\mathbb{R}^d$ into $\mathbb{R}^d$ is said to be locally Lipschitz continuous if for each compact subset $K$ of $\mathbb{R}^d$ there exists $M < \infty$ such that for all $x$ and $y$ in $K$

$$|b(x) - b(y)| \leq M \|x - y\|.$$

**Condition 5.4.1.**

(a) There exists a locally Lipschitz continuous function $b : \mathbb{R}^d \mapsto \mathbb{R}^d$ such that for each $\alpha \in \mathbb{R}^d$

$$\sup_{x \in \mathbb{R}^d} H_b(x, \alpha) < \infty.$$

(b) For any point $x_0 \in \mathbb{R}^d$ and any continuous function $y$ mapping $[0, 1]$ into $\mathbb{R}^d$, the equation

$$x(t) = x_0 + \int_0^t b(x(s)) \, ds + y(t)$$

has at least one solution $\{x(t), t \in [0, 1]\}$.

**Remark 5.4.2.** (i) Condition 5.4.1 is strictly weaker than Condition 5.3.1 since we can always take $b(x) \equiv 0$. In this case, $H_b(x, \alpha)$ reduces to the standard cumulant generating function $H(x, \alpha)$.

(ii) The local Lipschitz continuity of $b$ and Gronwall’s Inequality [Theorem A.6.4] imply that the solution $\{x(t), t \in [0, 1]\}$ in part (b) is unique. In addition, we note that for each $x$ and $\alpha$ in $\mathbb{R}^d$

$$H(x, \alpha) = H_b(x, \alpha) + \langle \alpha, b(x) \rangle.$$
5.4. A WEAKER VERSION OF CONDITION 5.3.1

Hence if \( b \) is bounded on \( \mathbb{R}^d \), then part (a) of Condition 5.4.1 yields Condition 5.3.1 and thus part (b) of Condition 5.4.1 is unnecessary. In this case, we may as well take \( b \equiv 0 \). The conclusion is that unless \( b \) is unbounded on \( \mathbb{R}^d \), Condition 5.4.1 offers no improvement over Condition 5.3.1.

(iii) A natural first choice for \( b(x) \) is the mean \( \tilde{b}(x) \equiv \int_{\mathbb{R}^d} y \mu(dy|x) \), provided that this quantity exists. For example, if for each \( x \in \mathbb{R}^d \), \( \mu(\cdot|x) \) is a Gaussian measure on \( \mathbb{R}^d \) with mean vector \( \bar{b}(x) \) and constant covariance matrix \( \Sigma \), then

\[
H(x, \alpha) = \langle \alpha, \tilde{b}(x) \rangle + \frac{1}{2}(\Sigma \alpha, \alpha) \quad \text{and} \quad H_\beta(x, \alpha) = \frac{1}{2}(\Sigma \alpha, \alpha).
\]

In this case, if \( \tilde{b} \) is locally Lipschitz continuous, then part (a) of Condition 5.4.1 holds. More generally, part (a) is valid if there exists a locally Lipschitz continuous function \( b \) such that \( \tilde{b} - b \) is bounded on \( \mathbb{R}^d \).

A sufficient condition that guarantees part (b) of Condition 5.4.1 is that \( b \) is globally Lipschitz continuous. An alternative sufficient condition that guarantees part (b) of Condition 5.4.1 is that \( b \) is locally Lipschitz continuous and that \( b \) is stable in a suitable sense; e.g., there exists a continuously differentiable function \( U \) mapping \( \mathbb{R}^d \) into \( \mathbb{R}^d \) such that

\[
\lim_{c \to \infty} \inf_{\{x \in \mathbb{R}^d : \|x\| \geq c\}} U(x) = \infty
\]

and \( \langle b(x), \nabla U(x) \rangle \leq -\varepsilon < 0 \) for all \( x \in \mathbb{R}^d \) lying outside some compact set. Thus \( U \) is a Lyapunov function. An example is \( b(x) \equiv -x^3 \) for \( x \in \mathbb{R} \), for which we may choose the Lyapunov function \( U(x) = x^2 \).

In the next chapter, we prove the Laplace principle for the random walk model with continuous statistics.
Chapter 6

Laplace Principle for the Random Walk Model with Continuous Statistics

6.1 Introduction

In Section 4.3 we introduced a random walk \( X^n = \{X^n(t), t \in [0, 1]\} \) having continuous paths in \( \mathbb{R}^d \) and found a stochastic control representation formula for

\[ W^n(x) = -\frac{1}{n} \log E_x \{ \exp [-n h(X^n)] \}, \]

where \( h \) is any bounded measurable function mapping \( C([0, 1] : \mathbb{R}^d) \) into \( \mathbb{R} \). The representation formula for \( W^n(x) \) was expressed in terms of the minimal cost function \( V^n(x) \) of a stochastic optimal control problem, as noted in Theorem 4.3.1. Our aim in this chapter is to use the representation formula in order to prove the Laplace principle for the sequence \( \{X^n, n \in \mathbb{N}\} \). The Laplace principle was proved for a special case of the random walk model in Chapter 3, where we assumed that the increments of the random walk are an i.i.d. sequence of random vectors. In contrast, in the more general model of Section 4.3, the distributions of the increments have a spatial dependence. This dependence leads to some complications in the proof of the Laplace principle, particularly in the lower bound.

After introducing a number of conditions, we will formulate the Laplace principle in Theorem 6.3.3. These conditions include a natural boundedness and continuity condition, Condition 6.2.1, on the stochastic kernel \( \mu(dy|x) \) defining the model. The term “continuous statistics” in the title of this chapter reflects the continuity of the mapping \( x \mapsto \mu(\cdot|x) \) that is part of Condition 6.2.1.

Our proof of the Laplace principle given here follows the general program outlined in Chapter 3. The proof of the upper bound starts with the representation formula \( W^n(x) = V^n(x) \) given in Theorem 4.3.1. We then consider an appropriate sequence of admissible control measures having bounded running costs, and we apply Theorem 5.3.5, which gives compactness and limit properties of the control measures and the associated controlled random walks. A straightforward calculation allows us to identify a rate function \( I_x \) on
$C([0, 1]: \mathbb{R}^d)$ for which the upper bound
\[
\limsup_{n \to \infty} \frac{1}{n} E_x \{ \exp[-n h(X^n)] \} \leq \inf_{\varphi \in C([0, 1]; \mathbb{R}^d)} \{ I_x(\varphi) + h(\varphi) \}
\]
is valid. The beauty of this approach is that the definition of the rate function need not be known ahead of time; the calculation literally hands it to us. The form of the rate function indicates how to construct a sequence of controls that are nearly optimal with respect to the infimum appearing in the definition of the minimal cost function $V^n(x)$. Using this sequence and the representation formula $W^n(x) = V^n(x)$, we prove the lower bound with the same rate function.

In Section 6.2 the Laplace principle upper bound is proved and the form of the rate function is identified. In Proposition 6.2.4 it is shown that the rate function has compact level sets. This proof exhibits an economy of the weak convergence approach in that it is a deterministic analogue of the method used to show the Laplace principle upper bound. In Section 6.3 we state the Laplace principle and discuss the conditions under which the lower bound will be proved. The proof of the latter is more difficult than the proof of the upper bound. The general strategy is described in Section 6.4, and the proof itself is carried out under two different sets of conditions in Sections 6.5 and 6.6. Finally, in Section 6.7 we present an extension of the Laplace principle which holds for a wider class of random walks than was considered previously. In Chapter 10 we will prove the Laplace principle for continuous-time diffusion processes and jump diffusion processes taking values in $\mathbb{R}^d$ and having continuous statistics. The proofs are based on an approximation argument that allows us to apply the extension of the Laplace principle given in Section 6.7.

Large deviation principles for the models considered in the present chapter have been studied in numerous works including [4, 28, 72, 89, 90, 91, 92, 93]. All of these works require assumptions that are more involved than those we will use and that are difficult to verify except in special cases.

### 6.2 Proof of the Laplace Principle Upper Bound and the Identification of the Rate Function

The random walk model was introduced in Section 4.3, and its compactness and limit properties were derived in Chapter 5. We will make use of these properties in the present section. The reader will recall that the random walk model is defined in terms of a stochastic kernel $\mu(dy|x)$ on $\mathbb{R}^d$ given $\mathbb{R}^d$. The distributions of the i.i.d. random vector fields $v_j(x)$ are given in terms of this stochastic kernel by $P\{v_j(x) \in dy\} \equiv \mu(dy|x)$. For $x \in \mathbb{R}^d$ and $n \in \mathbb{N}$ we then define $X^n_0 \equiv x$,

\[
X_j^n = X_j^{n-1} + \frac{1}{n} v_j(X^n_{j-1}) \text{ for } j \in \{0, 1, \ldots, n-1\},
\]

and

\[
X^n(t) = X_j^n + \left( t - \frac{j}{n} \right) v_j(X^n_{j-1}) \text{ for } t \in [j/n, (j+1)/n], \quad j = 0, 1, \ldots, n-1.
\]
6.2. PROOF OF UPPER BOUND

The piecewise linear interpolation $X^n = \{X^n(t), t \in [0, 1]\}$ takes values in $C([0, 1] : \mathbb{R}^d)$.

The next condition will be needed in the proofs of both of the Laplace principle bounds for the random walks $X^n$. For each $x$ and $\alpha$ in $\mathbb{R}^d$, $H(x, \alpha)$ denotes the cumulant generating function

$$H(x, \alpha) = \log \int_{\mathbb{R}^d} \exp(\alpha, y) \mu(dy|x).$$

(6.3)

**Condition 6.2.1.**

(a) For each $\alpha \in \mathbb{R}^d$, $\sup_{x \in \mathbb{R}^d} H(x, \alpha) < \infty$.

(b) The function mapping $x \in \mathbb{R}^d \mapsto \mu(\cdot|x) \in \mathcal{P}(\mathbb{R}^d)$ is continuous in the topology of weak convergence on $\mathcal{P}(\mathbb{R}^d)$.

Part (a) coincides with Condition 5.3.1, which is the hypothesis needed to prove the compactness and convergence results stated in Proposition 5.3.2 and Theorem 5.3.5.

Before proving the upper bound, we need to introduce some additional notation. The notation will be used to write the rate function in its conventional form for these process-level problems. We define $L(x, \cdot)$ to be the Legendre--Fenchel transform of $H(x, \cdot)$ with respect to the second variable; namely, for $x$ and $\beta$ in $\mathbb{R}^d$

$$L(x, \beta) = \sup_{\alpha \in \mathbb{R}^d} \{\langle \alpha, \beta \rangle - H(x, \alpha)\}.$$

(6.4)

In the weak convergence approach to large deviations, every rate function that we deal with can be written in terms of relative entropy. The function $L$ appears naturally in the rate function for the processes of this chapter because of the following variational formula, valid for each $x$ and $\beta$ in $\mathbb{R}^d$ [Lemma 6.2.3 (f)]:

$$L(x, \beta) = \inf \left\{ R(\gamma(\cdot)||\mu(\cdot|x)) : \gamma \in \mathcal{P}(\mathbb{R}^d), \int_{\mathbb{R}^d} y \gamma(dy) = \beta \right\}.$$

If $\gamma$ is a probability measure on $\mathbb{R}^d$ satisfying $R(\gamma(\cdot)||\mu(\cdot|x)) < \infty$, then by part (e) of Lemma 1.4.3 $\int_{\mathbb{R}^d} ||y|| \gamma(dy) < \infty$. Thus the last display implies that

$$R(\gamma(\cdot)||\mu(\cdot|x)) \geq L(x, \int_{\mathbb{R}^d} y \gamma(dy)).$$

(6.5)

Fix $x \in \mathbb{R}^d$ and let $h$ be any bounded continuous function mapping $C([0, 1] : \mathbb{R}^d)$ into $\mathbb{R}$. Our first goal is to prove a Laplace principle upper bound having the form

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{E}_x \{\exp[-n h(X^n)]\} \leq - \inf_{\varphi \in C([0, 1]: \mathbb{R}^d)} \{I_x(\varphi) + h(\varphi)\},$$

where the rate function $I_x$ must be identified. This is equivalent to proving the lower limit

$$\liminf_{n \to \infty} W^n(x) \geq \inf_{\varphi \in C([0, 1]: \mathbb{R}^d)} \{I_x(\varphi) + h(\varphi)\}.$$

(6.6)
After identifying the form of the rate function, we will verify that it has compact level sets. The key to the proof of the lower limit (6.6) is to use the stochastic control representation formula for $W^n(x)$ given in Theorem 4.3.1 and Corollary 5.2.1. This representation formula states that

$$W^n(x) = V^n(x)$$

$$\geq \inf_{\{\nu^n\}} \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot)\|\mu(\cdot|\bar{X}^n_j)) + h(\bar{X}^n) \right\}$$

$$= \inf_{\{\nu^n_j\}} \mathbb{E}_x \left\{ \int_0^1 R(\nu^n(\cdot|t)\|\mu(\cdot|\bar{X}^n(t))) \, dt + h(\bar{X}^n) \right\}.$$  \hspace{1cm} (6.7)

The infima in this display are each taken over all admissible control sequences $\{\nu^n_j\}$, each $\nu^n_j$ being a stochastic kernel on $\mathbb{R}^d$ given $(\mathbb{R}^d)^{j+1}$. Given an admissible control sequence we defined a sequence of controlled random vectors $\{\bar{X}^n_j\}$ by $\bar{X}^n_0 = X_0^n = x$ and

$$\bar{X}^n_{j+1} = \bar{X}^n_j + \frac{1}{n} \bar{Y}^n_j,$$

where

$$\bar{P}_x \{ \bar{Y}^n_j \in dy | \bar{X}^n_0, \bar{X}^n_1, \ldots, \bar{X}^n_j \} = \nu^n_j(dy|\bar{X}^n_0, \bar{X}^n_1, \ldots, \bar{X}^n_j).$$

The process $\bar{X}^n = \{\bar{X}^n(t), t \in [0,1]\}$, defined in equation (5.6), is the piecewise linear interpolation of these controlled random vectors. $\mathbb{E}_x$ denotes expectation with respect to $\bar{P}_x$, and the quantity $\nu^n(dy|t)$ appearing as the first argument of the relative entropy equals $\nu^n_j(\cdot|\bar{X}^n_0, \bar{X}^n_1, \ldots, \bar{X}^n_j)$. The formula in the third line of (6.7) involves a number of quantities that were introduced in Section 5.2. The stochastic kernel $\nu^n(dy|t)$ is defined in terms of an admissible control sequence $\{\nu^n_j, j = 0, 1, \ldots, n-1\}$ by

$$\nu^n(dy|t) = \begin{cases} 
\nu^n_j(dy) & \text{if } t \in \left[ j/n, (j+1)/n \right], j = 0, 1, \ldots, n-2 \\
\nu^n_{n-1}(dy) & \text{if } t \in \left[ (n-1)/n, 1 \right],
\end{cases}$$

where $\nu^n_j(dy) = \nu^n_j(dy|\bar{X}^n_0, \bar{X}^n_1, \ldots, \bar{X}^n_j)$. The process $\bar{X}^n = \{\bar{X}^n(t), t \in [0,1]\}$, defined in formula (5.5), is the piecewise constant interpolation of the controlled random vectors $\{\bar{X}^n_j, j = 0, 1, \ldots, n\}$. We also introduce the admissible control measure $\nu^n = \nu^n(dy \times dt) \equiv \nu^n(dy|t) \otimes dt$.

From its definition $|W^n(x)| \leq \|h\|_\infty$. It suffices to prove the lower limit (6.6) when $n$ is replaced by any subsequence along which the functions $W^n(x)$ converge. We will work with a fixed such subsequence for the remainder of the proof, indexing it, as usual, by $n \in \mathbb{N}$.

Given $\varepsilon > 0$, let $\{\nu^n, n \in \mathbb{N}\}$ be a sequence of admissible control measures satisfying for each $n$

$$V^n(x) + \varepsilon \geq \mathbb{E}_x \left\{ \int_0^1 R(\nu^n(\cdot|t)\|\mu(\cdot|\bar{X}^n(t))) \, dt + h(\bar{X}^n) \right\}. \hspace{1cm} (6.8)$$

In order to study the asymptotic properties of $\{\nu^n\}$, we would like to apply the convergence results stated in Theorem 5.3.5. Part (a) of Condition 6.2.1 coincides with Condition 5.3.1,
which is a hypothesis in this theorem. Since \( \sup_{n \in \mathbb{N}} V^n(x) = \sup_{n \in \mathbb{N}} W^n(x) \leq \|h\|_\infty \), the other hypothesis of the theorem is satisfied; namely the boundedness condition

\[
\sup_{n \in \mathbb{N}} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R\left( \nu^n_j(\cdot)\mu(\cdot | X^n_j) \right) \right\} < \infty
\]

on the sequence of running costs associated with \( \{\nu^n\} \). According to Theorem 5.3.5, there exists a subsequence of \( \{(\nu^n, X^n, \tilde{X}^n), n \in \mathbb{N}\} \) and a probability space \((\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}_x)\) such that along the subsequence

\[
(\nu^n, X^n, \tilde{X}^n) \xrightarrow{P} (\nu, X, \tilde{X})
\]

and such that \( \nu = \nu(dy \times dt) \) is a stochastic kernel on \( \mathbb{R}^d \times [0,1] \) given \( \tilde{\Omega} \). \( \tilde{X} = \{ \tilde{X}(t), t \in [0,1] \} \) maps \( \tilde{\Omega} \) into \( \mathcal{C}([0,1] : \mathbb{R}^d) \), and w.p.1 \( \tilde{X}(t) \) is an absolutely continuous function of \( t \in [0,1] \). As we stated in Lemma 5.3.4, there exists a stochastic kernel \( \nu(dy|t) \) on \( \mathbb{R}^d \) given \( [0,1] \times \tilde{\Omega} \) such that \( \tilde{P}_x \)-a.s., for \( \omega \in \tilde{\Omega} \)

\[
\nu(dy|t) = \nu(dy \times dt) = \nu(dy|t) \otimes dt.
\]

Furthermore, w.p.1 we have a.s. for \( t \in [0,1] \)

\[
\tilde{X}(t) = \int_{\mathbb{R}^d} y \nu(dy|t).
\]  

(6.9)

By the Skorohod Representation Theorem [Theorem A.3.9], we can assume without loss of generality that when we calculate limits of expectations involving the convergent subsequence

\[
(\nu^n, X^n, \tilde{X}^n) \xrightarrow{P} (\nu, X, \tilde{X}),
\]

the convergence occurs w.p.1. Invoking the Skorohod Representation Theorem requires that we pay attention to the metrics of the Polish spaces in which the various processes take values; for \( X^n \) and \( \tilde{X} \) the space is \( \mathcal{D}([0,1] : \mathbb{R}^d) \) while for \( X^n \) the space is \( \mathcal{D}([0,1] : \mathbb{R}^d) \). Since \( X^n \to X \) in the metric of \( \mathcal{D}([0,1] : \mathbb{R}^d) \) and since \( X \) is continuous w.p.1, it follows that \( X^n \to X \) uniformly w.p.1 [Theorem A.6.5 (c)]; i.e., \( \sup_{t \in [0,1]} \| \tilde{X}(t) - X^n(t) \| \to 0 \).

Our aim is to evaluate the limit inferior of \( W^n(x) \) using formula (6.8). As in the proof of Mogulskii’s Theorem in Section 3.3, it is convenient to rewrite the running cost in (6.8) as a relative entropy involving the admissible control measure \( \nu^n(dy \times dt) = \nu^n(dy|t) \otimes dt \). Let \( \lambda \) denote Lebesgue measure on \([0,1]\). Then part (f) of Lemma 1.4.3 yields

\[
\int_0^1 R\left( \nu^n(\cdot|t)\mu(\cdot | X^n(t)) \right) dt = R\left( \nu^n(dy \times dt)\mu(dy | X^n(t)) \otimes \lambda(dt) \right).
\]

We now evaluate the limit inferior of \( W^n(x) \) along the subsequence of \( n \in \mathbb{N} \) for which \( (\nu^n, X^n, \tilde{X}^n) \to (\nu, X, \tilde{X}) \). This evaluation is carried out in the following string of inequalities, each line of which will be explained afterwards:

\[
\lim_{n \to \infty} W^n(x) + \varepsilon = \lim_{n \to \infty} V^n(x) + \varepsilon
\]

\[
\geq \liminf_{n \to \infty} E_x \left\{ \int_0^1 R\left( \nu^n(\cdot|t)\mu(\cdot | X^n(t)) \right) dt + h(X^n) \right\}
\]  

(6.10)
\begin{align*}
&= \liminf_{n \to \infty} \mathbb{E}_x \left\{ R\left( \nu^n(dy \times dt) \| \mu(dy|\bar{X}^n(t)) \otimes \lambda(dt) \right) + h(\bar{X}^n) \right\} \\
&\geq \mathbb{E}_x \left\{ R\left( \nu(dy \times dt) \| \mu(dy|\bar{X}(t)) \otimes \lambda(dt) \right) + h(\bar{X}) \right\} \\
&= \mathbb{E}_x \left\{ \int_0^1 R\left( \nu(\cdot|t) \| \mu(\cdot|\bar{X}(t)) \right) dt + h(\bar{X}) \right\} \\
&\geq \mathbb{E}_x \left\{ \int_0^1 L\left( \bar{X}(t), \int_{\mathbb{R}^d} y \nu(dy|t) \right) dt + h(\bar{X}) \right\}.
\end{align*}

The first two lines of this display are given in formulas (6.7) and (6.8), and as noted above, line three is a consequence of part (f) of Lemma 1.4.3. We now collect the facts that are needed to obtain line four. With probability 1 \( \nu^n \Rightarrow \nu \). In addition, the function mapping \( x \in \mathbb{R}^d \mapsto \mu(dy|x) \in \mathcal{P}(\mathbb{R}^d) \) is continuous [Condition 6.2.1 (b)], and w.p.1 the sequence \( \{ \bar{X}^n, n \in \mathbb{N} \} \) converges to \( \bar{X} \) uniformly on \([0,1] \). Hence there exists a null set on the complement of which

\[ \mu(dy|\bar{X}^n(t)) \Rightarrow \mu(dy|\bar{X}(t)) \]

for each \( t \in [0,1] \). Theorem A.5.8 yields, on the complement of the same null set

\[ \mu(dy|\bar{X}^n(t)) \otimes \lambda(dt) \Rightarrow \mu(dy|\bar{X}(t)) \otimes \lambda(dt). \]

Since \( R(\cdot||\cdot) \) is lower semicontinuous [Lemma 1.4.3 (b)], we obtain w.p.1

\[ \liminf_{n \to \infty} R\left( \nu^n(dy \times dt) \| \mu(dy|\bar{X}^n(t)) \otimes \lambda(dt) \right) \geq R\left( \nu(dy \times dt) \| \mu(dy|\bar{X}(t)) \otimes \lambda(dt) \right) \].

Finally, since \( h \) is continuous on \( C([0,1]: \mathbb{R}^d) \) and w.p.1 the sequence \( \{ \bar{X}^n, n \in \mathbb{N} \} \) converges uniformly to \( \bar{X} \), we also have w.p.1

\[ \lim_{n \to \infty} h(\bar{X}^n) = h(\bar{X}). \]

Line four in (6.10) now follows from Fatou’s Lemma. Line five is a consequence of part (f) of Lemma 1.4.3 and the decomposition \( \nu(dy \times dt) = \nu(dy|t) \otimes \lambda(dt) \) given in Lemma 5.3.4. Line six is implied by the lower bound (6.5) involving relative entropy and \( L(x, \beta) \). Finally, the representation given in equation (6.9) yields the last line in the display.

Sending \( \varepsilon \to 0 \), we have proved that for the given convergent subsequence of \( W^n(x) \) there exists a subsubsequence for which

\[ \liminf_{n \to \infty} W^n(x) \geq \mathbb{E}_x \left\{ \int_0^1 L(\bar{X}(t), \dot{\bar{X}}(t)) dt + h(\bar{X}) \right\}. \tag{6.11} \]

This formula leads naturally to the definition of the rate function \( I_x(\varphi) \). For absolutely continuous functions \( \varphi \in C([0,1]: \mathbb{R}^d) \) satisfying \( \varphi(0) = x \), we define

\[ I_x(\varphi) \doteq \int_0^1 L(\varphi(t), \dot{\varphi}(t)) dt, \tag{6.12} \]
where \( L \) is the Legendre–Fenchel transform defined in (6.4). For all other \( \varphi \in C([0,1] : \mathbb{R}^d) \), set \( I_x(\varphi) = \infty \). Equation (6.11) yields the lower limit
\[
\liminf_{n \to \infty} W^n(x) \geq \inf_{\varphi \in C([0,1] : \mathbb{R}^d)} \{ I_x(\varphi) + h(\varphi) \}.
\]
We have proved that every convergent subsequence of the original sequence \( \{W^n(x), n \in \mathbb{N}\} \) has a subsequence satisfying this lower limit. An argument by contradiction applied to an arbitrary convergent subsequence establishes the same lower limit for the entire sequence \( \{W^n(x), n \in \mathbb{N}\} \). This proves the Laplace principle upper bound, which is stated in the next proposition. In Proposition 6.2.4 we will verify that \( I_x \) has compact level sets.

**Proposition 6.2.2.** We assume Condition 6.2.1. For \( x \in \mathbb{R}^d \) let \( I_x \) be the function on \( C([0,1] : \mathbb{R}^d) \) just defined. Then \( I_x \) is a rate function, and for all bounded continuous functions \( h \) mapping \( C([0,1] : \mathbb{R}^d) \) into \( \mathbb{R} \) we have the Laplace principle upper bound
\[
\limsup_{n \to \infty} \frac{1}{n} \log E_x \{ \exp[-n h(X^n)] \} \leq \inf_{\varphi \in C([0,1] : \mathbb{R}^d)} \{ I_x(\varphi) + h(\varphi) \}.
\]

We will prove that \( I_x(\varphi) \) is a rate function after we establish a number of properties of the functions \( H(x, \alpha) \) and \( L(x, \beta) \) defined in equations (6.3) and (6.4). Many of these properties are well known in the context of Cramér’s Theorem, which states the Laplace principle for sums of i.i.d. random vectors taking values in \( \mathbb{R}^d \) [Theorem 3.5.1]. The rate function in Cramér’s Theorem is the Legendre–Fenchel transform of the cumulant generating function of the underlying distribution of the random vectors and so is the analogue of \( L(x, \beta) \) in the current setup. Background on these topics can be found in Chapters VI—VIII of [45]. In the sequel we will refer to any function of the form of \( L(x, \beta) \) as a Cramér function.

The proof that \( I_x \) is a rate function uses only parts (b) and (c) of the following lemma. The remaining parts will be used in the proof of the Laplace principle lower bound. Part (g) is particularly important. There we construct a stochastic kernel that solves the variational formula
\[
L(x, \beta) = \inf \left\{ R(\gamma(\cdot)\|\mu(\cdot|x)) : \gamma \in \mathcal{P}(\mathbb{R}^d), \int_{\mathbb{R}^d} y \gamma(dy) = \beta \right\},
\]
the use of which we have just seen in the proof of the upper bound. Although the lemma is stated under Condition 6.2.1, the full force of this condition is needed only to show parts (a) and (b). Part (c) requires only Condition 6.2.1 (a). In order to prove the remaining parts, it suffices merely if the function \( H(x, \alpha) \) defined in equation (6.3) is finite for each \( x \) and \( \alpha \) in \( \mathbb{R}^d \). Since the proofs of parts (f) and (g) are lengthy, they are deferred to Appendix C.

In order to state the lemma, some notation and several definitions are needed. Given a probability measure \( \mu \) on \( \mathbb{R}^d \), \( S_\mu \) denotes the support of \( \mu \), which is the smallest closed set having \( \mu \)-probability 1. For a subset \( A \) of \( \mathbb{R}^d \), conv \( A \), aff \( A \), cl \( A \) and int \( A \) denote the convex hull of \( A \), the affine hull of \( A \), the closure of \( A \), and the interior of \( A \), respectively.
For a convex subset $C$ of $\mathbb{R}^d$, $\text{ri } C$ denotes the relative interior of $C$, which is the interior that results when $C$ is regarded as a subset of $\text{aff } C$. Of course, if the interior of $C$ is nonempty, or equivalently if the affine hull of $C$ equals $\mathbb{R}^d$, then $\text{ri } C = \text{int } C$. Finally, for $f$ a convex function on $\mathbb{R}^d$ $\text{dom } f$ denotes the effective domain of $f$, which is the set of $x \in \mathbb{R}^d$ for which $f(x) < \infty$. Background material on convex functions is given in Section D.2.

**Lemma 6.2.3.** Under Condition 6.2.1, the function $H(x, \alpha)$ defined in equation (6.3) and the function $L(x, \beta)$ defined in equation (6.4) have the following properties.

(a) For each $x \in \mathbb{R}^d$, $H(x, \alpha)$ is a finite convex function of $\alpha \in \mathbb{R}^d$ which is differentiable for all $\alpha$. In addition, $H(x, \alpha)$ is a continuous function of $(x, \alpha) \in \mathbb{R}^d \times \mathbb{R}^d$.

(b) For each $x \in \mathbb{R}^d$, $L(x, \beta)$ is a convex function of $\beta \in \mathbb{R}^d$. In addition, $L(x, \beta)$ is a nonnegative, lower semicontinuous function of $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$.

(c) $L(x, \beta)$ is uniformly superlinear in the sense that

$$
\lim_{N \to \infty} \inf_{x \in \mathbb{R}^d} \left( \inf_{\beta \in \mathbb{R}^d, \|\beta\| = N} \frac{1}{\|\beta\|} L(x, \beta) \right) = \infty.
$$

(d) For each $x \in \mathbb{R}^d$ $\text{ri}(\text{dom } L(x, \cdot))$ equals $\text{ri}(\text{conv } S_{\mu_{\gamma}(\cdot)})$. In particular, $L(x, \beta)$ equals $\infty$ for $x \in \mathbb{R}^d$ and $\beta \in (\text{cl}(\text{conv } S_{\mu_{\gamma}(\cdot)})^c$. For each $\beta \in \text{ri}(\text{dom } L(x, \cdot)) = \text{ri}(\text{conv } S_{\mu_{\gamma}(\cdot)})$ there exists $\alpha = \alpha(x, \beta) \in \mathbb{R}^d$ such that $\nabla_{\alpha} H(x, \alpha(x, \beta)) = \beta$. In addition

$$
L(x, \beta) = \langle \alpha(x, \beta), \beta \rangle - H(x, \alpha(x, \beta)).
$$

(e) Suppose in addition that for a given $x \in \mathbb{R}^d$ $\text{conv } S_{\mu_{\gamma}(\cdot)}$ has nonempty interior. Then $H(x, \alpha)$ is a strictly convex function of $\alpha \in \mathbb{R}^d$; $\text{int}(\text{dom } L(x, \cdot)) = \text{int}(\text{conv } S_{\mu_{\gamma}(\cdot)})$ is nonempty; for each $\beta \in \text{int}(\text{dom } L(x, \cdot))$ there exists a unique value of $\alpha = \alpha(x, \beta) \in \mathbb{R}^d$ such that $\nabla_{\alpha} H(x, \alpha(x, \beta)) = \beta$; and $L(x, \cdot)$ is differentiable on $\text{int}(\text{dom } L(x, \cdot))$.

(f) For each $x$ and $\beta$ in $\mathbb{R}^d$

$$
L(x, \beta) = \inf \left\{ \nabla_{\alpha} \|\gamma(\cdot)|\mu(\cdot|x)\| : \gamma \in \mathcal{P}(\mathbb{R}^d), \int_{\mathbb{R}^d} \gamma(dy) = \beta \right\},
$$

and the infimum is always attained. If $L(x, \beta) < \infty$, then the infimum is uniquely attained at some measure $\gamma$.

(g) There exists a stochastic kernel $\gamma(dy|x)$ on $\mathbb{R}^d$ given $\mathbb{R}^d$ satisfying for $x$ and $\beta$ in $\mathbb{R}^d$

$$
R(\gamma(\cdot)|\mu(\cdot|x)) = L(x, \beta) \quad \text{and} \quad \int_{\mathbb{R}^d} \gamma(dy|x) = \beta.
$$

In fact $\gamma(dy|x)$ can be chosen to be a measurable function $\gamma(dy|x, \beta)$ of both $x$ and $\beta$ in $\mathbb{R}^d$.

(h) If $\gamma \in \mathcal{P}(\mathbb{R}^d)$ satisfies $R(\gamma(\cdot)|\mu(\cdot|x)) < \infty$ for $x \in \mathbb{R}^d$, then $\int_{\mathbb{R}^d} \|y\| \gamma(dy) < \infty$ and

$$
R(\gamma(\cdot)|\mu(\cdot|x)) \geq L \left( x, \int_{\mathbb{R}^d} \gamma(dy) \right).
$$
6.2. PROOF OF UPPER BOUND

Proof. (a) Condition 6.2.1 guarantees that $H(x, \alpha)$ is finite for all $x$ and $\alpha$ in $\mathbb{R}^d$. The convexity of $H(x, \cdot)$ follows from Hölder’s inequality. The differentiability and continuity properties of $H$ are consequences of Condition 6.2.1 and the Lebesgue Dominated Convergence Theorem.

(b) The nonnegativity of $L(x, \beta)$ follows from the dual formula [Theorem D.2.6 (e)]

$$H(x, \alpha) = \sup_{\beta \in \mathbb{R}^d} \{ \langle \alpha, \beta \rangle - L(x, \beta) \},$$

which implies that

$$\inf_{\beta \in \mathbb{R}^d} L(x, \beta) = -H(x, 0) = 0.$$

According to part (a) of the present lemma, $H(x, \alpha)$ is a continuous function of $(x, \alpha) \in \mathbb{R}^d \times \mathbb{R}^d$. Since $L(x, \beta)$ is defined as a supremum, over $\alpha \in \mathbb{R}^d$, of a family of functions that are convex in $\beta$ for each $x$ and are continuous in $(x, \beta)$, it follows that $L(x, \beta)$ is a convex function of $\beta \in \mathbb{R}^d$ for each $x \in \mathbb{R}^d$ and is a lower semicontinuous function of $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$.

(c) For $\alpha \in \mathbb{R}^d$ we define

$$\bar{H}(\alpha) \doteq \sup_{x \in \mathbb{R}^d} H(x, \alpha).$$

According to part (a) of Condition 6.2.1, $\bar{H}(\alpha)$ is a finite convex function on $\mathbb{R}^d$ and thus is continuous [Theorem D.2.2 (b)]. For $M \in (0, \infty)$ and nonzero $\beta \in \mathbb{R}^d$, set $\alpha = M\beta/\|\beta\|$. Then for any $x \in \mathbb{R}^d$

$$L(x, \beta) \geq \langle \alpha, \beta \rangle - H(x, \alpha) \geq M\|\beta\| - \bar{H}(M\|\beta\|),$$

and so for all sufficiently large $N \in (0, \infty)$

$$\inf_{x \in \mathbb{R}^d} \inf_{\{\beta \in \mathbb{R}^d; \|\beta\| = N\}} \frac{1}{\|\beta\|} L(x, \beta) \geq M \frac{1}{N} \sup_{\{\alpha \in \mathbb{R}^d; \|\alpha\| = M\}} \bar{H}(\alpha) \geq \frac{M}{2}.$$ 

Since $M$ is arbitrary, the proof is done.

(d) In the case where $\text{aff} S_{\mu(x)} = \mathbb{R}^d$, the relative interior of $S_{\mu(x)}$ equals its interior. Theorem VIII.4.3 in [45] proves the first three assertions in part (d) as stated but with $\text{ri}(\text{conv} S_{\mu(x)})$ replaced by $\text{int}(\text{cc} S_{\mu(x)})$, where $\text{cc} S_{\mu(x)}$ denotes the closed convex hull of $S_{\mu(x)}$. By Theorem D.2.1 $\text{int}(\text{cc} S_{\mu(x)})$ equals $\text{ri}(\text{conv} S_{\mu(x)})$, and so the first three assertions in part (d) follow. In the case where $\text{aff} S_{\mu(x)}$ is a proper subset of $\mathbb{R}^d$, the proofs of the first three assertions in part (d) can be reduced to the earlier case by using a standard technique outlined in Problem VIII.6.9 of [45] (see Theorem VIII.4.4 in that reference). The third assertion in part (d) is also a consequence of a general fact, given in Theorem D.2.7, about the range of the gradient of a differentiable convex function. The equality

$$L(x, \beta) = \langle \alpha(x, x), \beta \rangle - H(x, \alpha(x, \beta))$$

is a consequence of part (b) of Theorem D.2.5 and part (c) of Theorem D.2.6.

(e) The set $\text{conv} S_{\mu(x)}$ has nonempty interior if and only if $\text{aff} S_{\mu(x)} = \mathbb{R}^d$. Under this assumption, $\text{int}(\text{dom} L(x, \cdot))$ equals $\text{int}(\text{conv} S_{\mu(x)})$ and is nonempty [part (d)]. In
addition, Proposition VIII.4.2 and Theorem VIII.4.3 in [45] prove that $H(x, \alpha)$ is a strictly convex function of $\alpha \in \mathbb{R}^d$ and that for each $\beta \in \text{int}(\text{dom} \, L(x, :)) = \text{int}(\text{conv} \, S_{\mu(x)})$ there exists a unique value of $\alpha = \alpha(x, \beta) \in \mathbb{R}^d$ such that $\nabla_{\alpha} H(x, \alpha(x, \beta)) = \beta$. The strict convexity of $H(x, \cdot)$ on $\mathbb{R}^d$ implies that $L(x, \cdot)$ is differentiable on $\text{int}(\text{dom} \, L(x, \cdot))$ [Theorem D.2.8].

(f) This is proved in Section C.5.

(g) This is proved in Section C.6.

(h) If $R(\gamma(\cdot)||\mu(\cdot|x)) < \infty$, then part (e) of Lemma 1.4.3 implies that $\int_{\mathbb{R}^d} \|y\| \gamma(dy) < \infty$. Hence part (h) of the present lemma is a consequence of the variational formula in part (f). The proof of the lemma is complete. \hfill \blacksquare

We end this section by showing that the function $I_x$ defined before the statement of Proposition 6.2.2 is a rate function. Since $L(x, \beta)$ is nonnegative, $I_x$ is also nonnegative, and thus to show that $I_x$ is a rate function, we must prove that it has compact level sets. Traditionally, the proof of the compactness of the level sets has been treated as a purely analytic issue divorced from the large deviation analysis. However, as we show, the two questions are in fact closely related. Indeed, our proof of the compactness of the level sets follows from a deterministic analogue of the argument just given to prove the Laplace principle upper bound. The reader may wish to compare this relatively streamlined proof with other proofs that such functions have compact level sets (see, e.g., Section 9.1 of [59]).

**Proposition 6.2.4.** Under Condition 6.2.1, the function $I_x$ defined in equation (6.12) has compact level sets in $\mathcal{C}([0, 1] : \mathbb{R}^d)$.

**Remark 6.2.5.** For an application in Chapter 10, we note the properties of $L(x, \beta)$ that are used in the proof of the proposition: the uniform superlinearity of $L(x, \beta)$, the nonnegativity and lower semicontinuity of $L(x, \beta)$ for $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$, and the convexity of $L(\cdot, \beta)$ for each $x \in \mathbb{R}^d$. These properties, proved in Lemma 6.2.3, are consequences of the convexity of $H(\cdot, \beta)$ for each $x \in \mathbb{R}^d$ and the fact that $H(x, \alpha)$ is a continuous function of $(x, \alpha) \in \mathbb{R}^d \times \mathbb{R}^d$. The latter two properties of $H(x, \alpha)$ are immediate consequences of Condition 6.2.1. \hfill \blacksquare

**Proof of Proposition 6.2.4.** We prove that for any $M < \infty$ the level set

$$Z(M) = \{ \varphi \in \mathcal{C}([0, 1] : \mathbb{R}^d) : I_x(\varphi) \leq M \}$$

is a compact subset of $\mathcal{C}([0, 1] : \mathbb{R}^d)$. Suppose that $\{ \varphi^n, n \in \mathbb{N} \}$ is any sequence in $Z(M)$. Thus for each $n \in \mathbb{N}$, $I_x(\varphi^n) \leq M$. We will first prove that there exists a function $\varphi \in \mathcal{C}([0, 1] : \mathbb{R}^d)$ such that for some subsequence of $n \in \mathbb{N}$

$$\lim_{n \to \infty} \sup_{t \in [0, 1]} \| \varphi^n(t) - \varphi(t) \| = 0. \quad (6.13)$$

This will show that $Z(M)$ is relatively compact. We will then prove that $I_x(\varphi) \leq M$. This will show that $Z(M)$ is also closed and thus compact.
6.2. PROOF OF UPPER BOUND

Since for each \( n \) \( I_x(\varphi^n) \leq M < \infty \), it follows that \( \varphi^n(0) = x \) and that \( \varphi^n \) is absolutely continuous and thus differentiable a.s. For Borel subsets \( A \) of \( \mathbb{R}^d \) and \( B \) of \([0, 1]\), we define probability measures \( \nu^n \) on \( \mathbb{R}^d \times [0, 1] \) by

\[
\nu^n(A \times B) = \int_B \nu^n(A|t) \, dt,
\]

where a.s. for \( t \in [0, 1] \) \( \nu^n(A|t) \triangleq \delta_{\varphi^n(t)}(A) \). Thus for each \( t \in [0, 1] \)

\[
\varphi^n(t) = x + \int_{\mathbb{R}^d \times [0, t]} y \, \nu^n(dy \times ds).
\]

Since

\[
\inf > M \geq \sup_{n \in \mathbb{N}} I_x(\varphi^n) = \sup_{n \in \mathbb{N}} \int_{\mathbb{R}^d \times [0, 1]} L(\varphi^n(t), y) \, \nu^n(dy \times dt),
\]

the uniform superlinearity and nonnegativity of \( L(x, \beta) \) [Lemma 6.2.3 (b)–(c)] imply that the sequence \( \{\nu^n, n \in \mathbb{N}\} \) is uniformly integrable; i.e.,

\[
\lim_{C \to \infty} \sup_{n \in \mathbb{N}} \int_{\{y \in \mathbb{R}^d : ||y|| > C\} \times [0, 1]} ||y|| \, \nu^n(dy \times dt) = 0.
\]

This in turn implies that \( \{\nu^n\} \) is tight. Thus there exists a weakly convergent subsequence and a probability measure \( \nu \) on \( \mathbb{R}^d \times [0, 1] \) such that \( \nu^n \Rightarrow \nu \). Since the second marginal of each measure \( \nu^n(dy \times dt) \) is Lebesgue measure on \([0, 1]\), the second marginal of \( \nu(dy \times dt) \) is also Lebesgue measure on \([0, 1]\). By Theorem A.5.4 there exists a stochastic kernel \( \nu(dy|t) \) on \( \mathbb{R}^d \) given \([0, 1]\) such that

\[
\nu(A \times B) = \int_B \nu(A|t) \, dt
\]

for all Borel subsets \( A \) of \( \mathbb{R}^d \) and \( B \) of \([0, 1]\). This display is summarized as \( \nu(dy \times dt) = \nu(dy|t) \otimes dt \).

Since \( \int_{\mathbb{R}^d \times [0, t]} ||y|| \, \nu(dy \times dt) < \infty \), we can define a function \( \varphi \in \mathcal{C}([0, 1] : \mathbb{R}^d) \) by

\[
\varphi(t) = x + \int_{\mathbb{R}^d \times [0, t]} y \, \nu(dy \times ds) = x + \int_0^t \left( \int_{\mathbb{R}^d} y \, \nu(dy|s) \right) \, ds.
\]

Then \( \varphi \) is absolutely continuous, and a.s. for \( t \in [0, 1] \)

\[
\hat{\varphi}(t) = \int_{\mathbb{R}^d} y \, \nu(dy|t).
\]

(6.14)

Let \( g \) be any bounded continuous function mapping \( \mathbb{R}^d \) into \( \mathbb{R} \). Since \( \nu^n \Rightarrow \nu \) and \( \nu \{\mathbb{R}^d \times \{t\}\} = 0 \) for all \( t \in [0, 1] \), Theorem A.3.11 implies that for each \( t \in [0, 1] \)

\[
\lim_{n \to \infty} \int_{\mathbb{R}^d \times [0, t]} g(y) \, \nu^n(dy \times ds) = \int_{\mathbb{R}^d \times [0, t]} g(y) \, \nu(dy \times ds).
\]

The uniform integrability of \( \{\nu^n\} \) allows us to replace \( g(y) \) in the last display by \( y \). It follows that for all \( t \in [0, 1] \)

\[
\lim_{n \to \infty} \varphi^n(t) = \varphi(t).
\]
Consider the sequence \( \{ \varphi^n \} \) indexed by \( n \in \mathbb{N} \) for which \( \nu^n \rightarrow \nu \). In order to prove the uniform limit (6.13), it suffices to show that \( \{ \varphi^n \} \) is relatively compact in \( C([0,1] : \mathbb{R}^d) \). Let \( \varepsilon > 0 \) be given. By the uniform integrability of \( \{ \nu^n \} \), there exists \( C \in (0, \infty) \) such that
\[
\sup_{n \in \mathbb{N}} \int_{\{y \in \mathbb{R}^d : \|y\| > C\} \times [0,1]} \|y\| \nu^n(dy \times ds) \leq \varepsilon/2.
\]
Then for any numbers \( t \) and \( u \) in \([0,1]\) satisfying \( C|u-t| \leq \varepsilon/2 \)
\[
\|\varphi^n(u) - \varphi^n(t)\| \leq \int_{\{y \in \mathbb{R}^d : \|y\| \leq C\} \times [t \wedge u, t \vee u]} \|y\| \nu^n(dy \times ds) + \int_{\{y \in \mathbb{R}^d : \|y\| > C\} \times [0,1]} \|y\| \nu^n(dy \times ds)
\]
\[
\leq C|u-t| + \varepsilon/2 \leq \varepsilon.
\]
Thus the sequence \( \{ \varphi^n \} \) is uniformly equicontinuous. Since \( \varphi^n(0) = x \) for all \( n \in \mathbb{N} \), the Arzelà-Ascoli Theorem [Theorem A.6.2] yields the relative compactness of \( \{ \varphi^n \} \). The uniform limit (6.13) follows. This shows that the level set \( Z(M) \) is relatively compact.

In order to complete the proof that \( Z(M) \) is compact, we must show that \( \varphi \) lies in \( Z(M) \); i.e., that \( I_x(\varphi) \leq M \). This is a consequence of the following string of inequalities:
\[
M \geq \liminf_{n \to \infty} I_x(\varphi^n) = \liminf_{n \to \infty} \int_{\mathbb{R}^d \times [0,1]} L(\varphi^n(t), y) \nu^n(dy \times dt)
\]
\[
\geq \int_{\mathbb{R}^d \times [0,1]} L(\varphi(t), y) \nu(dy \times dt) = \int_0^1 \left( \int_{\mathbb{R}^d} L(\varphi(t), y) \nu(dy|t) \right) dt
\]
\[
\geq \int_0^1 L(\varphi(t), \int_{\mathbb{R}^d} y \nu(dy|t)) dt = \int_0^1 L(\varphi(t), \dot{\varphi}(t)) dt = I_x(\varphi).
\]
The first equality is implied by the definition of \( \nu^n \). The second inequality is a consequence of the nonnegativity and lower semicontinuity of \( L(\cdot, \cdot) \) [Lemma 6.2.3 (b)], the uniform convergence \( \varphi^n \rightarrow \varphi \), the weak convergence \( \nu^n \rightharpoonup \nu \), and part (b) of Theorem A.3.13. The next equality follows from the representation \( \nu(dy \times dt) = \nu(dy|t) \otimes dt \). The next inequality is a consequence of the convexity of \( L(x, \cdot) \) for each \( x \in \mathbb{R}^d \) and Jensen’s Inequality. The final two equalities follow from equation (6.14) and the definition of \( I_x(\varphi) \). This completes the proof that \( I_x \) has compact level sets. ■

### 6.3 Statement of the Laplace Principle

We introduce two separate conditions either of which, when coupled with Condition 6.2.1, will yield the Laplace principle lower bound. After this, we state the Laplace principle for the processes \( \{X^n, n \in \mathbb{N}\} \).

Ideally, one would prefer conditions that are placed directly either on \( \mu(dy|x) \) or \( H(x, \alpha) \) and that are simple to verify. This is the case for Condition 6.3.1. Since Condition 6.3.2 is phrased in terms of the Legendre–Fenchel transform \( L(x, \beta) \), we present immediately after the statement of the Laplace principle a set of techniques for verifying when Condition 6.3.2 holds.
6.3. STATEMENT OF THE LAPLACE PRINCIPLE

The following condition is satisfied if for each \( x \in \mathbb{R}^d \) the support of the measure \( \mu(dy|x) \) is all of \( \mathbb{R}^d \). We recall that for \( x \in \mathbb{R}^d \) \( \text{ri}(\text{conv } S_{\mu(|x|)}) \) denotes the relative interior of the convex hull of the support of \( \mu(\cdot|x) \). Part (b) is a technical hypothesis that is needed in the proof of Lemma 6.5.3.

**Condition 6.3.1.**

(a) The sets \( \text{ri}(\text{conv } S_{\mu(|x|)}) \) are independent of \( x \in \mathbb{R}^d \).

(b) \( 0 \in \Sigma = \text{ri}(\text{conv } S_{\mu(|x|)}) \).

The next condition, Condition 6.3.2, will allow us to treat cases in which the sets \( \text{ri}(\text{conv } S_{\mu(|x|)}) \) depend on \( x \in \mathbb{R}^d \) and so Condition 6.3.1 does not hold. In Chapter 10 the Laplace principle for continuous-time diffusion processes and jump diffusion processes will be proved under conditions that are analogues of Conditions 6.2.1 and 6.3.1 or Conditions 6.2.1 and 6.3.2. In the continuous-time context, Condition 6.3.2 is particularly important since it allows the treatment of degenerate diffusions and degenerate jump diffusions. In order to state this condition, we must assume that for each \( x \) and \( \alpha \) in \( \mathbb{R}^d \)

\[
\int_{\mathbb{R}^d} \exp(\alpha, y) \mu(dy|x) < \infty. \tag{6.15}
\]

This guarantees the existence of the functions \( H(x, \alpha) \) and \( L(x, \beta) \) defined in equations (6.3) and (6.4), respectively. Of course (6.15) is valid if part (a) of Condition 6.2.1 holds. Proposition 6.3.4 gives a large collection of stochastic kernels for which the following condition is satisfied.

**Condition 6.3.2.** Let any compact subset \( \Delta \) of \( \mathbb{R}^d \) and any number \( \varepsilon \in (0, 1) \) be given. Then there are numbers \( \eta = \eta(\Delta, \varepsilon) \in (0, 1) \) and \( K = K(\Delta, \varepsilon) \in (0, \infty) \) so that whenever \( \xi \in \Delta, v \in \Delta \), and \( \gamma \in \mathbb{R}^d \) satisfy \( \|\xi - v\| \leq \eta \), there is a \( \tilde{\beta} \in \mathbb{R}^d \) such that

\[
L(\xi, \tilde{\beta}) - L(v, \gamma) \leq \varepsilon [1 + L(v, \gamma)] \tag{6.16}
\]

and

\[
\|\tilde{\beta} - \gamma\| \leq K \|\xi - v\| [1 + L(v, \gamma)]. \tag{6.17}
\]

We next state the Laplace principle. We emphasize that when Condition 6.3.1 holds, the only continuity assumption required for the Laplace principle is the continuity of the function mapping \( x \mapsto \mu(\cdot|x) \) as expressed in Condition 6.2.1. When Condition 6.3.1 is not satisfied, then we also need the Lipschitz-type smoothness condition expressed in Condition 6.3.2. In Example 6.3.5 following the statement of the theorem, we give an example in which both Conditions 6.3.1 and 6.3.2 fail and the Laplace principle with rate function \( I_x \) is not valid.

**Theorem 6.3.3.** We assume Condition 6.2.1 and either Condition 6.3.1 or Condition 6.3.2. For any \( x \in \mathbb{R}^d \), consider the piecewise-linearly interpolated processes \( \{X^n, n \in \mathbb{N}\} \) defined in equations (6.1) and (6.2) and satisfying \( X^n(0) = x \). For absolutely continuous functions \( \varphi \in C([0, 1] : \mathbb{R}^d) \) satisfying \( \varphi(0) = x \), we define

\[
I_x(\varphi) \triangleq \int_0^1 L(\varphi(t), \dot{\varphi}(t)) \, dt,
\]
where \( L \) is defined in equation (6.4). For all other functions \( \varphi \in C([0,1] : \mathbb{R}^d) \), set \( I_x(\varphi) = \infty \). Then the sequence \( \{X^n\} \) satisfies the Laplace principle on \( C([0,1] : \mathbb{R}^d) \) with rate function \( I_x \). In fact, the Laplace principle holds uniformly on compacts.

In order to simplify the exposition, we will focus on the proof of the nonuniform Laplace principle. The proof of the uniform version uses the same arguments but with additional notation. Let \( h \) be any bounded continuous function mapping \( C([0,1] : \mathbb{R}^d) \) into \( \mathbb{R} \). Then according to Proposition 1.2.7, it suffices to show that the rate function \( I_x \) has compact level sets uniformly on compacts and that whenever \( \{x_n, n \in \mathbb{N}\} \) is a sequence in \( \mathbb{R}^d \) converging to \( x \in \mathbb{R}^d \)

\[
\lim_{n \to \infty} \frac{1}{n} \log E_{x_n} \{\exp[-n h(X^n)]\} = -\inf_{x \in \mathbb{R}^d} \{I_x(\varphi) + h(\varphi)\}.
\]

With the replacement of \( x \) by \( x_n \) in the appropriate places, the reader can easily verify that the proofs of Proposition 5.3.2 and Theorem 5.3.5 go through without substantive change and that the limit in the last display is valid. Analogous modifications of the proof of Proposition 6.2.4 yield that \( I_x \) has compact level sets on compacts. Because these alterations are truly elementary, we have provided details in only one case; namely, in the proof of Theorem 8.4.3.

One of the hypotheses in the theorem is Condition 6.2.1, which states that for each \( \alpha \in \mathbb{R}^d \), \( \sup_{x \in \mathbb{R}^d} H(x, \alpha) < \infty \). In Remark 6.3.6, which appears at the end of this section, we comment on how this boundedness condition may be weakened.

The Laplace principle upper bound and the fact that \( I_x \) has compact level sets have already been proved in Proposition 6.2.2 and Proposition 6.2.4. The lower bound will be proved in the next two sections under Conditions 6.2.1 and 6.3.1 and under Conditions 6.2.1 and 6.3.2, respectively.

Before turning to the proof of the Laplace principle lower bound, we first discuss Conditions 6.3.1 and 6.3.2. Condition 6.3.1 includes the hypothesis that the sets \( \text{ri}(\text{conv} \ S_{\mu(x)}) \) are independent of \( x \in \mathbb{R}^d \). This condition is implicit in the assumptions made in the works [71, 89, 90, 91, 92]. In addition, these works assume a number of other technical conditions on \( H \) and \( L \) that are difficult to verify. The conditions that we assume are much simpler. As we have pointed out, it is also desirable to treat problems where the sets \( \text{ri}(\text{conv} \ S_{\mu(x)}) \) depend on \( x \). This is allowed by Condition 6.3.2.

It is only when \( \mu(dy|x) \) takes a simple form such as Gaussian or Bernoulli that one can check by explicit calculation whether or not Conditions 6.2.1 and 6.3.2 hold. The next proposition allows one to go considerably beyond these special cases. Part (a) gives a large class of stochastic kernels for which the conditions hold. Parts (b) and (c) show how one can construct, from stochastic kernels for which the conditions hold, more complex stochastic kernels for which they hold. After the proposition is stated, two illustrative examples are given. The proposition is not exhaustive because it omits other extensions which, for example, allow the weights \( \rho_j \) in part (b) to depend on the spatial parameter \( x \). The proposition is proved in Section C.7.

**Proposition 6.3.4.**
6.3. STATEMENT OF THE LAPLACE PRINCIPLE

(a) Dilation, translation, and rotation. Let \( \mu \in \mathcal{P}(\mathbb{R}^d) \) satisfy
\[
\int_{\mathbb{R}^d} \exp(\alpha, y) \mu(dy) < \infty \quad \text{for all } \alpha \in \mathbb{R}^d
\]
and define
\[
H(\alpha) \doteq \log \int_{\mathbb{R}^d} \exp(\alpha, y) \mu(dy).
\]
Assume that \( \Sigma \) is a bounded, Lipschitz continuous function mapping \( \mathbb{R}^d \) into the space of real \( d \times d \) matrices and that for each \( x \in \mathbb{R}^d \) \( \Sigma(x) \) is invertible; we denote by \( \Sigma(x)^T \) the transpose of \( \Sigma(x) \). Also assume that \( b \) is a bounded, Lipschitz continuous function mapping \( \mathbb{R}^d \) into \( \mathbb{R}^d \). For \( x \in \mathbb{R}^d \) and Borel subsets \( B \) of \( \mathbb{R}^d \), we define a stochastic kernel on \( \mathbb{R}^d \) given \( \mathbb{R}^d \) by
\[
\mu(B|x) \doteq \mu \left( \left( \Sigma(x)^T \right)^{-1} [B - b(x)] \right). \tag{6.18}
\]
Then for \( \mu(dy|x) \) Conditions 6.2.1 and 6.3.2 hold. In this case
\[
H(x, \alpha) = \langle \alpha, b(x) \rangle + H(\Sigma(x)\alpha).
\]

(b) Sums. Suppose that for some \( N \in \mathbb{N} \) and each \( j \in \{1, 2, \ldots, N\} \) \( \mu_j(dy|x) \) is a stochastic kernel on \( \mathbb{R}^d \) given \( \mathbb{R}^d \) for which Conditions 6.2.1 and 6.3.2 hold. Given real numbers \( \{\rho_j, j = 1, 2, \ldots, N\} \) satisfying \( \rho_j > 0 \) and \( \sum_{j=1}^N \rho_j = 1 \), we define
\[
\mu(dy|x) \doteq \sum_{j=1}^N \rho_j \mu_j(dy|x).
\]
Then for \( \mu(dy|x) \) Conditions 6.2.1 and 6.3.2 also hold. In this case
\[
H(x, \alpha) = \log \left( \sum_{j=1}^N \rho_j \exp[H_j(x, \alpha)] \right),
\]
where
\[
H_j(x, \alpha) \doteq \log \int_{\mathbb{R}^d} \exp(\alpha, y) \mu_j(dy|x).
\]

(c) Convolutions. Suppose that \( \mu_1(dy|x) \) and \( \mu_2(dy|x) \) are stochastic kernels on \( \mathbb{R}^d \) given \( \mathbb{R}^d \) for which Conditions 6.2.1 and 6.3.2 hold. For \( x \in \mathbb{R}^d \) and Borel subsets \( B \) of \( \mathbb{R}^d \) we define the convolution \( (\mu_1 * \mu_2)(\cdot|x) \) by
\[
(\mu_1 * \mu_2)(B|x) \doteq \int_{\mathbb{R}^d} \mu_1(B - z|x) \mu_2(dz|x).
\]
Then for \( \mu(\cdot|x) \doteq (\mu_1 * \mu_2)(\cdot|x) \) Conditions 6.2.1 and 6.3.2 also hold. In this case
\[
H(x, \alpha) = H_1(x, \alpha) + H_2(x, \alpha).
\]
We point out two elementary examples of stochastic kernels that are covered by this proposition. The measure

\[ \mu(B) = \frac{1}{(2\pi)^{d/2}} \int_B \exp\left(-\frac{\|y\|^2}{2}\right) \, dy \]

satisfies the integrability condition in part (a) of the proposition, and if we take \( \Sigma \) and \( b \) as in part (a), then \( \mu(dy|\xi) \) is a normal distribution with mean \( b(\xi) \) and covariance matrix \( \Sigma(\xi)\Sigma(\xi)^T \). For another example, choose bounded, Lipschitz continuous functions \( \{b_j; j = 1, 2, \ldots, N\} \) mapping \( \mathbb{R}^d \) into \( \mathbb{R}^d \) and real numbers \( \{\rho_j; j = 1, 2, \ldots, N\} \) as in part (b). From the choice \( \mu(dy) = \delta_0(dy) \) part (a) gives \( \mu_j(dy|\xi) = \delta_{b_j}(dy) \), and part (b) yields the sum \( \sum_{j=1}^N \rho_j \delta_{b_j}(dy) \).

If Condition 6.3.1 is not satisfied, then the Lipschitz–type smoothness condition expressed in Condition 6.3.2 is needed for the proof of the Laplace principle in Theorem 6.3.3. The following example shows that when Condition 6.3.1 is not satisfied, a condition of this kind is not surprising, at least if the Laplace principle is to hold with the usual rate function; i.e., the rate function defined in the statement of Theorem 6.3.3.

**Example 6.3.5.** We will give the example in terms of a diffusion process with small noise. An analogous problem that fits into the framework of the present chapter is easily constructed.

Let \( f_1 \) and \( f_2 \) be bounded, Lipschitz continuous functions mapping \( \mathbb{R} \) into \( \mathbb{R} \) and assume that \( f_2(x) = |x| \) for \( x \) in some neighborhood of \( 0 \). For \( n \in \mathbb{N} \) and \( t > 0 \) we consider the stochastic differential equation

\[
\begin{align*}
\frac{dY_1^n(t)}{dt} &= (f_1(Y_1^n(t)) + f_2(Y_2^n(t))) \, dt \\
\frac{dY_2^n(t)}{dt} &= n^{-1/2} \, dw(t),
\end{align*}
\]

(6.19)

where \( \{w(t), t \geq 0\} \) is a standard Wiener process on \( \mathbb{R} \) and \( (Y_1^n(0), Y_2^n(0)) = 0 \in \mathbb{R}^2 \). The process \( Y^n = \{(Y_1^n(t), Y_2^n(t)), t \geq 0\} \) is a degenerate diffusion process on \( \mathbb{R}^2 \). The results in Chapter 10 imply that the sequence \( \{Y^n, n \in \mathbb{N}\} \) satisfies the Laplace principle on \( C([0,1] : \mathbb{R}^2) \). The latter is proved under analogues of Conditions 6.2.1, 6.3.1, and 6.3.2. In order to simplify the present discussion we will refer to the latter conditions, with the understanding that the Chapter 10 analogues are actually intended. For \( x, \alpha, \) and \( \beta \) in \( \mathbb{R}^2 \) we define

\[ H(x, \alpha) \triangleq \alpha_1 (f_1(x_1) + f_2(x_2)) + \frac{1}{2} \alpha_2^2 \]

and

\[ L(x, \beta) \triangleq \sup_{\alpha \in \mathbb{R}^2} \{\langle \alpha, \beta \rangle - H(x, \alpha)\} = \begin{cases} \frac{1}{2} \beta_2^2 & \text{if } \beta_1 = f_1(x_1) + f_2(x_2), \beta_2 \in \mathbb{R}, \\ \infty & \text{otherwise}. \end{cases} \]

Finally, for absolutely continuous \( \varphi \in C([0,1] : \mathbb{R}^2) \) satisfying \( \varphi(0) = 0 \), we define

\[ I_0(\varphi) \triangleq \int_0^1 L(\varphi(t), \dot{\varphi}(t)) \, dt. \]

(6.20)
For all other \( \varphi \in \mathcal{C}([0,1]:\mathbb{R}^d) \), set \( I_0(\varphi) = \infty \). Although Condition 6.2.1 is satisfied, Condition 6.3.1 is not satisfied because the sets

\[
\text{ri}(\text{dom } L(x, \cdot)) = \{ (\beta_1, \beta_2) : \beta_1 = f_1(x_1) + f_2(x_2), \beta_2 \in \mathbb{R} \}
\]

depend on \( x \) and in general do not contain 0. However, because \( f_1 \) and \( f_2 \) are bounded and Lipschitz continuous, Condition 6.3.2 is valid. Theorem 10.2.6 implies that the sequence \( \{ Y^n, n \in \mathbb{N} \} \) satisfies the Laplace principle on \( \mathcal{C}([0,1]:\mathbb{R}^2) \) with rate function \( I_0 \).

In (6.19) we now choose \( f_1(x_1) = |x_1|^{1/2} \wedge 1 \). It is not hard to show that for each \( n \in \mathbb{N} \) this stochastic differential equation has a strong solution \( Y^n \) that is unique in the strong sense [61]. Furthermore, because the Wiener process \( \{ w(t), t \geq 0 \} \) is not identically zero on any interval \([0, \delta]\), \( \delta > 0 \), and \( f_2(x_2) = |x_2| \) for \( x_2 \) in some neighborhood of 0, there exists \( C > 0 \) such that w.p.1 \( x^n(t) \geq t^2/4 \) for all \( t \in [0,C] \). With the choice \( f_1(x_1) = |x_1|^{1/2} \wedge 1 \), Condition 6.2.1 continues to be satisfied. However, Condition 6.3.1 is not satisfied and, since the function mapping \( x_1 \mapsto |x_1|^{1/2} \wedge 1 \) is not Lipschitz continuous, \( L(x, \beta) \) does not satisfy Condition 6.3.2. Hence Theorem 10.2.6 cannot be applied to this diffusion process.

Nevertheless, let us assume that with the choice \( f_1(x_1) = |x_1|^{1/2} \wedge 1 \), the function \( I_0 \) defined in equation (6.20) were actually the rate function for \( \{ Y^n, n \in \mathbb{N} \} \). For the function \( \varphi(t) \equiv 0 \) for \( t \in [0,1] \), we have \( I_0(\varphi) = 0 \). If the Laplace principle lower bound, and thus the large deviation lower bound, were valid, then for any \( \delta > 0 \)

\[
\liminf_{n \to \infty} \frac{1}{n} \log P_0\{ d(Y^n, \varphi) < \delta \} \geq 0.
\]

However, the fact that w.p.1 \( Y^n(t) \geq t^2/4 \) for all \( t \in [0,C] \) implies that

\[
P_0\{ d(Y^n, \varphi) < \delta \} = 0
\]

for sufficiently small \( \delta > 0 \). This is a contradiction. We conclude that with the choice of the non-Lipschitz continuous function \( f_1(x_1) = |x_1|^{1/2} \wedge 1 \), for which Conditions 6.3.1 and 6.3.2 both fail, the Laplace principle with rate function \( I_0 \) does not hold. It would seem that the difficulty is due in part to the nonuniqueness of solutions of the “noiseless” or “mean” dynamical system \( \dot{x}_1 = |x_1|^{1/2} \wedge 1 + f_2(x_2), \dot{x}_2 = 0 \), and in part to the violation of Condition 6.3.1. This completes the example.  

One of the hypotheses of the Laplace principle stated in Theorem 6.3.3 is the global boundedness condition that for each \( \alpha \in \mathbb{R}^d \) \( \sup_{x \in \mathbb{R}^d} H(x, \alpha) < \infty \) [Condition 6.2.1 (a)]. We end this section by discussing ways in which this hypothesis can be weakened. The most basic continuous-time example in which such a weakening would be useful is a small noise diffusion process with a linear drift.

**Remark 6.3.6.** (a) For the Laplace principle lower bound, one may replace the global boundedness condition in part (a) of Condition 6.2.1 by the following local boundedness condition: for each compact subset \( K \) of \( \mathbb{R}^d \) and each \( \alpha \in \mathbb{R}^d \)

\[
\sup_{x \in K} H(x, \alpha) < \infty.
\]
In order to prove the Laplace principle lower bound under this weaker condition, one modifies the stochastic kernel \( \mu(dy|x) \) for \( x \) outside a sufficiently large compact set \( K \) and uses the fact that the random walk model defined by the original stochastic kernel \( \mu(dy|x) \) agrees in law with the random walk model defined by the modified stochastic kernel until the first exit time from \( K \).

(b) In Condition 5.4.1 we stated a strictly weaker version of the global boundedness condition in part (a) of Condition 6.2.1. This weaker version involves a locally Lipschitz continuous function \( b \) mapping \( \mathbb{R}^d \) into \( \mathbb{R}^d \) and the shifted cumulant generating function

\[
H_b(x, \alpha) = \log \int_{\mathbb{R}^d} \exp(\alpha, y - b(x)) \mu(dy|x).
\]

When \( b \equiv 0 \), \( H_b(x, \alpha) \) reduces to the cumulant generating function \( H(x, \alpha) \) and Condition 5.4.1 reduces to part (a) of Condition 6.2.1. We now point out that when Condition 5.4.1 replaces part (a) of Condition 6.2.1, the Laplace principle stated in Theorem 6.3.3 also remains true.

Under Condition 6.2.1, the Laplace principle upper bound was proved in Section 6.2 by using Theorem 5.3.5. In Theorem E.2.3 we prove that when Condition 5.4.1 replaces part (a) of Condition 6.2.1, all the conclusions of Theorem 5.3.5 remain true. Hence under this replacement of conditions, one may prove the Laplace principle upper bound exactly as in Section 6.2 by applying Theorem E.2.3 instead of Theorem 5.3.5.

As pointed out in part (a) of the present remark, in order to prove the Laplace principle lower bound in Theorem 6.3.3, it suffices that for each compact subset \( K \) of \( \mathbb{R}^d \) and each \( \alpha \in \mathbb{R}^d \)

\[
\sup_{x \in K} H(x, \alpha) < \infty. \tag{6.21}
\]

Since for each \( x \) and \( \alpha \) in \( \mathbb{R}^d \)

\[
H(x, \alpha) = H_b(x, \alpha) + b(x),
\]

(6.21) is a consequence of Condition 5.4.1. The proof of the Laplace principle lower bound also requires the conclusions of Theorem 5.3.5, which, as noted above, remain true when Condition 5.4.1 replaces part (a) of Condition 6.2.1. Hence under this replacement of conditions, the Laplace principle lower bound in Theorem 6.3.3 remains valid.

The proof that the function \( I_x \) has compact level sets in \( C([0, 1] : \mathbb{R}^d) \) uses a shifting technique analogous to that used to prove the compactness of appropriate sequences of controls in Proposition E.2.2. We omit the details. This completes the remark. ■

### 6.4 Strategy for the Proof of the Laplace Principle Lower Bound

Throughout this section, we assume Condition 6.2.1. Fix \( x \in \mathbb{R}^d \) and let \( h \) be any bounded continuous function mapping \( C([0, 1] : \mathbb{R}^d) \). In order to prove the Laplace principle lower bound, we must prove that the functions

\[
W^n(x) \doteq -\frac{1}{n} \log E_x \{\exp[- n h(X^n)]\}
\]
6.4. STRATEGY FOR PROOF OF LOWER BOUND

satisfy the upper limit

\[ \limsup_{n \to \infty} W^n(x) \leq \inf_{\varphi \in C([0,1]; \mathbb{R}^d)} \{ I_x(\varphi) + h(\varphi) \}. \]  \hspace{1cm} (6.22)

For a given \( \varepsilon > 0 \) there exists \( \psi \in C([0,1]; \mathbb{R}^d) \) satisfying

\[ I_x(\psi) + h(\psi) \leq \inf_{\varphi \in C([0,1]; \mathbb{R}^d)} \{ I_x(\varphi) + h(\varphi) \} + \varepsilon < \infty, \]

and since \( h \) is bounded, this implies that \( I_x(\psi) < \infty \). Hence \( \psi(0) = x \). Because \( \varepsilon > 0 \) is arbitrary, the Laplace principle lower bound follows once we derive the upper limit

\[ \limsup_{n \to \infty} W^n(x) \leq I_x(\psi) + h(\psi). \]  \hspace{1cm} (6.23)

We will carry this out by using the stochastic control representation formula given in Theorem 4.3.1; namely,

\[ W^n(x) = V^n(x) + \inf_{\{\nu^n_j\}} \mathbb{E}^x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot)\|\mu(\cdot | \tilde{X}_j^n)) + h(\tilde{X}_n) \right\}. \]  \hspace{1cm} (6.24)

The infimum is taken over all admissible control sequences \( \{\nu^n_j\} \), and as the first argument of the relative entropy, \( \nu^n_j(\cdot) = \nu^n_j(\cdot | \tilde{X}_j^n, \tilde{X}_1^n, \ldots, \tilde{X}_0^n) \). The process \( \tilde{X} \) is defined in formula (5.6) as the piecewise linear interpolation of the controlled random vectors \( \{\tilde{X}_j^n\} \).

The basic idea will be to utilize \( \psi \) to design admissible control sequences \( \{\nu^n_j\} \) and controlled random vectors \( \{\tilde{X}_j^n\} \) that can be used to prove the upper limit (6.23). A problem that will routinely occur in the proofs of this limit is that \( \psi \) may not be sufficiently well-behaved and must be approximated. The sense of the term “well-behaved” will of course vary with the particular situation. We begin our discussion by proving (6.23) under the temporary assumption that the derivative \( \dot{\psi}(t) \) is piecewise continuous with only finitely many jumps in \((0,1)\). In order to have \( \dot{\psi}(t) \) defined for all \( t \in [0,1) \), we replace the almost everywhere defined function \( \psi \) by its right continuous regularization. This convention is in force for the remainder of the chapter. An additional simplification is also achieved by temporarily assuming that \( L(x, \beta) \) is a continuous function on all of \( \mathbb{R}^d \times \mathbb{R}^d \). Both of these assumptions will be dropped in the following two sections. In general, \( L(x, \beta) \) is not continuous on all of \( \mathbb{R}^d \times \mathbb{R}^d \) and is not even finite on all of \( \mathbb{R}^d \times \mathbb{R}^d \). These issues, as well as a continuity condition that is valid for a general \( L \), are elucidated by Lemma 6.5.2 in the next section. In the proof of the Laplace principle lower bound to be given in the next few sections, a number of technical difficulties arise because of the state dependence of the stochastic kernel \( \mu(dy|x) \). This state dependence and the associated technical difficulties were absent from the proof of Mogulskii’s Theorem in Chapter 3.

Procedure 6.4.1: Construction of the Admissible Control Sequences and the Controlled Processes. We will explain the construction in detail since similar constructions will be needed later. For each \( n \in \mathbb{N} \) we define first the admissible control sequence \( \{\nu^n_j\} \) and then the controlled random vectors \( \{\tilde{X}_j^n\} \). For \( j \in \{0,1, \ldots, n-1\} \), points in
\((\mathbb{R}^d)^{j+1}\) are denoted by \(\xi = (\xi_0, \xi_1, \ldots, \xi_j)\), where each coordinate \(\xi_i\) is an element of \(\mathbb{R}^d\). The admissible control \(\nu_j^n\) applied at time \(j\) is a stochastic kernel on \(\mathbb{R}^d\) given \((\mathbb{R}^d)^{j+1}\) which we denote by \(\nu_j^n(dy|\xi)\). Part (g) of Lemma 6.2.3 guarantees that there exists a stochastic kernel \(\gamma_j^n(dy|\xi_j)\) on \(\mathbb{R}^d\) given \(\mathbb{R}^d\) satisfying for each \(\xi_j \in \mathbb{R}^d\)
\[
R(\gamma_j^n(\cdot|\xi_j)\|\mu(\cdot|\xi_j)) = L(\xi_j, \dot{\psi}(j/n)) \quad \text{and} \quad \int_{\mathbb{R}^d} y \gamma_j^n(dy|\xi_j) = \dot{\psi}(j/n).
\]

Now fix \(\delta > 0\); the definition of \(\nu_j^n(dy|\xi)\) depends on whether \(\max_{i=0,1,\ldots,j} \|\xi_i - \psi(i/n)\| \leq \delta\) or \(\max_{i=0,1,\ldots,j} \|\xi_i - \psi(i/n)\| > \delta\). Specifically, for \(\xi \in (\mathbb{R}^d)^{j+1}\) we define
\[
\nu_j^n(dy|\xi) = \begin{cases} \gamma_j^n(dy|\xi_j) & \text{if } \max_{i=0,1,\ldots,j} \|\xi_i - \psi(i/n)\| \leq \delta, \\ \mu(dy|\xi_j) & \text{if } \max_{i=0,1,\ldots,j} \|\xi_i - \psi(i/n)\| > \delta. \end{cases}
\]

Having defined the admissible control sequence \(\{\nu_j^n\}\), we set \(\bar{X}_0^n = x\), where \(x\) is a fixed point in \(\mathbb{R}^d\), and take \(\bar{Y}_j^n\) to be a random vector with conditional distribution
\[
P_x\{\bar{Y}_j^n \in dy|\bar{X}_0^n, \bar{X}_1^n, \ldots, \bar{X}_j^n\} = \nu_j^n(dy|\bar{X}_0^n, \bar{X}_1^n, \ldots, \bar{X}_j^n).
\]
We then define
\[
\bar{X}_{j+1}^n = \bar{X}_j^n + \frac{1}{n} \bar{Y}_j^n.
\]
In terms of the stopping time
\[
\tau^n \triangleq \frac{1}{n} \left( \min \left\{ i \in \{0,1,\ldots,n\} : \|\bar{X}_i^n - \psi(i/n)\| > \delta \right\} \right) \wedge n,
\]
the admissible controls \(\nu_j^n(\cdot) = \nu_j^n(\cdot|\bar{X}_0^n, \bar{X}_1^n, \ldots, \bar{X}_j^n)\) have the following properties. For \(j \in \{0,1,\ldots,n\tau^n - 1\}\)
\[
R(\nu_j^n(\cdot)\|\mu(\cdot|\bar{X}_j^n)) = L(\bar{X}_j^n, \dot{\psi}(j/n)) \quad \text{and} \quad \int_{\mathbb{R}^d} y \nu_j^n(dy) = \dot{\psi}(j/n). \quad (6.25)
\]
For \(j \in \{n\tau^n, n\tau^n + 1, \ldots, n - 1\}\) \(\nu_j^n(\cdot)\) equals \(\mu(\cdot|\bar{X}_j^n)\) and thus \(R(\nu_j^n(\cdot)\|\mu(\cdot|\bar{X}_j^n))\) equals 0. This completes the construction of the admissible control sequences and the controlled random vectors. ■

We now proceed with the proof of the upper limit (6.23). For each \(n \in \mathbb{N}\), associated with the admissible control sequence \(\{\nu_j^n\}\) just constructed are the admissible control measure \(\nu^n(dy \times dt) = \nu^n(dy|t) \otimes dt\), defined in equations (5.3) and (5.4), and the running cost
\[
\mathbb{E}_{\bar{x}} \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu_j^n(\cdot)\|\mu(\cdot|\bar{X}_j^n)) \right\}
\]

\[
= \mathbb{E}_{\bar{x}} \left\{ \frac{1}{n} \sum_{j=0}^{n\tau^n-1} R(\nu_j^n(\cdot)\|\mu(\cdot|\bar{X}_j^n)) + \frac{1}{n} \sum_{j=n\tau^n}^{n-1} R(\mu(\cdot|\bar{X}_j^n)\|\mu(\cdot|\bar{X}_j^n)) \right\}
= \mathbb{E}_{\bar{x}} \left\{ \frac{1}{n} \sum_{j=0}^{n\tau^n-1} L(\bar{X}_j^n, \dot{\psi}(j/n)) \right\}.
\]
6.4. STRATEGY FOR PROOF OF LOWER BOUND

We recall the temporary assumptions on $\psi$ and $L$. The boundedness of the sets $\{\|\psi(t)\|, t \in [0,1]\}$ and $\{\|\psi(t)\|, t \in [0,1]\}$, the definition of the stopping times $\tau^n$, and the continuity of $L(x, \beta)$ on all of $\mathbb{R}^d \times \mathbb{R}^d$ imply that the quantities $L(X^n_j, \psi(j/n))$ are uniformly bounded for $n \in \mathbb{N}$ and $j \in \{0, 1, \ldots, n\tau^n - 1\}$. It follows that the sequence of running costs in the last display is bounded. Since part (a) of Condition 6.2.1 coincides with Condition 5.3.1, we have verified the hypotheses of Theorem 5.3.5 and Proposition 5.3.8.

We now apply the compactness and convergence results stated there.

For $n \in \mathbb{N}$ we define the process $S^n = \{S^n(t), t \in [0,1]\}$ by

$$S^n(t) \equiv x + \int_{\mathbb{R}^d \times [0,t]} y \nu^n(dy \times ds).$$

Since $\tau^n$ takes values in the compact set $[0,1]$, Theorem 5.3.5, Proposition 5.3.8, and Prohorov's Theorem imply that any subsequence of the sequence $\{(\nu^n, X^n, S^n, \tau^n), n \in \mathbb{N}\}$ is tight and that given any subsequence, there exists a subsequence and a random variable $\tau$ taking values in $[0,1]$ such that

$$\mathbf{(\nu^n, X^n, S^n, \tau^n)} \xrightarrow{\mathcal{D}} (\nu, \bar{X}, \bar{X}, \tau)$$

as $n \to \infty$. With probability 1, for every $t \in [0,1]$ the limiting process $\bar{X} = \{\bar{X}(t), t \in [0,1]\}$ is related to the limiting stochastic kernel $\nu(dy \times dt) = \nu(dy(t) \otimes dt) by$

$$\bar{X}(t) - x = \int_{\mathbb{R}^d \times [0,t]} y \nu(dy \times ds) = \int_0^t \left( \int_{\mathbb{R}^d} y \nu(dy|s) \right) ds.$$

By using the Skorohod Representation Theorem, we can assume that w.p.1

$$\mathbf{(\nu^n, X^n, S^n, \tau^n)} \longrightarrow (\nu, \bar{X}, \bar{X}, \tau).$$

The next lemma is needed in order to complete the proof.

**Lemma 6.4.2.** With probability 1, $\tau = 1$ and $\bar{X}(t) = \psi(t)$ for all $t \in [0,1]$.

**Proof.** The equality $\bar{X}^n(0) = x$ for each $n \in \mathbb{N}$ implies that w.p.1 $\bar{X}(0) = x = \psi(0)$. Since for any $t \in [0, \tau]$ and all sufficiently large $n \in \mathbb{N}$ there exists $j^* \in \{0, 1, \ldots, n\tau^n - 1\}$ such that $t \in [j^*/n, (j^* + 1)/n)$, the definition of the admissible control sequence yields

$$x + \int_{\mathbb{R}^d \times [0,t]} y \nu^n(dy \times ds) = x + \frac{1}{n} \sum_{j=0}^{j^* - 1} \int_{\mathbb{R}^d} y \nu^n_j(dy) + (t - j^*/n) \int_{\mathbb{R}^d} y \nu^n_{j^*}(dy)$$

$$= x + \frac{1}{n} \sum_{j=0}^{j^* - 1} \psi(j/n) + (t - j^*/n) \psi(j^*/n).$$

The last term in this display is of the order $O(1/n)$. Hence sending $n \to \infty$, w.p.1 we obtain for every $t \in [0, \tau)$

$$\bar{X}(t) = x + \int_{\mathbb{R}^d \times [0,t]} y \nu(dy \times ds) = x + \int_0^t \psi(s) ds = \psi(t),$$
which extends by continuity to \( t = \tau \). The definition of the stopping times \( \tau^n \), the probability–1 convergence of \( \{X^n\} \) to \( X \) uniformly on \([0, 1] \), and the continuity of \( X \) and \( \psi \) imply that on the event \( \{\tau < 1\}\)

\[
\lim_{t \uparrow \tau} \|\bar{X}(t) - \psi(t)\| \geq \delta.
\]

However, if \( P_x\{\tau < 1\} \) were positive, then the previous two displays would contradict the continuity of \( \psi(t) \) and the probability–1 continuity of \( \bar{X}(t) \) at \( t = \tau \). We conclude that w.p.1 \( \tau = 1 \) and \( \bar{X}(t) = \psi(t) \) for all \( t \in [0, 1] \). \( \blacksquare \)

The following display involving the subsequence along which the convergence \((\nu^n, \bar{X}^n, \tau^n) \to (\nu, \psi, 1)\) holds will lead immediately to the desired upper limit (6.23):

\[
\limsup_{n \to \infty} W^n(x) = \limsup_{n \to \infty} V^n(x) = \int_0^1 L(\psi(t), \psi(t)) \, dt + h(\psi) = I_x(\psi) + h(\psi).
\]

Lines one and two of this display are a consequence of formula (6.24), and line three follows from a property of the admissible controls \( \nu^n_j \) specified in equation (6.25). In order to prove line four, we use the uniform continuity of \( L(\cdot, \cdot) \) on compact subsets of \( \mathbb{R}^d \times \mathbb{R}^d \); the piecewise continuity of \( \psi \); the probability–1 convergence, uniform over \( t \in [0, 1] \), of \( \bar{X}^n(t) \) to \( \bar{X}(t) = \psi(t) \); the probability–1 convergences \( \tau^n \to 1 \) and \( h(\bar{X}^n) \to h(\psi) \); and the Lebesgue Dominated Convergence Theorem. Line five follows from the definition of \( I_x(\psi) \). We have shown that every subsequence of the original sequence \( \{W^n(x), n \in \mathbb{N}\} \) has a subsequence satisfying the upper limit

\[
\limsup_{n \to \infty} W^n(x) \leq I_x(\psi) + h(\psi).
\]

An argument by contradiction applied to an arbitrary subsequence then gives the same upper limit for the entire sequence \( \{W^n(x), n \in \mathbb{N}\} \). By our choice of \( \psi \) we have completed the proof of the upper limit (6.22) under the temporary assumptions on \( \psi \) and \( L \).

Our task is to turn the calculations just given into a proof of the Laplace principle lower bound. It suffices to prove the upper limit (6.23) for any point \( x \in \mathbb{R}^d \), any bounded continuous function \( h \) mapping \( C([0, 1] : \mathbb{R}^d) \) into \( \mathbb{R} \), and any \( \psi \in C([0, 1] : \mathbb{R}^d) \) satisfying \( I_x(\psi) < \infty \). We first address the issue of proving (6.23) for an arbitrary \( \psi \in C([0, 1] : \mathbb{R}^d) \) satisfying \( I_x(\psi) < \infty \), while retaining for a little while longer the assumption that \( L \) is continuous on all of \( \mathbb{R}^d \times \mathbb{R}^d \). The condition \( I_x(\psi) < \infty \) guarantees that \( \psi \) is absolutely
continuous and \( \psi(0) = x \). The absolute continuity, in turn, implies that \( \dot{\psi}(t) \) exists a.s. for \( t \in [0, 1] \). However, we cannot continue assuming that \( \dot{\psi} \) is piecewise continuous.

The problem is that some additional regularity of \( \psi \) the argument used to derive (6.27) fails. Even if one could design a control scheme that would yield the convergence of \( \bar{X}^n \to \psi \) with the appropriate cost for each \( n \), there could still be a difficulty. Indeed, since in general the set \( \{ \| \dot{\psi}(t) \|, t \in [0, 1] \} \) might no longer be bounded, the Lebesgue Dominated Convergence Theorem might no longer be applicable. For this reason we adopt an approximation procedure involving a special class of functions \( \mathcal{N} \).

**Approximation Procedure 6.4.3.** We seek a class of functions \( \mathcal{N} \subset C([0, 1] : \mathbb{R}^d) \) having the following properties. For each \( \psi \in C([0, 1] : \mathbb{R}^d) \) satisfying \( I_x(\psi) < \infty \) and for each \( \delta > 0 \), there exists \( \psi^* \in \mathcal{N} \) such that

\[
\liminf_{n \to \infty} \frac{1}{n} \log E_x \{ \exp[-n h(X^n)] \} \geq -I_x(\psi^*) - h(\psi^*),
\]

or \( \| \psi^* - \psi \|_{\infty} \leq \delta \), \( I_x(\psi^*) \leq I_x(\psi) + \delta \).

Let us suppose that such a class \( \mathcal{N} \) can be found. Given \( \varepsilon > 0 \), we choose \( \psi \in C([0, 1] : \mathbb{R}^d) \) such that

\[
I_x(\psi) + h(\psi) \leq \inf_{\varphi \in C([0, 1] : \mathbb{R}^d)} \{ I_x(\varphi) + h(\varphi) \} + \varepsilon < \infty.
\]

Since \( h \) is bounded, this implies that \( I_x(\psi) < \infty \). Since \( h \) is continuous, there exists \( \psi^* \in \mathcal{N} \) so that

\[
h(\psi^*) \leq h(\psi) + \varepsilon \quad \text{and} \quad I_x(\psi^*) \leq I_x(\psi) + \varepsilon.
\]

It follows that

\[
\liminf_{n \to \infty} \frac{1}{n} \log E_x \{ \exp[-n h(X^n)] \} \\
\geq -I_x(\psi^*) - h(\psi^*) \\
\geq -I_x(\psi) - h(\psi) - 2\varepsilon \\
\geq - \inf_{\varphi \in C([0, 1] : \mathbb{R}^d)} \{ I_x(\varphi) + h(\varphi) \} - 3\varepsilon.
\]

Sending \( \varepsilon \to 0 \) yields the Laplace principle lower bound.

When proving a Laplace principle lower bound in this chapter and in the next, we will identify a class \( \mathcal{N} \) of functions \( \psi^* \) for which the Approximation Procedure 6.4.3 can be carried out. One chooses this class so that for each \( \psi^* \in \mathcal{N} \) one may easily construct an admissible control sequence such that for \( \psi = \psi^* \) (6.27) is valid and thus the lower bound (6.28) holds.

We return to the random walk model under consideration. The work earlier in this section suggests taking \( \mathcal{N} \) to be a subset of functions whose derivatives are piecewise constant with only finitely many jumps in \((0, 1)\). Besides the Approximation Procedure 6.4.3, we must also consider the continuity assumption on \( L \) that was used in order to obtain line four of (6.27). It is immediate from Lemma 6.5.2, which will be stated in the
next section, that we cannot in general expect such continuity to hold. However, the lack
of continuity of \(L\) will be seen not to matter when Condition 6.3.1 holds; i.e., when the
sets \(\text{ri}(\text{conv} \ S_{\mu(\{x\}))}\) are independent of \(x \in \mathbb{R}^d\) and contain the origin. For situations in
which this is not the case, we will assume Condition 6.3.2 and prove the Laplace principle
lower bound by means of a perturbation argument. This will be carried out in Section
6.6.

In the next section we give the first proof of the Laplace principle lower bound for the
random walk model.

6.5 Proof of the Laplace Principle Lower Bound Under Conditions 6.2.1 and 6.3.1

In this section the following proposition will be proved.

**Proposition 6.5.1.** We assume Conditions 6.2.1 and 6.3.1. For \(x \in \mathbb{R}^d\), we define \(I_x\)
as in Theorem 6.3.3. Then for all bounded continuous functions \(h\) mapping \(C([0, 1] : \mathbb{R}^d)\)
into \(\mathbb{R}\) we have the Laplace principle lower bound

\[
\liminf_{n \to \infty} \frac{1}{n} \log E_x \{\exp[-n h(X^n)]\} \geq - \inf_{\varphi \in C([0, 1] : \mathbb{R}^d)} \{I_x(\varphi) + h(\varphi)\}.
\]

Before proving the proposition, we point out an equivalent form of Condition 6.3.1
which is somewhat easier to implement. This condition states that the sets \(\text{ri}(\text{conv} \ S_{\mu(\{x\}))}\)
are independent of \(x \in \mathbb{R}^d\) and that \(0 \in \Sigma \equiv \text{ri}(\text{conv} \ S_{\mu(\{x\}))}); thus \(\Sigma\) is convex. It follows
that the affine hull of \(\Sigma\), \(\text{aff} \Sigma\), is a linear subspace of \(\mathbb{R}^d\). Now consider the case where
the interior of \(\Sigma\), relative to \(\mathbb{R}^d\), is the empty set. Then the interior of \(\Sigma\), relative to \(\text{aff} \Sigma\),
is \(\Sigma\). Furthermore, according to the definition

\[
X^n_{j+1} = X^n_j + \frac{1}{n}v_j(X^n_j),
\]

for \(n \in \mathbb{N}\) the random vectors \(\{X^n_j, j = 0, 1, \ldots, n\}\) all take values in the affine subspace

\[
\Sigma_x \triangleq \{u \in \mathbb{R}^d : u = w + x, w \in \text{aff} \Sigma\},
\]

where \(X^n_0 = x\) is the initial condition. Thus the paths of the piecewise–linearly interpo-
lated process \(X^n\) lie in \(C([0, 1] : \Sigma_x)\). Since by translation we can assume that the initial
point \(x\) equals 0, we conclude that when Condition 6.3.1 holds, the Laplace principle lower
bound can be proved by considering \(\text{aff} \Sigma\) in place of \(\mathbb{R}^d\). Equivalently, we can modify
Condition 6.3.1, without loss of generality, to include the condition that the interior of \(\Sigma\)
relative to \(\mathbb{R}^d\) is nonempty; i.e., that \(\Sigma\) is a nonempty open convex subset of \(\mathbb{R}^d\). Hence
we can and will assume for the remainder of this section that for each \(x \in \mathbb{R}^d\)

\[
0 \in \Sigma \triangleq \text{int}(\text{conv} S_{\mu(\{x\})}).
\]
6.5. **FIRST PROOF OF LOWER BOUND**

Part (a) of Lemma 6.2.3 states the continuity of \( H(x, \alpha) \) on \( \mathbb{R}^d \times \mathbb{R}^d \), a property which is a consequence of Condition 6.2.1. According to the next lemma, under Condition 6.2.1 and part (a) of Condition 6.3.1 this continuity property of \( H(x, \alpha) \) carries over to the continuity of the Legendre–Fenchel transform \( L(x, \beta) \) for \( (x, \beta) \in \mathbb{R}^d \times \Sigma \).

**Lemma 6.5.2.** Under Condition 6.2.1 and part (a) of Condition 6.3.1, the function \( L(x, \beta) \) defined in equation (6.4) has the following properties.

(a) For each \( x \in \mathbb{R}^d \) the set \( \text{ri}(\text{dom } L(x, \cdot)) = \text{ri}(\text{conv } S_{\mu(x)}) \) equals \( \Sigma \).

(b) \( L(x, \beta) \) is a continuous function of \( (x, \beta) \in \mathbb{R}^d \times \Sigma \).

**Comments on the Proof.** (a) Part (d) of Lemma 6.2.3 states that for each \( x \in \mathbb{R}^d \)
\( \text{ri}(\text{dom } L(x, \cdot)) = \text{ri}(\text{conv } S_{\mu(x)}) \). According to part (a) of Condition 6.3.1 this set equals \( \Sigma \).

(b) This requires a detailed derivation that is given in Section C.8. ■

The continuity of \( L(\cdot, \cdot) \) on \( \mathbb{R}^d \times \Sigma \) will play a key role in the proof of the Laplace principle lower bound. The first step in the proof is to identify the class \( \mathcal{N} \) of functions \( \psi \) for which the lower bound will be directly proved via (6.27).

**Identification of the class \( \mathcal{N} \).** \( \mathcal{N} \) consists of all functions \( \psi^* \in \mathcal{C}([0, 1] : \mathbb{R}^d) \) satisfying the following two conditions:

(a) \( \dot{\psi}^*(t) \) is piecewise constant with only finitely many jumps in the interval \((0, 1)\).

(b) \( \dot{\psi}^*(t) \in \Sigma \) for all \( t \in [0, 1] \).

In the proof of the Laplace principle lower bound, the only difference in the present section as compared to the previous section is that there \( L(\cdot, \cdot) \) is continuous on \( \mathbb{R}^d \times \mathbb{R}^d \) while here \( L(\cdot, \cdot) \) is continuous on \( \mathbb{R}^d \times \Sigma \). This affects the proof in the previous section only in verifying the uniform boundedness of the running costs

\[
\bar{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot) \mid \mu(\cdot | X^n_j)) \right\} = \bar{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} L(X^n_j, \dot{\psi}(j/n)) \right\}
\]

appearing in equation (6.26) and in verifying the third equality in formula (6.27) with \( \psi \equiv \psi^* \). However, in the present section these running costs are uniformly bounded in \( n \) since for any compact subset \( K \) of \( \mathbb{R}^d \)

\[
\sup_{x \in K} \sup_{t \in [0, 1]} L(x, \dot{\psi}^*(t)) < \infty,
\]

and the third equality in (6.27) is valid because of the continuity of \( L(\cdot, \cdot) \) on \( \mathbb{R}^d \times \Sigma \). With these modifications one proves exactly as in the previous section the upper limit

\[
\limsup_{n \to \infty} W^n(x) \leq I_x(\psi^*) + h(\psi^*)
\]
and thus the lower bound
\[
\liminf_{n \to \infty} \frac{1}{n} \log E_x \{ \exp \left[ -n h(X^n) \right] \} \geq -I_x(\psi^*) - h(\psi^*). \tag{6.29}
\]

In order to finish the proof of the Laplace principle lower bound under Conditions 6.2.1 and 6.3.1, we must verify that the Approximation Procedure 6.4.3 can be carried out. Let \( \psi \in \mathcal{C}([0,1]) \) satisfy \( I_x(\psi) < \infty \). For each \( \delta > 0 \) we must produce a function \( \psi^* \in \mathcal{N} \) such that the lower bound (6.29) holds and such that
\[
\| \psi^* - \psi \|_\infty \leq \delta \quad \text{and} \quad I_x(\psi^*) \leq I_x(\psi) + \delta. \tag{6.30}
\]
The lower bound (6.29) has already been checked for any \( \psi^* \in \mathcal{N} \). The verification of the two inequalities in the last display is carried out in the next two lemmas. For \( \theta > 0 \), define the closed convex set
\[
\Sigma^{-\theta} \triangleq \{ \beta \in \Sigma : \| \beta - b \| \geq \theta \quad \text{for all} \quad b \in \Sigma^c \}.
\]
In other words, \( \Sigma^{-\theta} \) consists of all points in \( \Sigma \) that are at least a distance \( \theta \) from the complement of \( \Sigma \). Since \( \Sigma \) is open, any point \( \beta \in \Sigma \) lies in \( \Sigma^{-\theta(\beta)} \) for some \( \theta(\beta) > 0 \).

**Lemma 6.5.3.** Assume Conditions 6.2.1 and 6.3.1. For \( x \in \mathbb{R}^d \), let \( \psi \in \mathcal{C}([0,1]) \) satisfy \( I_x(\psi) < \infty \). Then for each \( \delta > 0 \) there exist \( \theta \in (0, \infty) \) and \( \zeta \in \mathcal{C}([0,1]) \) such that
\[
\| \zeta - \psi \|_\infty \leq \delta/2 \quad \text{and} \quad I_x(\zeta) \leq I_x(\psi) + \delta/2
\]
and a.s. for \( t \in [0,1] \) \( \zeta(t) \in \Sigma^{-\theta} \) and \( \| \zeta(t) \| \leq 1/\theta \).

**Remark 6.5.4.** In the next chapter we will want to adapt the proof of Lemma 6.5.3 to the related but different setting that arises in Lemma 7.5.4. For that purpose, we note the four properties of \( L(x, \beta) \) that are needed: for some \( c \in (0, \infty) \)
\[
\sup_{t \in [0,1]} \sup_{\beta \in K(0,c)} L(\psi(t), \beta) < \infty,
\]
where \( K(0,c) \) is a closed ball of radius \( c \) centered at 0 and contained in \( \Sigma \); the uniform superlinearity of \( L(x, \beta) \); the nonnegativity of \( L(x, \beta) \); and the convexity of \( L(x, \beta) \) with respect to \( \beta \in \mathbb{R}^d \) for each \( x \in \mathbb{R}^d \). We emphasize that the continuity of \( L(x, \beta) \) is not required. ■

**Proof of Lemma 6.5.3.** We will obtain \( \zeta \) from \( \psi \) by a time change. Since \( I_x(\psi) < \infty \), \( \psi \) is absolutely continuous and \( \psi(0) = x \). Define \( \Psi \) to be the set of \( t \in (0,1) \) such that \( \psi(t) \) lies in the closure of \( \Sigma \), \( \text{cl} \ \Sigma \). Since \( \psi \) is absolutely continuous and \( L(x, \beta) = \infty \) for \( (x, \beta) \in \mathbb{R}^d \times (\text{cl} \ \Sigma)^c \) [Lemma 6.5.2 (a)], the Lebesgue measure of \( \Psi \) is 1.

We have agreed that without loss of generality the set \( \Sigma \) in Condition 6.3.1 can be taken to be a nonempty open convex subset of \( \mathbb{R}^d \). Since by part (b) of this condition \( 0 \) lies in \( \Sigma \), \( c \in (0,1) \) can be chosen so that
\[
K(0,c) \triangleq \{ \beta \in \mathbb{R}^d : \| \beta \| \leq c \} \subset \Sigma.
\]
The continuity of $L(\cdot, \cdot)$ on $\mathbb{R}^d \times \Sigma$ implies that
\[
\Gamma = \sup_{t \in [0,1]} \sup_{\beta \in K(0,c)} L(\psi(t), \beta)
\]
is finite. Let $\lambda \in (0, 1)$ be given. Defining
\[
D_\lambda \doteq \left\{ t \in \Psi : \|\dot{\psi}(t)\| > 1/\lambda \right\} \quad \text{and} \quad E_\lambda \doteq \left\{ t \in \Psi \setminus D_\lambda : \dot{\psi}(t) \not\in \Sigma^{-\lambda} \right\},
\]
we construct a time-rescaling map $S^\lambda : [0, 1] \mapsto [0, \infty)$ as follows. For $t \in [0, 1]$, define $S^\lambda(0) \doteq 0$ and
\[
\dot{S}^\lambda(t) \doteq \begin{cases} 
\|\dot{\psi}(t)\|/|c(1 - \lambda)| & \text{if } t \in D_\lambda \\
1/(1 - \lambda) & \text{if } t \in E_\lambda \\
1 & \text{if } t \in [0, 1] \setminus (D_\lambda \cup E_\lambda).
\end{cases}
\]
$S^\lambda$ is continuous, and since $c \in (0, 1)$ and $\lambda \in (0, 1)$, $\dot{S}^\lambda(t) \geq 1$ for all $t \in (0, 1)$. Thus $S^\lambda$ is a strictly increasing map of $[0, 1]$ onto the interval $[0, S^\lambda(1)]$, where $S^\lambda(1) \geq 1$. For $t \in [0, 1]$, we define the inverse function $T^\lambda$ by $T^\lambda(S^\lambda(t)) \doteq t$. For $t \in [0, S^\lambda(1)]$, we then define the absolutely continuous function
\[
\zeta^\lambda(t) \doteq \dot{\psi}(T^\lambda(t)),
\]
which as a function of $t \in [0, 1]$ may be thought of as a slowed version of $\psi$. The function $\zeta$ in the statement of the lemma will be taken to be $\zeta^\lambda(t)$ for a suitable choice of $\lambda$. We will need the following three properties of $S^\lambda$, which hold a.s. for $t \in [0, 1]$: 
\[
\dot{S}^\lambda(T^\lambda(t)) T^\lambda(t) = 1, \quad \zeta^\lambda(t) = \dot{\psi}(T^\lambda(t))/\dot{S}^\lambda(T^\lambda(t)), \quad \text{and} \quad \zeta^\lambda(S^\lambda(t)) = \dot{\psi}(t)/\dot{S}^\lambda(t).
\]
We first prove that for a suitable choice of $\lambda \in (0, 1)$, $\|\zeta^\lambda - \psi\|_\infty \leq \delta/2$. Since
\[
\|\zeta^\lambda - \psi\|_\infty = \sup_{t \in [0, 1]} \|\psi(T^\lambda(t)) - \psi(t)\| = \sup_{s \in [0, 1]} \|\psi(s) - \psi(S^\lambda(s))\|,
\]
it suffices to prove that $\lim_{\lambda \to 0} S^\lambda(s) = s$ uniformly for $s \in [0, 1]$. The fact that $I_x(\psi) < \infty$ implies
\[
\lim_{\lambda \to 0} \int_0^1 1_{D_\lambda}(s) L(\psi(s), \dot{\psi}(s)) \, ds = 0.
\]
Hence by the nonnegativity and uniform superlinearity of $L$ [Lemma 6.2.3 (b)–(c)]
\[
\lim_{\lambda \to 0} \int_0^1 1_{D_\lambda}(s) \|\dot{\psi}(s)\| \, ds = 0. \tag{6.31}
\]
Since $\dot{S}^\lambda(u) \geq 1$ for all $u \in (0, 1)$, it follows that
\[
\sup_{s \in [0, 1]} |S^\lambda(s) - s| \leq \int_0^1 (\dot{S}^\lambda(u) - 1) \, du
\]
\[
\leq \frac{1}{c(1 - \lambda)} \int_0^1 1_{D_\lambda}(u) \|\dot{\psi}(u)\| \, du + \frac{\lambda}{1 - \lambda} \int_0^1 1_{E_\lambda}(u) \, du,
\]
which converges to 0 as $\lambda \to 0$. We conclude that there exists $\lambda_1 \in (0, 1)$ such that

$$\|\zeta^\lambda - \psi\|_\infty \leq \delta/2 \text{ whenever } \lambda \in (0, \lambda_1].$$

We now estimate $I_x(\zeta^\lambda) - I_x(\psi)$ for $\lambda \in (0, \lambda_1]$. For $t \in D_\lambda$ the nonnegativity of $L$ and the definition of $\Gamma$ imply that

$$L\left(\psi(t), \dot{\psi}(t) / \dot{S}^\lambda(t)\right) \dot{S}^\lambda(t) = L\left(\psi(t), \dot{\psi}(t)\right) \leq L\left(\psi(t), c(1 - \lambda)\dot{\psi}(t) / \|\dot{\psi}(t)\|\right) \|\dot{\psi}(t)\| \frac{1}{c(1 - \lambda)}$$

$$\leq \Gamma \|\dot{\psi}(t)\| \frac{1}{c(1 - \lambda)},$$

while for each $x$ and $\beta$ in $\mathbb{R}^d$ the convexity of $L(x, \cdot)$ yields

$$L(x, (1 - \lambda)\beta) / (1 - \lambda) \leq L(x, \beta) + \lambda L(x, 0) / (1 - \lambda).$$

Hence for $t \in E_\lambda$

$$L\left(\psi(t), \dot{\psi}(t) / \dot{S}^\lambda(t)\right) \dot{S}^\lambda(t) - L\left(\psi(t), \dot{\psi}(t)\right) \leq \Gamma \lambda / (1 - \lambda).$$

Finally, a.s. for $t \in [0, 1] \setminus (D_\lambda \cup E_\lambda)$ the left hand side of the last display equals 0. We now combine these inequalities with the inequality $S^\lambda(1) \geq 1$, the nonnegativity of $L$, and the equality $\psi(t) = \zeta^\lambda(S^\lambda(t))$, deducing that

$$I_x(\zeta^\lambda) - I_x(\psi)$$

$$\leq \int_0^{S^\lambda(1)} L\left(\zeta^\lambda(s), \dot{\zeta}^\lambda(s)\right) ds - \int_0^1 L\left(\psi(s), \dot{\psi}(s)\right) ds$$

$$= \int_0^1 L\left(\zeta^\lambda(S^\lambda(t)), \dot{\zeta}^\lambda(S^\lambda(t))\right) \dot{S}^\lambda(t) dt - \int_0^1 L\left(\psi(s), \dot{\psi}(s)\right) ds$$

$$= \int_0^1 L\left(\psi(t), \dot{\psi}(t) / \dot{S}^\lambda(t)\right) \dot{S}^\lambda(t) dt - \int_0^1 L\left(\psi(s), \dot{\psi}(s)\right) ds$$

$$\leq \Gamma \left(\frac{1}{c(1 - \lambda)} \int_0^1 \|\dot{\psi}(s)\| ds + \frac{\lambda}{1 - \lambda}\right).$$

It follows from (6.31) that

$$\limsup_{\lambda \to 0} \left(I_x(\zeta^\lambda) - I_x(\psi)\right) \leq 0.$$

We conclude that there exists $\lambda_2 \in (0, \lambda_1]$ such that $I_x(\zeta^\lambda) - I_x(\psi) \leq \delta/2$ for all $\lambda \in (0, \lambda_2]$. For $\lambda \in (0, \lambda_2]$, the set

$$\Psi_\lambda = \left\{ t \in [0, 1] : T^\lambda(t) \in \Psi \right\} = \left\{ t \in [0, 1] : \dot{\psi}(T^\lambda(t)) \in \text{cl } \Sigma \right\}$$

has Lebesgue measure 1. For $t \in \Psi_\lambda$ such that $T^\lambda(t) \in D_\lambda$

$$\|\dot{\zeta}^\lambda(t)\| = c(1 - \lambda) \leq 1/\lambda.$$
By the choice of $c$, $\hat{\zeta}^{\lambda}(t) \in K(0, c(1 - \lambda)) \subset K(0, c) \subset \Sigma$; thus $\hat{\zeta}^{\lambda}(t) \in \Sigma^{-c\lambda}$. On the other hand, for $t \in \Psi_{\lambda}$ such that $T^{\lambda}(t) \in E_{\lambda}$

$$\hat{\zeta}^{\lambda}(t) = (1 - \lambda) \dot{\psi}(T^{\lambda}(t)) \in (1 - \lambda)(\text{cl } \Sigma).$$

Since $T^{\lambda}(t) \in E_{\lambda}$ implies $T^{\lambda}(t) \notin D_{\lambda}$, we also have $\|\hat{\zeta}^{\lambda}(t)\| \leq 1/\lambda$. Thus $\hat{\zeta}^{\lambda}(t)$ lies in the compact set $[(1 - \lambda)(\text{cl } \Sigma)] \cap K(0, 1/\lambda)$. Since by assumption 0 lies in the open convex set $\Sigma$, the compact set $[(1 - \lambda)(\text{cl } \Sigma)] \cap K(0, 1/\lambda)$ is a subset of $\Sigma$. Hence there exists $\theta_{1} > 0$ with the property that $\hat{\zeta}^{\lambda}(t) \in \Sigma^{-\theta_{1}}$ for all $t \in \Psi_{\lambda}$ for which $T^{\lambda}(t) \in E_{\lambda}$. Finally, for $t \in \Psi_{\lambda}$ such that $T^{\lambda}(t) \in [0, 1] \setminus (D_{\lambda} \cup E_{\lambda})$

$$\hat{\zeta}^{\lambda}(t) \in \Sigma^{-\lambda} \quad \text{and} \quad \|\hat{\zeta}^{\lambda}(t)\| \leq 1/\lambda.$$

Defining $\theta = \min\{c\lambda, \theta_{1}, \lambda\}$, we conclude that for all $t \in \Psi_{\lambda}$ $\hat{\zeta}^{\lambda}(t) \in \Sigma^{-\theta}$ and $\|\hat{\zeta}^{\lambda}(t)\| \leq 1/\theta$. We complete the proof of the lemma by taking any $\lambda \in (0, \lambda_{2}]$ and defining $\zeta(t) \equiv \zeta^{\lambda}(t)$ for $t \in [0, 1]$.

The next lemma is the final result in this section. The class $\mathcal{N}$ is defined after the statement of Lemma 6.5.2.

**Lemma 6.5.5.** Assume Conditions 6.2.1 and 6.3.1 and let $\delta > 0$ be given. For a given function $\psi \in C([0, 1] : \mathbb{R}^{d})$ satisfying $I_{x}(\psi) < \infty$, let $\zeta \in C([0, 1] : \mathbb{R}^{d})$ be the absolutely continuous function constructed in the previous lemma, Lemma 6.5.3. The following conclusions hold.

(a) There exists $\psi^{*} \in \mathcal{N}$ such that

$$\|\psi^{*} - \zeta\|_{\infty} \leq \delta /2 \quad \text{and} \quad I_{x}(\psi^{*}) \leq I_{x}(\zeta) + \delta /2.$$ 

(b) $\|\psi^{*} - \psi\|_{\infty} \leq \delta$ and $I_{x}(\psi^{*}) \leq I_{x}(\psi) + \delta$.

**Proof.** (a) For $c > 0$, $K(0, c)$ denotes the closed ball of radius $c$ centered at 0. Given $\lambda \in (0, 1)$ such that $\lambda^{-1} \in \mathcal{N}$, we define $\psi^{\lambda}(t), t \in [0, 1]$, by setting $\psi^{\lambda}(0) \equiv \zeta(0)$ and specifying the derivative

$$\psi^{\lambda}(t) \equiv \frac{1}{\lambda} \int_{k\lambda}^{(k+1)\lambda} \hat{\zeta}(s) \, ds = \frac{1}{\lambda} (\zeta(k\lambda + \lambda) - \zeta(k\lambda))$$

for $t \in [k\lambda, (k + 1)\lambda), k = 0, 1, \ldots, \lambda^{-1} - 1$. Since $\Sigma$ is convex, $\psi^{\lambda}$ is in $\mathcal{N}$, and since $\zeta$ is absolutely continuous on $[0, 1]$, $\lim_{\lambda \to 0} \psi^{\lambda}(t) = \dot{\zeta}(t)$ a.s. for $t \in [0, 1]$.

The continuity of $\zeta$ implies that for all sufficiently small $\lambda \in (0, 1)$ $\|\psi^{\lambda} - \zeta\|_{\infty} \leq \delta /2$. Since $\dot{\zeta}(t)$ lies in the compact convex set $\Sigma^{-\theta} \cap K(0, 1/\theta)$ a.s. for $t \in [0, 1]$, $\psi^{\lambda}(t)$ lies in this set for all $t \in [0, 1]$. By possibly reducing $\lambda$ further, we obtain the estimate

$$I_{x}(\psi^{\lambda}) = \int_{0}^{1} L\left(\psi^{\lambda}(t), \dot{\psi}^{\lambda}(t)\right) \, dt \leq \int_{0}^{1} L\left(\zeta(t), \dot{\zeta}(t)\right) \, dt + \delta /2 = I_{x}(\zeta) + \delta /2$$
from the uniform convergence of $\psi^\lambda$ to $\zeta$, from the a.s. convergence of $\psi^\lambda$ to $\zeta$, and from the uniform continuity of $L(\cdot, \cdot)$ on $Q \times (\Sigma^- \cap K(0, 1/\theta))$, where $Q$ is any compact subset of $\mathbb{R}^d$. Taking $\psi^* \equiv \psi^\lambda$ for sufficiently small $\lambda \in (0, 1)$ completes the proof of part (a).

(b) Combining Lemma 6.5.3 with part (a) of the present lemma yields

$$
\|\psi^* - \psi\|_\infty \leq \|\psi^* - \zeta\|_\infty + \|\zeta - \psi\|_\infty \leq \delta \quad \text{and} \quad I_x(\psi^*) \leq I_x(\zeta) + \delta/2 \leq I_x(\psi) + \delta.
$$

This is what we want to prove. ■

Since part (b) of Lemma 6.5.5 yields the two inequalities in (6.30), we have finished the proof of Proposition 6.5.1, the Laplace principle lower bound under Conditions 6.2.1 and 6.3.1. The proof of the Laplace principle stated in Theorem 6.3.3 under these two conditions is now complete.

A key step in the proof of Proposition 6.5.1 is the construction, in Lemma 6.5.5, of the function $\psi^* \in \mathcal{N}$. This is made possible by Condition 6.3.1. The latter may be replaced by any other conditions on $\mu(dy|x)$ which allow the construction of such a function $\psi^*$.

In the next section we give the second proof of the Laplace principle lower bound for the random walk model.

### 6.6 Proof of the Laplace Principle Lower Bound Under Conditions 6.2.1 and 6.3.2

In this section we will prove the following proposition, in which we state the Laplace principle lower bound under a condition different from that in the previous section.

**Proposition 6.6.1.** We assume Conditions 6.2.1 and 6.3.2. For $x \in \mathbb{R}^d$ we define $I_x$ as in Theorem 6.3.3. Then for all bounded continuous functions $h$ mapping $C([0, 1]: \mathbb{R}^d)$ into $\mathbb{R}$ we have the Laplace principle lower bound

$$
\liminf_{n \to \infty} \frac{1}{n} \log E_x\{\exp[-n h(X^n)]\} \geq - \inf_{\varphi \in C([0, 1]: \mathbb{R}^d)} \{I_x(\varphi) + h(\varphi)\}.
$$

A key step in the proof of the Laplace principle lower bound in the previous section was the regularization argument given in Lemma 6.5.3. However, when Condition 6.3.1 does not hold, this argument does not directly apply and the analysis is considerably more difficult. In the present section we will prove the Laplace principle lower bound via a perturbation argument. The idea is to modify the problem in such a way that the function corresponding to $L$ in the rate function is continuous on all of $\mathbb{R}^d \times \mathbb{R}^d$. In this way, the method for proving the Laplace principle lower bound in the previous section may be used. We then obtain the Laplace principle lower bound for the original problem by making the perturbation suitably small. Such an argument was used in [37] in the proof of a process-level result and in Section 4 of [88] in the proof of the large deviation lower bound in Cramér's Theorem.
According to Corollary 1.2.5, it suffices to prove the Laplace principle lower bound for any bounded, Lipschitz continuous function $h$ mapping $\mathbb{R}^d$ into $\mathbb{R}$. In order to do this, we must verify for each $x \in \mathbb{R}^d$ that the functions

$$W^n(x) \doteq -\frac{1}{n} \log E_x \{ \exp[-n h(X^n)] \}$$

satisfy the upper limit

$$\limsup_{n \to \infty} W^n(x) \leq \inf_{\varphi \in C([0,1];\mathbb{R}^d)} \{ I_x(\varphi) + h(\varphi) \}. \quad (6.32)$$

We carry this out by introducing small perturbations of the processes $\{X^n, n \in \mathbb{N}\}$.

For $\sigma > 0$, let $\{Y_{j,\sigma}, j \in \mathbb{N}_0\}$ be a sequence of i.i.d. Gaussian random vectors with mean 0 and covariance matrix $\sigma I$ which are independent of the i.i.d. random vector fields $\{v_j(x), j \in \mathbb{N}_0, x \in \mathbb{R}^d\}$. For $n \in \mathbb{N}$, $j \in \{0, 1, \ldots, n - 1\}$, and $\sigma > 0$, we consider the sequences of random vectors defined by $X^n_0 = x$, $U^n_0 = 0$, and

$$X^n_{j+1} \doteq X^n_j + \frac{1}{n} v_j(X^n_j), \quad U^n_{j+1,\sigma} \doteq U^n_{j,\sigma} + \frac{1}{n} Y_{j,\sigma}.$$ 

We now define piecewise linear processes $X^n = \{X^n(t), t \in [0, 1]\}$, $U^n_\sigma = \{U^n_\sigma(t), t \in [0, 1]\}$, and $Z^n_\sigma = \{Z^n_\sigma(t), t \in [0, 1]\}$ by

$$X^n(t) \doteq X^n_j + \left( t - \frac{j}{n} \right) v_j(X^n_j) \quad \text{for } t \in [j/n, (j + 1)/n], \ j = 0, 1, \ldots, n - 1,$n

$$U^n_\sigma(t) \doteq U^n_{j,\sigma} + \left( t - \frac{j}{n} \right) Y_{j,\sigma} \quad \text{for } t \in [j/n, (j + 1)/n], \ j = 0, 1, \ldots, n - 1,$n

and

$$Z^n_\sigma(t) \doteq X^n(t) + U^n_\sigma(t) \quad \text{for } t \in [0, 1].$$

$Z^n_\sigma$ is the piecewise linear interpolation of the random vectors $\{X^n_j, U^n_{j,\sigma}\}$, which satisfy

$$X^n_{j+1} + U^n_{j+1,\sigma} = X^n_j + U^n_{j,\sigma} + \frac{1}{n} \left( v_j(X^n_j) + Y_{j,\sigma} \right).$$

Although the sequence $\{X^n + U^n_{j,\sigma}\}$ does not form a Markov chain, $Z^n_\sigma$ has a form similar to that of $X^n$. Let $\rho_\sigma \in \mathcal{P}(\mathbb{R}^d)$ denote the $N(0, \sigma I)$ Gaussian measure on $\mathbb{R}^d$ that is the distribution of $Y_{0,\sigma}$; i.e., for Borel subsets $A$ of $\mathbb{R}^d$

$$\rho_\sigma(A) \doteq \frac{1}{(2\pi \sigma^2)^{d/2}} \int_A \exp(-\|y\|^2/2\sigma^2) \, dy.$$

The role played by the distributions

$$\mu(dy|x) = P\{v_j(x) \in dy\}$$

in the definition of $X^n$ is, in the definition of $Z^n_\sigma$, played by the distributions

$$P\{v_j(x) + Y_{j,\sigma} \in dy\} = (\mu \ast \rho_\sigma)(dy|x),$$
where $*$ denotes convolution. Since the support of each of these measures is all of $\mathbb{R}^d$, we will be able to apply Lemma 6.5.5. Furthermore, for $\sigma$ small $Z^n_{\sigma}$ can be viewed as a small perturbation of $X^n$.

Let $M$ denote the Lipschitz constant of $h$ and define $B \doteq 2\|h\|_{\infty}$. Then

$$h(Z^n_{\sigma}) = h(X^n + U^n_{\sigma}) \geq h(X^n) - (M\|U^n_{\sigma}\|_{\infty} \wedge B).$$

Since $X^n$ and $U^n_{\sigma}$ are independent,

$$\frac{1}{n} \log E_x\{\exp[-n h(Z^n_{\sigma})]\} \leq \frac{1}{n} \log E_x\{\exp[-n h(X^n)] \cdot \exp[n ((M\|U^n_{\sigma}\|_{\infty} \wedge B))]\}$$

$$= \frac{1}{n} \log E_x\{\exp[-n h(X^n)]\} + \frac{1}{n} \log E_x\{\exp[n ((M\|U^n_{\sigma}\|_{\infty} \wedge B))]\}$$

$$= -W^n(x) + \frac{1}{n} \log E_x\{\exp[n ((M\|U^n_{\sigma}\|_{\infty} \wedge B))]\}.$$

This implies that

$$\limsup_{n \to \infty} W^n(x) \leq \limsup_{n \to \infty} \left( -\frac{1}{n} \log E_x\{\exp[-n h(Z^n_{\sigma})]\} \right)$$

$$+ \limsup_{n \to \infty} \frac{1}{n} \log E_x\{\exp[n ((M\|U^n_{\sigma}\|_{\infty} \wedge B))]\}. \tag{6.33}$$

We will prove the desired upper limit (6.32) by obtaining suitable bounds on the last two quantities in this display.

The first step is to note that the stochastic kernel $\mu(\cdot|x) \doteq \rho_{\sigma}(\cdot)$ for all $x \in \mathbb{R}^d$ satisfies Condition 6.2.1 with cumulant generating function

$$H_{\sigma}(\alpha) \doteq \log \int_{\mathbb{R}^d} \exp\langle \alpha, y \rangle \rho_{\sigma}(dy) = \frac{\sigma^2}{2} \|\alpha\|^2.$$ 

$H_{\sigma}(\alpha)$ has the Legendre–Fenchel transform

$$L_{\sigma}(\beta) \doteq \sup_{\alpha \in \mathbb{R}^d} \{\langle \alpha, \beta \rangle - H_{\sigma}(\alpha)\} = \frac{1}{2\sigma^2} \|\beta\|^2.$$

By Proposition 6.2.2 the sequence $\{U^n_{\sigma}, n \in \mathbb{N}\}$ satisfies the Laplace principle upper bound with rate function

$$I_{0,\sigma}(\varphi) \doteq \int_0^1 \frac{1}{2\sigma^2} \|\hat{\varphi}(s)\|^2 ds$$

if $\varphi \in C([0,1]:\mathbb{R}^d)$ is absolutely continuous and $\varphi(0) = 0$; $I_{0,\sigma}(\varphi) \doteq \infty$ in all other cases. We thus obtain

$$\limsup_{n \to \infty} \frac{1}{n} \log E_x\{\exp[n ((M\|U^n_{\sigma}\|_{\infty} \wedge B))]\} \leq -\inf_{\varphi \in C([0,1]:\mathbb{R}^d)} \{I_{0,\sigma}(\varphi) - M\|\varphi\|_{\infty} \wedge B\}$$

$$\leq -\inf_{\varphi \in C([0,1]:\mathbb{R}^d)} \{I_{0,\sigma}(\varphi) - M\|\varphi\|_{\infty}\}.$$
6.6. SECOND PROOF OF LOWER BOUND

If we restrict to \( \varphi \in \mathcal{C}([0,1] : \mathbb{R}^d) \) having the form \( \varphi(t) = vt \) for some \( v \in \mathbb{R}^d \) and all \( t \in [0,1] \), then

\[
\inf_{\varphi \in \mathcal{C}([0,1] : \mathbb{R}^d)} \{ I_{0,\sigma}(\varphi) - M \| \varphi \|_\infty \} \leq \inf_{v \in \mathbb{R}^d} \left\{ \frac{\|v\|^2}{2\sigma^2} - M\|v\| \right\}.
\]

We claim in fact that the two infima are equal. Indeed, given \( \varphi \in \mathcal{C}([0,1] : \mathbb{R}^d) \) for which \( I_{0,\sigma}(\varphi) < \infty \), choose \( \tau \in [0,1] \) such that \( v = \varphi(\tau) \) satisfies \( \|v\| = \|\varphi\|_\infty \). Since \( \varphi(0) = 0 \),

\[
\|v\| \leq \int_0^\tau \|\dot{\varphi}(s)\| \, ds \leq \int_0^1 \|\dot{\varphi}(s)\| \, ds \leq \left( \int_0^1 \|\dot{\varphi}(s)\|^2 \, ds \right)^{1/2}
\]

and thus

\[
I_{0,\sigma}(\varphi) - M \|\varphi\|_\infty \geq \frac{\|v\|^2}{2\sigma^2} - M\|v\|.
\]

This yields the claim. We have shown that

\[
\limsup_{n \to \infty} \frac{1}{n} \log E_x \{ \exp[n(M\|U^n_\sigma\|_\infty) \wedge B] \} \leq \inf_{\varphi \in \mathcal{C}([0,1] : \mathbb{R}^d)} \{ I_{0,\sigma}(\varphi) - M \|\varphi\|_\infty \} - \inf_{r \geq 0} \left\{ \frac{r^2}{2\sigma^2} - Mr \right\} = -\frac{M^2\sigma^2}{2},
\]

which tends to 0 as \( \sigma \to 0 \). Combining this with (6.33), we see that the desired upper limit (6.32) is proved once we show that

\[
\limsup_{n \to \infty} \left( \frac{1}{n} \log E_x \{ \exp[-n h(Z^n_\sigma)] \} \right) \leq \inf_{\varphi \in \mathcal{C}([0,1] : \mathbb{R}^d)} \{ I_x(\varphi) + h(\varphi) \} + \theta(\sigma), \quad (6.34)
\]

where \( \theta(\sigma) \to 0 \) as \( \sigma \to 0 \).

We prove (6.34) via the weak convergence approach. Define

\[
W^n_\sigma(x) = -\frac{1}{n} \log E_x \{ \exp[-n h(Z^n_\sigma)] \},
\]

where \( E_x \) denotes expectation conditioned on \( Z^n_\sigma(0) = x \). For a given \( \varepsilon > 0 \), we choose \( \psi \in \mathcal{C}([0,1] : \mathbb{R}^d) \) satisfying

\[
I_x(\psi) + h(\psi) \leq \inf_{\varphi \in \mathcal{C}([0,1] : \mathbb{R}^d)} \{ I_x(\varphi) + h(\varphi) \} + \varepsilon < \infty.
\]

The boundedness of \( h \) implies that \( I_x(\psi) < \infty \). Because \( \varepsilon > 0 \) is arbitrary, the upper limit (6.34) follows if we prove

\[
\limsup_{n \to \infty} W^n_\sigma(x) \leq I_x(\psi) + h(\psi) + \theta(\sigma) = \int_0^1 L(\psi(t), \dot{\psi}(t)) \, dt + h(\psi) + \theta(\sigma), \quad (6.35)
\]

where \( \theta(\sigma) \to 0 \) as \( \sigma \to 0 \).

We would like to apply Theorem 4.3.1 to derive a representation formula for \( W^n_\sigma(x) \). While \( Z^n_\sigma \) is the piecewise linear interpolation of the random vectors \( \{X^n_j + U^n_{j\sigma}\} \), the
Theorem cannot directly be applied since this sequence does not form a Markov chain. However, \( h(Z^n_t) \) is a bounded continuous function of the \( \mathbb{R}^{2d} \)-valued process \((X^n_t, U^n_{t \sigma})\), for which the theorem can be applied. In comparison with the last section, the only difference is that the state space of the controlled process is now \( \mathbb{R}^{2d} \) rather than \( \mathbb{R}^d \). In addition, since the respective increments \( \nu_j \) and \( \Upsilon_j \) of \( X^n \) and \( U^n_{t \sigma} \) are independent, the role formerly played by the stochastic kernel \( \mu(dy|x) \) is now played by the product stochastic kernel\( (\mu \times \rho_{t \sigma})(dy \times dz|x) \) on \( \mathbb{R}^{2d} \). For \( x \in \mathbb{R}^d \) the latter is defined by

\[
(\mu \times \rho_{t \sigma})(dy \times dz|x) \equiv \mu(dy|x) \times \rho_{t \sigma}(dz).
\]

In the associated stochastic control problem the control applied at time \( j \in \{0, 1, \ldots, n-1\} \) is a stochastic kernel on \( \mathbb{R}^{2d} \) given \((\mathbb{R}^{2d})^{j+1}\). It is useful to reflect in the notation the natural partitioning of the state space \( \mathbb{R}^{2d} = \mathbb{R}^d \times \mathbb{R}^d \). Thus we will write the control applied at time \( j \) as \( \nu_j^n(dy \times dz|\xi, \mu) \), where \( \xi = (\xi_0, \xi_1, \ldots, \xi_j) \in (\mathbb{R}^d)^{j+1} \) and \( u = (u_0, u_1, \ldots, u_j) \in (\mathbb{R}^d)^{j+1} \); each coordinate \( \xi_i \) and \( u_i \) is an element of \( \mathbb{R}^d \). A sequence \( \{\nu_j^n, j = 0, 1, \ldots, n-1\} \) of such controls is called an admmissible control sequence on \( \mathbb{R}^{2d} \).

For \( j \in \{0, 1, \ldots, n-1\} \), the dynamics of the controlled system are given by

\[
\begin{align*}
X^n_{j+1} &= X^n_j + \frac{1}{n} \Upsilon_j^n \quad \text{and} \quad \bar{U}^n_{j+1} = \bar{U}^n_j + \frac{1}{n} \Upsilon_j^n
\end{align*}
\]

with initial conditions \( X^n_0 = x \) and \( \bar{U}^n_0 = 0 \). We also define

\[
Z^n_j = X^n_j + \bar{U}^n_j \quad \text{for each} \ j \in \{0, 1, \ldots, n\}.
\]

For \( j \in \{0, 1, \ldots, n-1\} \) the conditional distribution of the pair \((\bar{Y}^n_j, \bar{T}^n_j)\) is defined by

\[
\begin{align*}
\mathbb{P}_x\{(\bar{Y}^n_j, \bar{T}^n_j) \in dy \times dz|(\bar{X}^n_0, \ldots, \bar{X}^n_j), (\bar{U}^n_0, \ldots, \bar{U}^n_j)\} \\
&\equiv \nu_j^n(dy \times dz|(\bar{X}^n_0, \ldots, \bar{X}^n_j), (\bar{U}^n_0, \ldots, \bar{U}^n_j)).
\end{align*}
\]

We also introduce three processes,

\[
\bar{X}^n = \{\bar{X}^n(t), t \in [0, 1]\}, \bar{U}^n = \{\bar{U}^n(t), t \in [0, 1]\}, \quad \text{and} \quad \bar{Z}^n = \{\bar{Z}^n(t), t \in [0, 1]\},
\]

which are the respective piecewise linear interpolations of the random vectors

\[
\{\bar{X}^n_j, j = 0, 1, \ldots, n\}, \{\bar{U}^n_j, j = 0, 1, \ldots, n\}, \quad \text{and} \quad \{\bar{Z}^n_j, j = 0, 1, \ldots, n\}.
\]

These linear interpolations are defined as in previous sections. Note that \( \bar{Z}^n = \bar{X}^n + \bar{U}^n \). For all \( n \in \mathbb{N} \) the random variables and stochastic processes introduced in this paragraph are defined on a common probability space \( (\Omega, \mathbb{F}, \mathbb{P}_x) \).

We can express the minimal cost function of the stochastic control problem as

\[
V^n_{t \sigma}(x) \equiv \inf_{\{\nu_j^n\}} \mathbb{E}_x\left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu_j^n(\cdot)(\mu \times \rho_{t \sigma})(\cdot|\bar{X}^n_j)) + h(\bar{Z}^n) \right\}.
\]
6.6. SECOND PROOF OF LOWER BOUND

The infimum is taken over all admissible control sequences \( \{\nu_j^n\} \); \( \bar{E}_x \) denotes expectation with respect to \( \bar{P}_x \); as the first argument of the relative entropy

\[
\nu_j^n(\cdot) = \nu_j^n(\cdot| (\bar{X}_j^n, \bar{X}_j^n, \ldots, \bar{X}_j^n), (\bar{U}_0^n, \bar{U}_1^n, \ldots, \bar{U}_j^n));
\]

\( \bar{X}_j^n \), \( \bar{U}_j^n \), and \( \bar{Z}^n \) are the controlled random variables and the controlled process that are associated with a particular admissible control sequence \( \{\nu_j^n\} \); and \( h \) is the same function appearing in the definition of \( W^n_\sigma(x) \). Theorem 4.3.1 implies that for each \( n \in N \) and \( x \in \mathbb{R}^d \) \( W^n_\sigma(x) \) equals the minimal cost function \( V^n_\sigma(x) \).

In order to prove the upper limit (6.35) and thus the upper limit (6.34), we can restrict the infimum in the definition of \( V^n_\sigma(x) \) to an appropriate subset of admissible control sequences. For each \( j \in \{0, 1, \ldots, n - 1\} \) let \( \nu_j^{1,n}(\cdot|\bar{\xi}) \) and \( \nu_j^{2,n}(\cdot|\bar{\xi}) \) be stochastic kernels on \( \mathbb{R}^d \) given \( (\mathbb{R}^d)^{j+1} \). It will turn out to be sufficient to consider only controls having the product form

\[
\nu_{j,\text{prod}}^n(dy \times dz|\xi) = \nu_j^{1,n}(dy|\xi) \times \nu_j^{2,n}(dz|\xi),
\]

in which the parameter \( u \) is absent. Each of the sequences \( \{\nu_j^{1,n}\} \) and \( \{\nu_j^{2,n}\} \) will be called an admissible control sequence on \( \mathbb{R}^d \).

Corollary C.3.3 implies that for controls of product form the running cost partitions nicely, in the sense that

\[
\bar{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R\left( \nu_{j,\text{prod}}^n(\cdot) \right) \right\} = \bar{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \left[ R\left( \nu_j^{1,n}(\cdot) \right) + R\left( \nu_j^{2,n}(\cdot) \right) \right] \right\}.
\]

If in the definition of \( V^n_\sigma(x) \) the infimum is restricted to admissible control sequences of product form (6.37), then we obtain

\[
V^n_\sigma(x) \leq \inf_{\{\nu_j^{1,n}\}, \{\nu_j^{2,n}\}} \bar{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \left[ R\left( \nu_j^{1,n}(\cdot) \right) + R\left( \nu_j^{2,n}(\cdot) \right) \right] \right\}.
\]

The infimum is taken over all admissible control sequences \( \{\nu_j^{1,n}\} \) and \( \{\nu_j^{2,n}\} \) on \( \mathbb{R}^d \), and as the first arguments of the relative entropy \( \nu_j^{1,n}(\cdot) = \nu_j^{1,n}(\cdot|\bar{X}_0^n, \bar{X}_1^n, \ldots, \bar{X}_j^n) \) and \( \nu_j^{2,n}(\cdot) = \nu_j^{2,n}(\cdot|\bar{X}_0^n, \bar{X}_1^n, \ldots, \bar{X}_j^n) \).

For \( x \) and \( \beta \) in \( \mathbb{R}^d \) we consider the Legendre–Fenchel transforms of

\[
H(x, \alpha) \doteq \log \int_{\mathbb{R}^d} \exp(\langle \alpha, y \rangle) \mu(dy|x) \quad \text{and} \quad H_\sigma(\alpha) \doteq \log \int_{\mathbb{R}^d} \exp(\langle \alpha, y \rangle) \rho_\sigma(dy) = \frac{\sigma^2}{2} \|\alpha\|^2.
\]

These functions are defined by

\[
L(x, \beta) \doteq \sup_{\alpha \in \mathbb{R}^d} \{ \langle \alpha, \beta \rangle - H(x, \alpha) \} \quad \text{and} \quad L_\sigma(\beta) \doteq \sup_{\alpha \in \mathbb{R}^d} \{ \langle \alpha, \beta \rangle - H_\sigma(\alpha) \} = \frac{1}{2\sigma^2} \|\beta\|^2.
\]
CHAPTER 6. RANDOM WALK WITH CONTINUOUS STATISTICS

The key step in the proof of (6.35) is to obtain an upper limit of the form

$$\limsup_{n \to \infty} V^n_\sigma(x) \leq \int_0^1 L_\sigma(\psi(t), \dot{\psi}(t)) \, dt + h(\psi) + \theta(\sigma),$$

where $L_\sigma(x, \beta)$ is a suitable perturbation of $L(x, \beta)$ and $\theta(\sigma) \to 0$ as $\sigma \to 0$. For $x \in \mathbb{R}^d$, we denote by $\mu_\sigma(\cdot | x)$ the distribution of the sum $\nu_0(x) + \chi_{0, \sigma}$. Thus for each Borel subset $A$ of $\mathbb{R}^d$

$$\mu_\sigma(A|x) = (\mu * \rho_\sigma)(A|x) = \int_{\mathbb{R}^d} \mu(A - z|x) \rho_\sigma(dz).$$

The function $L_\sigma(x, \beta)$ is defined to be the Legendre–Fenchel transform of

$$H_\sigma(x, \alpha) = \log \int_{\mathbb{R}^d} \exp(\langle \alpha, y \rangle) \mu_\sigma(dy|x)$$

$$= \log \int_{\mathbb{R}^d \times \mathbb{R}^d} \exp(\langle \alpha, w + z \rangle) \mu(dw|x) \rho_\sigma(dz)$$

$$= H(x, \alpha) + H_\sigma(\alpha) = H(x, \alpha) + \frac{\sigma^2}{2} \| \alpha \|^2.$$

A number of properties of $L_\sigma(x, \beta)$ are given in the next lemma.

**Lemma 6.6.2.** We assume Conditions 6.2.1 and 6.3.2. For $\sigma > 0$ and $x$ and $\beta$ in $\mathbb{R}^d$ we define

$$L_\sigma(x, \beta) = \sup_{\alpha \in \mathbb{R}^d} \left\{ \langle \alpha, \beta \rangle - H(x, \alpha) - \frac{\sigma^2}{2} \| \alpha \|^2 \right\}.$$

The following conclusions hold.

(a) $L_\sigma(x, \beta) = \inf_{\zeta \in \mathbb{R}^d} \{ L(x, \beta - \zeta) + \frac{1}{2\sigma^2} \| \zeta \|^2 \}$ and $L_\sigma(x, \beta) \leq L(x, \beta)$.

(b) $L_\sigma(x, \beta)$ is a finite, nonnegative, continuous function of $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$.

(c) The properties (6.16) and (6.17) of $L$ assumed in Condition 6.3.2 are inherited by $L_\sigma$. Thus given any compact subset $\Delta$ of $\mathbb{R}^d$ and any number $\varepsilon \in (0, 1)$, we can find $\eta = \eta(\Delta, \varepsilon) \in (0, 1)$ and $\tilde{K} = \tilde{K}(\Delta, \varepsilon) \in (0, \infty)$ so that whenever $\xi \in \Delta$, $\nu \in \Delta$, and $\beta \in \mathbb{R}^d$ satisfy $\| \xi - \nu \| \leq \eta$, there is a $\tilde{\beta} \in \mathbb{R}^d$ such that

$$L_\sigma(\xi, \tilde{\beta}) - L_\sigma(\nu, \beta) \leq \varepsilon [1 + L_\sigma(\nu, \beta)] \quad (6.40)$$

and

$$\| \tilde{\beta} - \beta \| \leq \tilde{K} \| \xi - \nu \| [1 + L_\sigma(\nu, \beta)]. \quad (6.41)$$

(d) Given any $\psi \in C([0, 1] : \mathbb{R}^d)$ satisfying $I_x(\psi) < \infty$ and any $\varepsilon > 0$, there exists $\psi^* \in \mathcal{N}$ satisfying $\| \psi^* - \psi \|_\infty \leq \varepsilon$ and

$$\int_0^1 L_\sigma(\psi^*(t), \dot{\psi}^*(t)) \, dt \leq \int_0^1 L_\sigma(\psi(t), \dot{\psi}(t)) \, dt + \varepsilon \leq I_x(\psi) + \varepsilon.$$

The class $\mathcal{N}$ is defined after the statement of Lemma 6.5.2.
Proof. (a) For $x \in \mathbb{R}^d$ and $\sigma > 0$ $L(x, \beta)$ and $\|\beta\|^2/2\sigma$ are the respective Legendre-Fenchel transforms of the finite convex functions $H(x, \alpha)$ and $\sigma\|\alpha\|^2/2$. Corollary D.4.2 yields the equality in part (a), which in turn yields the subsequent inequality.

(b) We have assumed that $H(x, \alpha)$ and $\mu(dy|x)$ satisfy Condition 6.2.1. Hence

$$H_\sigma(x, \alpha) \doteq H(x, \alpha) + \frac{\sigma^2}{2}\|\alpha\|^2$$

and

$$\mu_\sigma(A|x) \doteq \int_{\mathbb{R}^d} \mu(A - z|x) \rho_\sigma(dz)$$

also satisfy Condition 6.2.1. In addition, for each $x \in \mathbb{R}^d$ the support of $\mu_\sigma(\cdot|x)$ is all of $\mathbb{R}^d$, and thus $\text{ri}(\text{dom } L_\sigma(x, \cdot)) = \mathbb{R}^d$. It follows that $L_\sigma(x, \beta) < \infty$ for all $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$ and that $\mu_\sigma(\cdot|x)$ satisfies part (a) of Condition 6.3.1 with $\Sigma = \mathbb{R}^d$. As with the continuity of $L(x, \beta)$ asserted in part (b) of Lemma 6.5.2, Lemma C.8.1 implies that $L_\sigma(x, \beta)$ is continuous for all $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$. The nonnegativity of $L_\sigma(x, \beta)$ follows from part (b) of Lemma 6.2.3.

(c) For each $v$ and $\beta$ in $\mathbb{R}^d$ and any $\varepsilon \in (0, \varepsilon/4)$, there is a $b \in \mathbb{R}^d$ such that

$$L_\sigma(v, \beta) \geq L(v, \beta - b) + \frac{1}{2\sigma^2}\|b\|^2 - \varepsilon. \quad (6.42)$$

According to Condition 6.3.2 with $\gamma \doteq \beta - b$, we can find $\eta = \eta(\Delta, \varepsilon) \in (0, 1)$ and $K = K(\Delta, \varepsilon) < \infty$ such that whenever $\xi \in \Delta$ and $v \in \Delta$ satisfy $\|\xi - v\| \leq \eta$, there is a $\bar{\beta} \in \mathbb{R}^d$ such that

$$\|\bar{\beta} - \beta\| = \|\bar{\beta} - b - (\beta - b)\| \leq K \|\xi - v\| \left[ 1 + L(v, \beta - b) \right]$$

$$\leq K \|\xi - v\| \left[ 1 + L_\sigma(v, \beta) + \varepsilon \right]$$

$$\leq 2K \|\xi - v\| \left[ 1 + L_\sigma(v, \beta) \right]$$

and

$$L(\xi, \bar{\beta} - b) - L(v, \beta - b) \leq \frac{\varepsilon}{2} \left[ 1 + L(v, \beta - b) \right].$$

The first of these displays yields (6.41) with $\bar{K} \doteq 2K$. The inequality (6.40) is a consequence of the definition of $L_\sigma(\xi, \bar{\beta})$, formula (6.42), and the following string of inequalities:

$$L_\sigma(\xi, \bar{\beta}) - L_\sigma(v, \beta) \leq L(\xi, \bar{\beta} - b) + \frac{1}{2\sigma^2}\|b\|^2 - L(v, \beta - b) - \frac{1}{2\sigma^2}\|b\|^2 + \varepsilon$$

$$\leq L(\xi, \bar{\beta} - b) - L(v, \beta - b) + \varepsilon$$

$$\leq \frac{\varepsilon}{2} \left[ 1 + L(v, \beta - b) \right] + \varepsilon$$

$$\leq \frac{\varepsilon}{2} \left[ 1 + L(v, \beta - b) + \frac{1}{2\sigma^2}\|b\|^2 \right] + \varepsilon$$

$$\leq \varepsilon \left[ 1 + L_\sigma(v, \beta) \right].$$

This completes the proof of part (c).
(d) In part (b) we noted that $H(x,\alpha)$ and $\mu_\sigma(dy|x)$ satisfy Conditions 6.2.1 and 6.3.1. Hence we can apply Lemma 6.5.5, which yields $\psi^* \in \mathcal{N}$ satisfying $\|\psi^* - \psi\|_\infty \leq \varepsilon$ and

$$\int_0^1 L_\sigma(\psi^*(t), \dot{\psi}^*(t)) dt \leq \int_0^1 L_\sigma(\psi(t), \dot{\psi}(t)) dt + \varepsilon.$$ 

By part (a) of the present lemma

$$\int_0^1 L_\sigma(\psi(t), \dot{\psi}(t)) dt + \varepsilon \leq \int_0^1 L(\psi(t), \dot{\psi}(t)) dt + \varepsilon = I_\varepsilon(\psi) + \varepsilon.$$ 

Combining the last two displays completes the proof of part (d). ■

We now turn to the proof that

$$\limsup_{m \to \infty} V^\sigma_m(x) \leq \int_0^1 L_\sigma(\psi(t), \dot{\psi}(t)) dt + h(\psi) + \theta(\sigma),$$

(6.43)

where $\theta(\sigma) \to 0$ as $\sigma \to 0$. Since $L_\sigma(x, \beta) \leq L(x, \beta)$ for all $x$ and $\beta$ in $\mathbb{R}^d$ [Lemma 6.6.2 (a)], we will obtain the upper limit (6.35) for $W^\sigma_m(x) = V^\sigma_m(x)$.

In order to prove (6.43), we will design an admissible control sequence that is based on the function $\psi^*$ in part (d) of Lemma 6.6.2 and has the following two properties: the running costs are nearly optimal and with probability converging to 1 the processes $\tilde{Z}^n = \tilde{X}^n + \tilde{U}^n$ enter a small neighborhood of $\psi$ as $n \to \infty$. The basic problem in fulfilling both of these requirements is that for a fixed admissible control sequence the running cost depends on $\tilde{X}^n$ through $R(\nu_{\tilde{j}}^1 \mu_{(\tilde{j})})$ and $R(\nu_{\tilde{j}}^2 \rho_{(\tilde{j})})$ while the terminal cost depends on $\tilde{Z}^n$. We will first choose an admissible control sequence that will yield the desired upper bound on the running costs and then show that the Lipschitz properties of $L_\sigma$ expressed by formulas (6.40) and (6.41) allow us to control the terminal cost.

Let $\psi^*$ be the function associated by part (d) of Lemma 6.6.2 with the given function $\psi \in C([0,1] : \mathbb{R}^d)$ and the choice $\varepsilon \equiv \sigma$. We will work with the numbers $\eta = \eta(\Delta, \sigma) \in (0,1)$ and $\tilde{K} = \tilde{K}(\Delta, \sigma)$ associated by part (c) of Lemma 6.6.2 with $\sigma$ and with

$$\Delta \equiv \bigcup_{t \in [0,1]} \{y \in \mathbb{R}^d : \|y - \psi^*(t)\| \leq 1\}.$$

$\Delta$ is compact since $\psi^*$ is continuous. For each $n \in \mathbb{N}$ we define suitable admissible control sequences $\{\nu_{\tilde{j}}^1 \}, \{\nu_{\tilde{j}}^2 \}$ and controlled random vectors $\{\tilde{X}^n_{\tilde{j}} \}, \{\tilde{U}^n_{\tilde{j}} \}$ by adapting the construction given in Procedure 6.4.1. The admissible controls applied at time $\tilde{j}$ will be stochastic kernels $\nu_{\tilde{j}}^1(dy|x)$ and $\nu_{\tilde{j}}^2(dy|x)$ on $\mathbb{R}^d$ given $\mathbb{R}^d$ the definition of these stochastic kernels will depend on whether $\max_{i=0,1,\ldots,\tilde{j}} \|\xi_{\tilde{i}} - \psi^*(i/n)\| \leq \eta$ or $\max_{i=0,1,\ldots,\tilde{j}} \|\xi_{\tilde{i}} - \psi^*(i/n)\| > \eta$. If $\max_{i=0,1,\ldots,\tilde{j}} \|\xi_{\tilde{i}} - \psi^*(i/n)\| \leq \eta$, then $\|\xi_{\tilde{j}} - \psi^*(j/n)\| \leq \eta$ and $\Delta$ contains $\xi_{\tilde{j}}$ and $\psi^*(j/n)$. In this case part (c) of Lemma 6.6.2 guarantees that with

$$\varepsilon \equiv \sigma, \xi \equiv \xi_{\tilde{j}}, \psi \equiv \psi^*(j/n),$$

and $\beta \equiv \dot{\psi}^*(j/n)$,

there exists $\tilde{\beta}^n_{\tilde{j}} \in \mathbb{R}^d$ such that (6.40) and (6.41) hold with $\tilde{\beta}^n_{\tilde{j}}$; i.e.,

$$L_\sigma(\xi_{\tilde{j}}, \tilde{\beta}^n_{\tilde{j}}) - L_\sigma(\psi^*(j/n), \dot{\psi}(j/n)) \leq \sigma \left[1 + L_\sigma(\psi^*(j/n), \dot{\psi}^*(j/n))\right]$$

(6.44)
and
\[
\|\hat{\beta}_j^n - \hat{\psi}^*(j/n)\| \leq K \|\xi_j - \psi^*(j/n)\| \left[1 + L_\sigma(\psi^*(j/n), \hat{\psi}^*(j/n))\right].
\]  
(6.45)

Since \(L_\sigma(\cdot, \cdot)\) is continuous, \(\hat{\beta}_j^n\) can be picked as a measurable function of \(\xi_j\). In turn, according to part (a) of the lemma, there exist \(\beta_j^{1,n}\) and \(\beta_j^{2,n}\) in \(\mathbb{R}^d\) satisfying \(\beta_j^{1,n} + \beta_j^{2,n} = \hat{\beta}_j^n\) and
\[
L(\xi_j, \beta_j^{1,n}) + \frac{1}{2\sigma} \|\beta_j^{2,n}\|^2 \leq L_\sigma(\xi_j, \hat{\beta}_j^n) + \sigma.
\]  
(6.46)

We claim that \(\beta_j^{1,n}\) and \(\beta_j^{2,n}\) can be picked as measurable functions of \(\xi_j\). Indeed, one can show that the infimum in the definition of \(L_\sigma(\xi_j, \hat{\beta}_j^n)\) is uniquely attained at some \(\zeta\) and that if \(L(\cdot, \cdot)\) is continuous, then \(\zeta\) is a measurable function of \(\xi_j\) and \(\hat{\beta}_j^n\). In the general case where \(L(\cdot, \cdot)\) is only lower semicontinuous, one can prove the same result by approximating \(L\) from below by continuous functions.

We now apply parts (f) and (g) of Lemma 6.2.3. There exists a stochastic kernel \(\gamma_j^{1,n}(dy|\xi_j)\) on \(\mathbb{R}^d\) given \(\mathbb{R}^d\) satisfying for each \(\xi_j \in \mathbb{R}^d\)
\[
R(\gamma_j^{1,n}(\cdot|\xi_j)\|\mu(\cdot|\xi_j)) = L(\xi_j, \beta_j^{1,n}) \quad \text{and} \quad \int_{\mathbb{R}^d} y \gamma_j^{1,n}(dy|\xi_j) = \beta_j^{1,n},
\]
and there exists a probability measure \(\gamma_j^{2,n}\) on \(\mathbb{R}^d\) satisfying
\[
R(\gamma_j^{2,n}(\cdot)\|\rho_\sigma) = \frac{1}{2\sigma} \|\beta_j^{2,n}\|^2 \quad \text{and} \quad \int_{\mathbb{R}^d} y \gamma_j^{2,n}(dy) = \beta_j^{2,n}.
\]

For \(\xi \in (\mathbb{R}^d)^{j+1}\) we now define stochastic kernels \(\nu_j^{1,n}(dy|\xi)\) and \(\nu_j^{2,n}(dy|\xi)\) by
\[
\nu_j^{1,n}(dy|\xi) = \begin{cases} 
\gamma_j^{1,n}(dy|\xi_j) & \text{if max}_{i=0,1,\ldots,j} \|\xi_i - \psi^*(i/n)\| \leq \eta \\
\mu(dy|\xi_j) & \text{if max}_{i=0,1,\ldots,j} \|\xi_i - \psi^*(i/n)\| > \eta
\end{cases}
\]
and
\[
\nu_j^{2,n}(dy|\xi) = \begin{cases} 
\gamma_j^{2,n}(dy) & \text{if max}_{i=0,1,\ldots,j} \|\xi_i - \psi^*(i/n)\| \leq \eta \\
\rho_\sigma & \text{if max}_{i=0,1,\ldots,j} \|\xi_i - \psi^*(i/n)\| > \eta
\end{cases}
\]
Both \(\nu_j^{1,n}(dy|\xi)\) and \(\nu_j^{2,n}(dy|\xi)\) depend on \(\sigma\), but this dependence will not be indicated in the notation.

Having defined the stochastic kernels \(\nu_j^{1,n}(dy|\xi)\) and \(\nu_j^{2,n}(dy|\xi)\), we set \(\bar{X}_0^n = x\) and \(\bar{U}_0^n = 0\), where \(x\) is a fixed point in \(\mathbb{R}^d\), and we take \(\bar{Y}_j^n\) and \(\bar{T}_j^n\) to be random vectors with the conditional distribution
\[
P_x\{(\bar{Y}_j^n, \bar{T}_j^n) \in dy \times dz | \bar{X}_0^n, \ldots, \bar{X}_j^n, \bar{U}_0^n, \ldots, \bar{U}_j^n\} = \nu_j^{1,n}(dy|\bar{X}_0^n, \ldots, \bar{X}_j^n) \times \nu_j^{2,n}(dz|\bar{X}_0^n, \ldots, \bar{X}_j^n).
\]  
(6.47)

We then define
\[
\bar{X}_{j+1}^n \stackrel{d}{=} \bar{X}_j^n + \frac{1}{n} \bar{Y}_j^n \quad \text{and} \quad \bar{U}_{j+1}^n \stackrel{d}{=} \bar{U}_j^n + \frac{1}{n} \bar{T}_j^n.
\]
In terms of the stopping time

\[ \kappa^n = \frac{1}{n} \left( \min \left\{ i \in \{0, 1, \ldots, n\} : \| \hat{X}^n_i - \psi^*(i/n) \| > \eta \right\} \wedge n \right), \]

(6.48)

the admissible controls \( \nu_j^{1,n}(\cdot) = \nu_j^{1,n}(\cdot|X^n_0, \ldots, X^n_j) \) and \( \nu_j^{2,n}(\cdot) = \nu_j^{2,n}(\cdot|X^n_0, \ldots, X^n_j) \) have the following properties: for \( j \in \{0, 1, \ldots, n\kappa^n - 1\} \)

\[ R\left( \nu_j^{1,n}(\cdot) \| \mu(\cdot|X^n_j) \right) = L\left( X^n_j, \beta_j^{1,n} \right), \int_{\mathbb{R}^d} y \nu_j^{1,n}(dy) = \beta_j^{1,n} \]

(6.49)

and

\[ R\left( \nu_j^{2,n}(\cdot) \| \rho_\sigma \right) = \frac{1}{2\sigma} \left\| \beta_j^{2,n} \right\|^2, \int_{\mathbb{R}^d} y \nu_j^{2,n}(dy) = \beta_j^{2,n}. \]

(6.50)

For \( j \in \{n\kappa^n, n\kappa^n + 1, \ldots, n - 1\} \) \( \nu_j^{1,n}(\cdot) \) equals \( \mu(\cdot|X^n_j) \) and \( \nu_j^{2,n}(\cdot) \) equals \( \rho_\sigma(\cdot) \). Hence both \( R(\nu_j^{1,n}(\cdot) \| \mu(\cdot|X^n_j)) \) and \( R(\nu_j^{2,n}(\cdot) \| \rho_\sigma) \) equal 0. In addition, for \( j \in \{0, 1, \ldots, n\kappa^n - 1\} \), since \( \max_{i=0,1,\ldots,j} \| \hat{X}^n_i - \psi^*(i/n) \| \leq \eta, (6.44), (6.45), \) and (6.46) hold with \( \xi_j = \hat{X}^n_j; \) i.e.,

\[ L_\sigma\left( \hat{X}^n_j, \beta_j^{2,n} \right) - L_\sigma\left( \psi^*(j/n), \psi^*(j/n) \right) \leq \sigma \left[ 1 + L_\sigma\left( \psi^*(j/n), \psi^*(j/n) \right) \right], \]

(6.51)

\[ \left\| \beta_j^{2,n} - \psi^*(j/n) \right\| \leq \bar{K} \| \hat{X}^n_j - \psi^*(j/n) \| \left[ 1 + L_\sigma\left( \psi^*(j/n), \psi^*(j/n) \right) \right], \]

(6.52)

and

\[ L\left( X^n_j, \beta_j^{1,n} \right) + \frac{1}{2\sigma} \left\| \beta_j^{2,n} \right\|^2 \leq L_\sigma\left( \hat{X}^n_j, \beta_j^{2,n} \right) + \sigma. \]

(6.53)

To the admissible control sequences \( \{\nu_j^{1,n}(dy|\xi)\} \) and \( \{\nu_j^{2,n}(dy|\xi)\} \) just defined, we associate the admissible control sequence \( \{\nu_j^{n\text{prod}}(dy|\xi)\} \) on \( \mathbb{R}^{2d} \) of product form by setting

\[ \nu_j^{n\text{prod}}(dy \times dz|\xi) \equiv \nu_j^{1,n}(dy|\xi) \times \nu_j^{2,n}(dz|\xi). \]

(6.54)

The corresponding admissible control measures must also be constructed. For \( n \in \mathbb{N} \) and \( t \in [0, 1] \) set

\[ \nu_j^{1,n}(dy|t) \overset{\text{def}}{=} \begin{cases} \nu_j^{1,n}(dy) & \text{if } t \in [j/n, (j + 1)/n), \ j = 0, 1, \ldots, n - 2 \\ \nu_{n-1}^{1,n}(dy) & \text{if } t \in [(n - 1)/n, 1] \end{cases} \]

and

\[ \nu_j^{2,n}(dz|t) \overset{\text{def}}{=} \begin{cases} \nu_j^{2,n}(dz) & \text{if } t \in [j/n, (j + 1)/n), \ j = 0, 1, \ldots, n - 2 \\ \nu_{n-1}^{2,n}(dz) & \text{if } t \in [(n - 1)/n, 1] \end{cases} \]

where \( \nu_j^{1,n}(\cdot) = \nu_j^{1,n}(\cdot|X^n_0, X^n_1, \ldots, X^n_j) \) and \( \nu_j^{2,n}(\cdot) = \nu_j^{2,n}(\cdot|X^n_0, \ldots, X^n_j) \). We now define admissible control measures \( \nu_j^{1,n} \) and \( \nu_j^{2,n} \) on \( \mathbb{R}^d \times [0, 1] \) by

\[ \nu_j^{1,n}(A \times B) \overset{\text{def}}{=} \int_B \nu_j^{1,n}(A|t) dt \quad \text{and} \quad \nu_j^{2,n}(A \times B) \overset{\text{def}}{=} \int_B \nu_j^{2,n}(A|t) dt \]

for Borel subsets \( A \) of \( \mathbb{R}^d \) and \( B \) of \( [0, 1] \). Finally, for Borel subsets \( A_1 \) and \( A_2 \) of \( \mathbb{R}^d \) and \( B \) of \( [0, 1] \) we define the admissible control measure \( \nu_{\text{prod}} \) on \( \mathbb{R}^{2d} \times [0, 1] \) by

\[ \nu_{\text{prod}}(A_1 \times A_2 \times B) \overset{\text{def}}{=} \int_B \nu_j^{1,n}(A_1|t) \nu_j^{2,n}(A_2|t) dt. \]
The conditional distribution of the pair \((\hat{Y}_j^n, \hat{\eta}_j^n)\) is specified in equation (6.47) in terms of the admissible controls \(\nu_j^{1,n}\) and \(\nu_j^{2,n}\). Because of the product-form definition of \(\nu_{j,\text{prod}}^n\), this conditional distribution can also be expressed as

\[
P_x\{(\hat{Y}_j^n, \hat{\eta}_j^n) \in dy \times dz | \hat{X}_0^n, \ldots, \hat{X}_j^n, \hat{U}_0^n, \ldots, \hat{U}_j^n \} = \nu_{j,\text{prod}}^n(dy \times dz | \hat{X}_0^n, \ldots, \hat{X}_j^n).
\]

The controlled process \((\hat{X}_n, \hat{U}^n)\), which is defined in terms of \(\hat{Y}_j^n\) and \(\hat{\eta}_j^n\), stands in relationship to the admissible control measure \(\nu_{\text{prod}}^n\) in the manner prescribed by Proposition 5.3.2 and Theorem 5.3.5. Hence we will be able to apply these results to study the limits of \(\{\hat{X}_n, n \in \mathbb{N}\}\) and \(\{\hat{U}^n, n \in \mathbb{N}\}\) after we show that the hypotheses of the proposition and the theorem are satisfied; namely, Condition 5.3.1 and the bound

\[
\sup_{n \in \mathbb{N}} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R\left(\nu_{j,\text{prod}}^n(\cdot)(\mu \times \rho_\sigma)(\cdot | \hat{X}_j^n)\right) \right\} < \infty. \tag{6.55}
\]

In order to verify Condition 5.3.1, we use the fact for \(\alpha_1^2\) and \(\alpha_2^2\) in \(\mathbb{R}^d\) the cumulant generating function of the product stochastic kernel \((\mu \times \rho_\sigma)(dy \times dz|x)\) is given by

\[
\log \int_{\mathbb{R}^d} \exp \left[ (\alpha_1^1, y) + (\alpha_2^2, z) \right] \mu(dy|x) \rho_\sigma(dz) = H(x, \alpha_1^1) + H_\sigma(\alpha_2^2).
\]

Since we have assumed Condition 6.2.1 on \(H(x, \alpha)\), it follows that for each \(\alpha_1^1\) and \(\alpha_2^2\) in \(\mathbb{R}^d\)

\[
\sup_{x \in \mathbb{R}^d} \left[ H(x, \alpha_1^1) + H_\sigma(\alpha_2^2) \right] < \infty, \tag{6.56}
\]

verifying Condition 5.3.1. The bound (6.55) is a consequence of the following display:

\[
\begin{align*}
\hat{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R\left(\nu_{j,\text{prod}}^n(\cdot)(\mu \times \rho_\sigma)(\cdot | \hat{X}_j^n)\right) \right\} \quad &\tag{6.57} \\
= \hat{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \left[ R\left(\nu_j^{1,n}(\cdot)(\mu)(\cdot | \hat{X}_j^n)\right) + R\left(\nu_j^{2,n}(\cdot)(\rho_\sigma)(\cdot | \hat{X}_j^n)\right) \right] \right\} \\
= \hat{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{nk-1} \left[ R\left(\nu_j^{1,n}(\cdot)(\mu)(\cdot | \hat{X}_j^n)\right) + R\left(\nu_j^{2,n}(\cdot)(\rho_\sigma)(\cdot | \hat{X}_j^n)\right) \right] \right\} \\
= \hat{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{nk-1} \left( L\left(\hat{X}_j^n, \beta_j^{1,n} \right) + \frac{1}{2\sigma} \left\| \beta_j^{2,n} \right\|^2 \right) \right\} \\
\leq \hat{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{nk-1} L_\sigma\left(\hat{X}_j^n, \beta_j^n \right) \right\} + \sigma \\
\leq \frac{1}{n} \sum_{j=0}^{n-1} \left( L_\sigma\left(\psi^*(j/n), \hat{\psi}^*(j/n) \right) + \sigma \left[ 1 + L_\sigma\left(\psi^*(j/n), \hat{\psi}^*(j/n) \right) \right] \right) + \sigma.
\end{align*}
\]

The first equality in this display is equation (6.38). By the definition of the control sequences \(\nu_j^{1,n}\) and \(\nu_j^{2,n}\), the terms in the running cost corresponding to \(j \in \{nk^n, nk^n+\}

$1, \ldots, n - 1$ all equal 0. This gives the second equality in the last display. The remainder of the display follows from formulas (6.49), (6.50), (6.53), and (6.51), the nonnegativity of $L_\sigma$, and the inequality $\kappa^n \leq 1$. Because $L_\sigma(\cdot, \cdot)$ is continuous on $\mathbb{R}^d \times \mathbb{R}^d$ and the sets \{||\psi^*(t)||, t \in [0, 1]\} and \{||\psi^*(t)||, t \in [0, 1]\} are bounded, the display yields (6.55).

We now proceed with the proof of the upper limit (6.43), which will lead to the Laplace principle lower bound in Proposition 6.6.1. Formula (6.56) and the bound (6.55) on the running costs associated with $\{\nu_{\text{prod}}^n\}$ allow us to apply Proposition 5.3.2 and Theorem 5.3.5 to study the limits of $\{\bar{X}^n\}$ and $\{\bar{U}^n\}$. Specifically, the proposition implies that the sequence $\{\nu_{\text{prod}}^n\}$ is tight and, in fact, has the uniform integrability property

$$\lim_{C \to \infty} \sup_{n \in \mathbb{N}} E_x \left\{ \int_{\{(y,z) \in \mathbb{R}^d \times [0,1] : ||(y,z)|| > C\}} ||(y,z)|| \nu_{\text{prod}}^n(dy \times dz \times dt) \right\} = 0. \tag{6.58}$$

Since $\nu_{j,\text{prod}}^n$ is product measure and $0 \leq \kappa^n \leq 1$, Theorem 5.3.5 implies that given any subsequence of $\{(\nu^{1,n}, \nu^{2,n}, \bar{X}^n, \bar{U}^n, \kappa^n), n \in \mathbb{N}\}$, there exists a further subsequence that satisfies

$$(\nu^{1,n}, \nu^{2,n}, \bar{X}^n, \bar{U}^n, \kappa^n) \xrightarrow{\mathcal{P}} (\nu^1, \nu^2, \bar{X}, \bar{U}, \kappa) \tag{6.59}$$

as $n \to \infty$. The limiting quantities $\nu^1$ and $\nu^2$ are stochastic kernels, $\bar{X}$ and $\bar{U}$ are processes taking values in $C([0,1] : \mathbb{R}^d)$, and $\kappa$ is a random variable taking values in $[0,1]$. The upper limit (6.43) will be proved for the subsequence of $n \in \mathbb{N}$ for which the convergence in (6.59) holds. An argument by contradiction applied to an arbitrary subsequence of $V_{\sigma}^n(x)$ will then give (6.43) for the entire sequence $\{V_{\sigma}^n(x), n \in \mathbb{N}\}$. By using the Skorohod Representation Theorem, we can assume that the convergence in (6.59) occurs w.p.1. All of the limit quantities $\bar{X}, \bar{U}, \nu^1, \nu^2, \kappa$ depend on $\sigma$, but this dependence will not be indicated in the notation.

With probability 1, Theorem 5.3.5 also implies that for every $t \in [0,1]$

$$\bar{X}(t) = x + \int_{\mathbb{R}^d \times [0,t]} y \nu^1(dy \times ds) \quad \text{and} \quad \bar{U}(t) = \int_{\mathbb{R}^d \times [0,t]} y \nu^2(dy \times ds).$$

Both of these processes take values in $C([0,1] : \mathbb{R}^d)$. With probability 1, we have for all $t \in [0,1]$

$$\bar{Z}(t) = \lim_{n \to \infty} \bar{Z}^n(t) = \lim_{n \to \infty} \bar{X}^n(t) + \lim_{n \to \infty} \bar{U}^n(t) = \bar{X}(t) + \bar{U}(t).$$

Hence, w.p.1 it follows that for all $t \in [0,1]$

$$\bar{Z}(t) = x + \int_{\mathbb{R}^d \times [0,t]} y \left[ \nu^1(dy \times ds) + \nu^2(dy \times ds) \right]. \tag{6.60}$$

A fact needed later in the proof is that $\sup_{t \in [0,1]} ||\bar{U}(t)|| \xrightarrow{\mathcal{P}} 0$ as $\sigma \to 0$. This is established in the next lemma.

**Lemma 6.6.3.** For any $\alpha > 0$

$$\lim_{\sigma \to 0} P_x \left\{ \sup_{t \in [0,1]} ||\bar{U}(t)|| \geq \alpha \right\} = 0.$$
Proof. We work with the subsequence of $n \in \mathbb{N}$ for which $(\nu_{1,n}, \nu_{2,n}, \bar{X}^n, \bar{U}^n, \kappa^n) \overset{\mathcal{D}}{\to} (\nu^1, \nu^2, \bar{X}, \bar{U}, \kappa)$. Defining processes $S^n = \{S^n(t), t \in [0, 1]\}$ by

$$S^n(t) = \int_{\mathbb{R}^d \times [0, t]} y \nu^{2n}(dy \times ds),$$

we claim that

$$\bar{E}_x \left\{ \left( \sup_{t \in [0, 1]} \| \bar{U}(t) \| \right)^2 \right\} \leq \liminf_{n \to \infty} \bar{E}_x \left\{ \left( \sup_{t \in [0, 1]} \| S^n(t) \| \right)^2 \right\}.$$

Indeed, (6.55) implies that

$$\sup_{n \in \mathbb{N}} \bar{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu_j^{2n}) \| \rho_\sigma \right\} < \infty.$$

Hence by Proposition 5.3.8 $S^n \overset{\mathcal{D}}{\to} \bar{U}$ as processes taking values in $\mathcal{C}([0, 1] : \mathbb{H}^d)$. The inequality (6.61) now follows from the Skorohod Representation Theorem and Fatou’s Lemma. For any $\alpha > 0$

$$P_x \left\{ \sup_{t \in [0, 1]} \| \bar{U}(t) \| \geq \alpha \right\}$$

$$\leq \frac{1}{\alpha^2} \bar{E}_x \left\{ \left( \sup_{t \in [0, 1]} \| \bar{U}(t) \| \right)^2 \right\}$$

$$\leq \frac{1}{\alpha^2} \liminf_{n \to \infty} \bar{E}_x \left\{ \left( \sup_{t \in [0, 1]} \| \int_{\mathbb{R}^d \times [0, t]} y \nu^{2n}(dy \times ds) \right)^2 \right\}$$

$$\leq \frac{1}{\alpha^2} \liminf_{n \to \infty} \bar{E}_x \left\{ \left( \frac{1}{n} \sum_{j=0}^{n-1} \left\| \int_{\mathbb{R}^d} y \nu_j^{2n}(dy) \right\|^2 \right) \right\}$$

$$\leq \frac{1}{\alpha^2} \liminf_{n \to \infty} \bar{E}_x \left\{ \left( \frac{1}{n} \sum_{j=0}^{n-1} \| \beta_j^{2n} \|^2 \right) \right\}$$

$$\leq \frac{2\sigma}{\alpha^2} \left( \liminf_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} \left( L_\sigma \left( \psi^*(j/n), \dot{\psi}^*(j/n) \right) + \sigma \left( 1 + L_\sigma \left( \psi^*(j/n), \dot{\psi}^*(j/n) \right) \right) \right) \right)$$

$$= \frac{2\sigma}{\alpha^2} \left( (1 + \sigma) \int_0^1 L_\sigma \left( \psi^*(t), \dot{\psi}^*(t) \right) dt + 2\sigma \right)$$

$$\leq \frac{2\sigma}{\alpha^2} \left( (1 + \sigma) (I_x(\psi) + \sigma) + 2\sigma \right).$$

The second line in this display follows from Chebyshev’s Inequality, the third line from (6.61), the fourth line from Jensen’s Inequality, the fifth line from the definition of $\nu_j^{2n}$, the sixth line from the last three lines of (6.57) and the nonnegativity of $L$, the seventh line from the continuity of $L_\sigma(\cdot, \cdot)$ on $\mathbb{R}^d \times \mathbb{H}^d$ [Lemma 6.6.2 (b)] and the Lebesgue Dominated
Convergence Theorem, and the eighth line from part (d) of Lemma 6.6.2 with \( \varepsilon \equiv \sigma \). It follows that

\[
\lim_{\sigma \to 0} P_x \left\{ \sup_{t \in [0,1]} \| U(t) \| \geq \alpha \right\} \leq \lim_{\sigma \to 0} \frac{2\sigma}{\alpha^2} \left( (1 + \sigma)(I_x(\psi) + \sigma) + 2\sigma \right) = 0,
\]

as claimed. \( \blacksquare \)

We now return to the proof of the upper limit

\[
\limsup_{n \to \infty} V_n(x) \leq \int_0^1 L_\sigma(\psi(t), \dot{\psi}(t)) dt + h(\psi) + \theta(\sigma), \tag{6.62}
\]

where \( \theta(\sigma) \to 0 \) as \( \sigma \to 0 \). As noted earlier, since \( L_\sigma(x, \beta) \leq L(x, \beta) \) for all \( x \) and \( \beta \) in \( \mathbb{R}^d \) [Lemma 6.6.2 (a)], this will yield the upper limit (6.35) for \( W_\sigma^n(x) = V_\sigma^n(x) \), completing the proof of the Laplace principle lower bound. In order to prove (6.62), we use formula (6.39), which is obtained by restricting the infimum in the definition of \( V_\sigma^n(x) \) to admissible control sequences of product form (6.37). Formula (6.39) implies that

\[
V_\sigma^n(x) \leq \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \left[ R\left( \nu^{1,n}_j(\cdot) \| \mu(\cdot | X^n_j) \right) + R\left( \nu^{2,n}_j(\cdot) \| \rho_\sigma(\cdot) \right) \right] + h(\mathcal{Z}^n) \right\} \tag{6.63}
\]

for the particular admissible control sequences \( \{\nu^{1,n}_j\} \) and \( \{\nu^{2,n}_j\} \) constructed after the proof of Lemma 6.6.2.

The following display involving the convergent subsequence \( (\nu^{1,n}, \nu^{2,n}, X^n, U^n, \kappa^n) \) yields a suitable upper bound on the running cost in (6.63):

\[
\limsup_{n \to \infty} \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \left[ R\left( \nu^{1,n}_j(\cdot) \| \mu(\cdot | X^n_j) \right) + R\left( \nu^{2,n}_j(\cdot) \| \rho_\sigma(\cdot) \right) \right] \right\}
\leq \limsup_{n \to \infty} \left( \frac{1}{n} \sum_{j=0}^{n-1} \left( L_\sigma(\psi^*(j/n), \dot{\psi}^*(j/n)) + \sigma \left[ 1 + L_\sigma(\psi^*(j/n), \dot{\psi}^*(j/n)) \right] \right) \right) + \sigma
\leq (1 + \sigma) \int_0^1 L_\sigma(\psi^*(t), \dot{\psi}^*(t)) dt + 2\sigma
\leq (1 + \sigma) \int_0^1 L_\sigma(\psi(t), \dot{\psi}(t)) dt + 2\sigma + \sigma(1 + \sigma).
\]

The second line follows from (6.57); the third line from the fact that \( \dot{\psi}^* \) is piecewise constant with only finitely many jumps in \([0,1]\), the continuity of \( L_\sigma(\cdot, \cdot) \) on \( \mathbb{R}^d \times \mathbb{R}^d \) [Lemma 6.6.2 (b)], and the Lebesgue Dominated Convergence Theorem; and the fourth line from part (d) of Lemma 6.6.2 with \( \varepsilon \equiv \sigma \).

The upper limit (6.62) is proved once we show that

\[
\limsup_{n \to \infty} \mathbb{E}_x \left\{ h(\mathcal{Z}^n) \right\} \leq h(\psi) + \bar{\theta}(\sigma),
\]
where \( \hat{\theta}(\sigma) \to 0 \) as \( \sigma \to 0 \). By part (d) of Lemma 6.6.2 with \( \varepsilon = \hat{\sigma} \), sup\( t \in [0,1] \| \psi^*(t) - \psi(t) \| \leq \sigma \). Since \( h \) is bounded and Lipschitz continuous, it suffices to prove that for any \( \alpha > 0 \)
\[
\lim_{\sigma \to 0} \limsup_{n \to \infty} P_x \left\{ \sup_{t \in [0,1]} \| \mathbf{Z}^n(t) - \psi^*(t) \| \geq \alpha \right\} = 0. 
\] (6.64)

By the properties (6.49) and (6.50) of the sequence of controls and by (6.52), for all \( j \in \{0,1,\ldots,n\kappa^n - 1\} \)
\[
\left\| \int_{\mathbb{R}^d} y \left[ \nu_j^1(dy) + \nu_j^2(dy) \right] - \dot{\psi}^*(j/n) \right\| 
= \left\| 2 \beta_j^1 + \beta_j^2 - \dot{\psi}^*(j/n) \right\| 
\leq K \left\| X_j - \psi^*(j/n) \right\| \left[ 1 + L_\sigma \left( \psi^*(j/n), \dot{\psi}^*(j/n) \right) \right].
\]

Pick any number \( t \in [0,\kappa] \). In the last display, we multiply through by \( 1/n \), sum over \( j \in \{0,1,\ldots,\lceil nt \rceil \wedge (n\kappa^n - 1)\} \), and send \( n \to \infty \); \( \lfloor nt \rfloor \) denotes the largest integer less than or equal to \( nt \). Then we use equation (6.60) for \( \mathbf{Z}(t) \). It follows that w.p.1, for all \( t \in [0,\kappa] \)
\[
\left\| \mathbf{Z}(t) - \psi^*(t) \right\| 
= \left\| \int_{\mathbb{R}^d \times [0,t]} y \left[ \nu^1(dy \times ds) + \nu^2(dy \times ds) \right] - \left( \psi^*(t) - x \right) \right\| 
\leq K \int_0^t \left\| \mathbf{X}(s) - \psi^*(s) \right\| \left[ 1 + L_\sigma \left( \psi^*(s), \dot{\psi}^*(s) \right) \right] ds 
\leq K \int_0^t \left\| \mathbf{X}(s) - \mathbf{Z}(s) \right\| \left[ 1 + L_\sigma \left( \psi^*(s), \dot{\psi}^*(s) \right) \right] ds 
+ K \int_0^t \left\| \mathbf{Z}(s) - \psi^*(s) \right\| \left[ 1 + L_\sigma \left( \psi^*(s), \dot{\psi}^*(s) \right) \right] ds 
\leq \hat{K} M \sup_{s \in [0,\kappa]} \left\| \mathbf{Z}(s) - \mathbf{X}(s) \right\| + \hat{K} M \int_0^t \left\| \mathbf{Z}(s) - \psi^*(s) \right\| ds,
\]

where
\[
M \doteq \sup_{s \in [0,1]} \left[ 1 + L_\sigma \left( \psi^*(s), \dot{\psi}^*(s) \right) \right].
\]

Define the constant \( \hat{K}' \doteq \hat{K} M e^{K M} \in (0,\infty) \). Since \( 0 \leq \kappa \leq 1 \), Gronwall’s Inequality [Theorem A.6.4] implies that w.p.1
\[
\sup_{s \in [0,\kappa]} \left\| \mathbf{Z}(s) - \psi^*(s) \right\| \leq \hat{K}' \sup_{s \in [0,\kappa]} \left\| \mathbf{Z}(s) - \mathbf{X}(s) \right\| = \hat{K}' \sup_{s \in [0,\kappa]} \left\| \mathbf{U}(s) \right\|.
\]

By Lemma 6.6.3, for any \( \alpha > 0 \)
\[
\lim_{\sigma \to 0} P_x \left\{ \sup_{t \in [0,\kappa]} \left\| \mathbf{Z}(t) - \psi^*(t) \right\| \geq \alpha \right\} \leq \lim_{\sigma \to 0} P_x \left\{ \sup_{t \in [0,\kappa]} \left\| \mathbf{U}(t) \right\| \geq \alpha / \hat{K}' \right\} = 0 \] (6.65)
and since w.p.1 $\tilde{X} = -\tilde{U} + \tilde{Z}$,

$$\lim_{\sigma \to 0} \tilde{P}_{x} \left\{ \sup_{t \in [0,\kappa]} \| \tilde{X}(t) - \psi^{*}(t) \| \geq \alpha \right\}$$

(6.66)

$$\leq \lim_{\sigma \to 0} \tilde{P}_{x} \left\{ \sup_{t \in [0,\kappa]} \| \tilde{U}(t) \| \geq \alpha/2 \right\} + \lim_{\sigma \to 0} \tilde{P}_{x} \left\{ \sup_{t \in [0,\kappa]} \| \tilde{Z}(t) - \psi^{*}(t) \| \geq \alpha/2 \right\} = 0.$$  

We write $\kappa = \kappa_{\sigma}$ for the probability–1 limit of the stopping times $\kappa^{n}$ defined in formula (6.48) in terms of $\eta$. We claim that

$$\lim_{\sigma \to 0} \tilde{P}_{x} \{ \kappa_{\sigma} < 1 \} = 0.$$  

(6.67)

Indeed, the definition of $\kappa^{n}$, the probability–1 convergence of $\tilde{X}^{n}$ to $\tilde{X}$ uniformly on $[0,1]$, and the continuity of $\tilde{X}$ and $\psi$ imply that on the event $\{ \kappa_{\sigma} < 1 \}$

$$\lim_{t \uparrow \kappa_{\sigma}} \| \tilde{X}(t) - \psi^{*}(t) \| \geq \eta.$$  

In addition, because of the probability–1 continuity of $\tilde{X}$, on this event we also have

$$\sup_{t \in [0,\kappa_{\sigma}]} \| \tilde{X}(t) - \psi^{*}(t) \| \geq \lim_{t \uparrow \kappa_{\sigma}} \| \tilde{X}(t) - \psi^{*}(t) \| \geq \eta.$$  

Hence if $\alpha \in (0, \eta/2)$, then by equation (6.66)

$$\lim_{\sigma \to 0} \tilde{P}_{x} \{ \kappa_{\sigma} < 1 \} \leq \lim_{\sigma \to 0} \tilde{P}_{x} \left\{ \sup_{t \in [0,\kappa_{\sigma}]} \| \tilde{X}(t) - \psi^{*}(t) \| \geq \alpha \right\} = 0.$$  

This proves the limit (6.67).

It now follows from equation (6.65) that

$$\lim_{\sigma \to 0} \tilde{P}_{x} \left\{ \sup_{t \in [0,\kappa_{\sigma}]} \| \tilde{Z}(t) - \psi^{*}(t) \| \geq \alpha \right\} = 0.$$  

Since w.p.1 $\tilde{Z}^{n} \rightarrow \tilde{Z}$ uniformly on $[0,1]$, this proves (6.64) and therefore the upper limit (6.62). We have finished the proof of Proposition 6.6.1, the Laplace principle lower bound under Conditions 6.2.1 and 6.3.2. The proof of the Laplace principle stated in Theorem 6.3.3 under these two conditions is now complete.

In the next section, we present an extension of Theorem 6.3.3 that will be applied in Chapter 10 in order to derive the Laplace principle for a class of continuous–time Markov processes with continuous statistics.

### 6.7 An Extension of Theorem 6.3.3 To Be Applied in Chapter 10

We need an extension of Theorem 6.3.3 that will be applied in Chapter 10 to a random walk model that is slightly different from the random walk model $\{ X^{n}, n \in \mathbb{N} \}$ considered
so far in the present chapter. The extension, stated in Theorem 6.7.5, will be the key step in the proof of the Laplace principle for a class of continuous-time Markov processes including diffusions and jump processes. The theorem will deal with sequences of processes that satisfy Laplace-type limits with the normalizing sequence \( n \in \mathbb{N} \) replaced by other sequences. We start by giving the relevant definition.

**Definition 6.7.1.** Let \( \{Y^n, n \in \mathbb{N}\} \) be a sequence of random variables taking values in a Polish space \( \mathcal{X} \), \( \{c_n, n \in \mathbb{N}\} \) a sequence of real numbers tending to \( \infty \), and \( I \) a rate function on \( \mathcal{X} \). The sequence \( \{Y^n\} \) is said to satisfy the **Laplace principle on \( \mathcal{X} \)** with rate function \( I \) and normalizing sequence \( \{c_n\} \) if for all bounded continuous functions \( h \) mapping \( \mathcal{X} \) into \( \mathbb{R} \)

\[
\lim_{n \to \infty} \frac{1}{c_n} \log E\{\exp[-c_n h(Y^n)]\} = - \inf_{x \in \mathcal{X}} \{h(x) + I(x)\}.
\]

Throughout this section we work with a fixed sequence \( \{c_n, n \in \mathbb{N}\} \) of positive integers that tend to \( \infty \). The random walk model \( \{X^n, n \in \mathbb{N}\} \) was defined in Section 4.3 in terms of a stochastic kernel \( \mu(dy|x) \) on \( \mathbb{R}^d \) given \( \mathbb{R}^d \). We now replace this stochastic kernel by a sequence of stochastic kernels \( \{\mu^n(dy|x), n \in \mathbb{N}\} \) on \( \mathbb{R}^d \) given \( \mathbb{R}^d \) satisfying the following property:

- There exists a stochastic kernel \( \mu(dy|x) \) on \( \mathbb{R}^d \) given \( \mathbb{R}^d \) such that for each \( x \in \mathbb{R}^d \)
  \( \mu^n(\cdot|x) \to \mu(\cdot|x) \) as \( n \to \infty \).

For each \( n \in \mathbb{N} \), let \( \{v^n_j(x), x \in \mathbb{R}^d, j \in \mathbb{N}_0\} \) be an i.i.d. sequence of random vector fields having the common distribution

\[
P\{v^n_j(x) \in dy\} \doteq \mu^n(dy|x).
\]

For \( x \in \mathbb{R}^d, n \in \mathbb{N} \), and \( j \in \{0, 1, \ldots, c_n - 1\} \), we define \( X^n_0 \doteq x \) and

\[
X^n_{j+1} = X^n_j + \frac{1}{c_n} v^n_j(X^n_j).
\]  \hfill (6.68)

We also define \( X^n = \{X^n(t), t \in [0, 1]\} \) by

\[
X^n(t) \doteq X^n_j + \left(t - \frac{j}{c_n}\right) v^n_j(X^n_j) \quad \text{for} \ t \in [j/c_n, (j+1)/c_n], \ j = 0, 1, \ldots, c_n - 1. \quad (6.69)
\]

Besides the appearance of \( \{c_n\} \) in place of the original sequence \( n \in \mathbb{N} \), these random quantities \( \{X^n_j\} \) and \( X^n \) differ from the analogous quantities in Section 4.3 only in one way: the random vector fields having the common distribution \( \mu(\cdot|x) \) have been replaced by the random vector fields having the common distribution \( \mu^n(\cdot|x) \). Since \( \mu^n(\cdot|x) \) may be viewed as a perturbation of \( \mu(\cdot|x) \), one would expect that under further conditions the Laplace principle for the new stochastic processes \( \{X^n, n \in \mathbb{N}\} \) defined in (6.68) and (6.69) would hold with the normalizing sequence \( \{c_n\} \) and with the same rate function \( I_x \) as in Theorem 6.3.3. This is in fact the content of Theorem 6.7.5 to be stated below.
For each \( n \in \mathbb{N} \) and \( x \) and \( \alpha \) in \( \mathbb{R}^d \), we define the cumulant generating function

\[
H^n(x, \alpha) = \log \int_{\mathbb{R}^d} \exp(\alpha \cdot y) \mu^n(dy|x).
\]

As in the earlier sections, \( H(x, \alpha) \) denotes the cumulant generating function

\[
H(x, \alpha) = \log \int_{\mathbb{R}^d} \exp(\alpha \cdot y) \mu(dy|x).
\]

In order to prove the Laplace principle upper bound, we need Condition 6.2.1 (as in Theorem 6.3.3) together with the following additional condition. Together these two conditions guarantee the existence of \( H(x, \alpha) \) and \( H^n(x, \alpha) \).

**Condition 6.7.2.**

(a) For each \( \alpha \in \mathbb{R}^d \)

\[
\sup_{n \in \mathbb{N}} \sup_{x \in \mathbb{R}^d} H^n(x, \alpha) < \infty.
\]

(b) For each \( \alpha \in \mathbb{R}^d \) and each compact subset \( K \) of \( \mathbb{R}^d \)

\[
\lim_{n \to \infty} \sup_{x \in K} |H^n(x, \alpha) - H(x, \alpha)| = 0.
\]

According to the next lemma, Conditions 6.2.1 and 6.7.2 imply that for each \( x \in \mathbb{R}^d \)

\[
\mu^n(\cdot|x) \Rightarrow \mu(\cdot|x) \text{ as } n \to \infty.
\]

This is the property of the stochastic kernels \( \{\mu^n(\cdot|x), n \in \mathbb{N}\} \) assumed at the beginning of this section.

**Lemma 6.7.3.** We assume Conditions 6.2.1 and 6.7.2. If \( \{x_n, n \in \mathbb{N}\} \) is any sequence in \( \mathbb{R}^d \) converging to some \( x \in \mathbb{R}^d \), then \( \mu^n(\cdot|x_n) \Rightarrow \mu(\cdot|x) \) as \( n \to \infty \). In particular, for each \( x \in \mathbb{R}^d \)

\[
\mu^n(\cdot|x) \Rightarrow \mu(\cdot|x).
\]

**Proof.** Condition 6.2.1 implies that \( H(x, \alpha) \) is continuous for \( (x, \alpha) \in \mathbb{R}^d \times \mathbb{R}^d \). Part (b) of Condition 6.7.2 implies that for each \( \alpha \in \mathbb{R}^d \)

\[
\lim_{n \to \infty} H^n(x_n, \alpha) = H(x, \alpha).
\]

Thus the sequence of moment generating functions of \( \{\mu^n(\cdot|x_n), n \in \mathbb{N}\} \), which is the sequence \( \{\exp[H^n(x_n, \alpha)]\} \), converges for each \( \alpha \in \mathbb{R}^d \) to the moment generating function \( \exp[H(x, \alpha)] \) of \( \mu(\cdot|x) \). Since the latter is a continuous function of \( \alpha \in \mathbb{R}^d \), the weak convergence \( \mu^n(\cdot|x_n) \Rightarrow \mu(\cdot|x) \) follows from part (b) of the continuity theorem for moment generating functions [Theorem A.3.20].

As in Theorem 6.3.3, the Laplace principle lower bound will follow from Condition 6.2.1 and from either Condition 6.3.2 and the following replacement for Condition 6.3.1, in which relative interior is replaced by interior. This is used to deal with the dependence of \( S_{\mu^n(\cdot|x)} \) upon \( n \).
6.7. AN EXTENSION OF THEOREM 6.3.3

Condition 6.7.4.
(a) The sets \( \text{int}(\text{conv } S_{\rho(x)}) \) are independent of \( x \in \mathbb{R}^d \).
(b) \( 0 \in \Sigma \triangleq \text{int}(\text{conv } S_{\rho(x)}) \).

We now state the Laplace principle for \( \{X^n\} \). For \( x \) and \( \beta \) in \( \mathbb{R}^d \) let \( L(x, \beta) \) denote the Legendre–Fenchel transform

\[
L(x, \beta) \doteq \sup_{\alpha \in \mathbb{R}^d} \{ \langle \alpha, \beta \rangle - H(x, \alpha) \}.
\]

(6.70)

Theorem 6.7.5. We assume Condition 6.7.2, Condition 6.2.1, and either Condition 6.7.4 or Condition 6.3.2. Let \( \{c_n, n \in \mathbb{N}\} \) be a sequence of positive integers that tend to \( \infty \). For any \( x \in \mathbb{R}^d \), consider the stochastic processes \( \{X^n, n \in \mathbb{N}\} \) defined in equations (6.68) and (6.69) and satisfying \( X^n(0) = x \). For absolutely continuous \( \varphi \in C([0,1] : \mathbb{R}^d) \) satisfying \( \varphi(0) = x \), we define

\[
I_x(\varphi) \doteq \int_0^1 L(\varphi(t), \dot{\varphi}(t)) dt,
\]

where \( L \) is defined in (6.70). For all other \( \varphi \in C([0,1] : \mathbb{R}^d) \) set \( I_x(\varphi) \doteq \infty \). Then the sequence \( \{X^n\} \) satisfies the Laplace principle on \( C([0,1] : \mathbb{R}^d) \) with rate function \( I_x \) and normalizing sequence \( \{c_n\} \). In fact, the Laplace principle holds uniformly on compacts.

As we have done earlier in this chapter, we will focus on the proof of the nonuniform Laplace principle. The uniform version can be proved with minor modifications. That \( I_x \)
has compact level sets in \( C([0,1] : \mathbb{R}^d) \) is proved in Proposition 6.2.4. The remainder of this section is devoted to the proofs of the Laplace principle bounds in Theorem 6.7.5. They will be proved by modifying the proofs of the corresponding bounds in Theorem 6.3.3. As we will see, the modifications are minor and the details of the proofs are straightforward. They are provided for completeness and can be skipped if desired. It is worthwhile to remark on the roles played by the two parts of Condition 6.7.2 in the proof of the Laplace principle. Exactly as in Section 6.2, part (a) is used in the proof of the Laplace principle upper bound to show the uniform integrability of sequences of controls. Part (b) is used in the proof of the Laplace principle lower bound to show that under Condition 6.3.1 the sequence \( \{L^n(x, \beta), n \in \mathbb{N}\} \) converges to \( L(x, \beta) \) uniformly for \( (x, \beta) \) in compact subsets of \( \mathbb{R}^d \times \Sigma \).

We will prove the Laplace principle bounds using a stochastic control representation formula. In the present setting an admissible control sequence is a sequence \( \{\nu^n_j, j = 0, 1, \ldots, c_n - 1\} \) in which each \( \nu^n_j \) is a stochastic kernel on \( \mathbb{R}^d \) given \( (\mathbb{R}^d)^{j+1} \). Associated with such a sequence is a sequence of controlled random vectors \( \{\tilde{X}^n_j, j = 0, 1, \ldots, c_n\} \), which are defined exactly like in Section 4.3 except that the index set is \( \{0, 1, \ldots, c_n\} \) instead of \( \{0, 1, \ldots, n\} \). Because of the presence of the new normalizing sequence \( \{c_n, n \in \mathbb{N}\} \), we must make minor changes in the definition of the admissible control measures. Let \( \{\nu^n_j, j = 0, 1, \ldots, c_n - 1\} \) be an admissible control sequence. For \( n \in \mathbb{N} \) and \( t \in [0,1] \) we define the stochastic kernel

\[
\nu^n(dy|t) \doteq \begin{cases} 
\nu^n_j(dy) & \text{if } t \in [j/c_n, (j + 1)/c_n), j = 0, 1, \ldots, c_n - 2 \\
\nu^n_{c_n-1}(dy) & \text{if } t \in [(c_n - 1)/c_n, 1],
\end{cases}
\]
where \( \nu^n_j(dy) = \nu^n_j(dy|\bar{X}_0^n, \bar{X}_1^n, \ldots, \bar{X}_j^n) \). We then define the admissible control measure \( \nu^n \) on \( \mathbb{R}^d \times [0, 1] \) by

\[
\nu^n(A \times B) = \int_B \nu^n(A|t) \, dt
\]

for Borel subsets \( A \) of \( \mathbb{R}^d \) and \( B \) of \( [0, 1] \).

With these modifications, the representations given in Theorem 4.3.1 and Corollary 5.2.1 remain valid for the new stochastic process \( X^n \) except that in the definition (5.2) [equivalently, (5.7)] of \( V^n(x) \) \( \mu^n(\cdot|x) \) replaces \( \mu(\cdot|x) \) and the normalizing sequence \( n \in \mathbb{N} \) is replaced by \( \{c_n\} \). In other words, for any bounded measurable function \( h \) mapping \( C([0, 1]; \mathbb{R}^d) \) into \( \mathbb{R} \) and for all \( n \in \mathbb{N} \) and \( x \in \mathbb{R}^d \)

\[
W^n(x) = -\frac{1}{c_n} \log E_x \{\exp(-c_n h(X^n))\}
\]

equals

\[
V^n(x) = \inf_{\{\nu^n_j\}} \bar{E}_x \left\{ \frac{1}{c_n} \sum_{j=0}^{c_n-1} R\left(\nu^n_j(\cdot)\|\mu^n(\cdot|\bar{X}_j^n)\right) + h(\bar{X}^n) \right\}
\]

\[
= \inf_{\{\nu^n_j\}} \bar{E}_x \left\{ \int_0^1 R\left(\nu^n(\cdot|t)\|\mu^n(\cdot|X^n(t))\right) \, dt + h(\bar{X}^n) \right\}.
\]

The infima are each taken over all admissible control sequences \( \{\nu^n_j\} \), and as the first argument of the relative entropy \( \nu^n_j(\cdot) = \nu^n_j(\cdot|\bar{X}_0^n, \bar{X}_1^n, \ldots, \bar{X}_j^n) \).

In the proof of Theorem 6.3.3, the convergence results stated in Theorem 5.3.5 played a crucial role. We want to carry these results over to the present situation. As a hypothesis, we have part (a) of Condition 6.7.2, which replaces part (a) of Condition 6.2.1. In place of the bound (5.12), we assume that

\[
\sup_{n \in \mathbb{N}} \bar{E}_x \left\{ \frac{1}{c_n} \sum_{j=0}^{c_n-1} R\left(\nu^n_j(\cdot)\|\mu^n(\cdot|\bar{X}_j^n)\right) \right\} < \infty.
\]

The proof of Theorem 5.3.5 and the proof of Proposition 5.3.2, on which Theorem 5.3.5 relies, may be easily adapted to the present situation with only notational changes. The bound in the last display will be satisfied by the admissible control sequences that arise in the proof of the Laplace principle. It follows that if the bound in the last display holds, then given any subsequence of \( \{(\nu^n, \bar{X}^n, \tilde{X}^n), n \in \mathbb{N}\} \), we can extract a further subsequence that satisfies \( (\nu^n, \bar{X}^n, \tilde{X}^n) \xrightarrow{D} (\nu, \bar{X}, \tilde{X}) \) as \( n \to \infty \). The limiting stochastic kernel \( \nu \) has the probability–1 decomposition \( \nu(dy \times dt) = \nu(dy|t) \otimes dt \). Furthermore, the limiting process \( \bar{X} = \{\tilde{X}(t), t \in [0, 1]\} \) is related to \( \nu \) as in part (b) of Theorem 5.3.5. As usual, the Skorohod Representation Theorem allows us to assume that the convergent subsequence \( (\nu^n, \bar{X}^n, \tilde{X}^n) \) converges to \( (\nu, \bar{X}, \tilde{X}) \) w.p.1.

The Laplace principle upper bound in Theorem 6.3.3 is proved in Section 6.2. The proof carries over to the present situation provided we can prove the string of inequalities...
in (6.10) when \( \mu(\cdot \mid \bar{X}^n(t)) \) in the second and third lines of this display is replaced by \( \mu^n(\cdot \mid \bar{X}^n(t)) \). This reduces to showing that

\[
\liminf_{n \to \infty} \mathbb{E}_x \left\{ \int_0^1 R \left( \nu^n(\cdot \mid t) \| \mu^n(\cdot \mid \bar{X}^n(t)) \right) dt \right\} \geq \mathbb{E}_x \left\{ \int_0^1 R \left( \nu(\cdot \mid t) \| \mu(\cdot \mid \bar{X}(t)) \right) dt \right\}, \tag{6.73}
\]

where \( \{ \nu^n, n \in \mathbb{N} \} \) is a sequence of admissible control measures having the form \( \nu^n(dy \times dt) = \nu^n(dy \mid t) \otimes dt \) and converging weakly w.p.1 to \( \nu(dy \times dt) = \nu(dy \mid t) \otimes dt \). In order to prove (6.73), we recall from Lemma 6.7.3 that if \( \{ x_n, n \in \mathbb{N} \} \) is any sequence in \( \mathbb{R}^d \) converging to \( x \in \mathbb{R}^d \), then \( \mu^n(\cdot \mid x_n) \rightharpoonup \mu(\cdot \mid x) \). Since w.p.1 the sequence \( \{ \bar{X}^n, n \in \mathbb{N} \} \) converges to \( \bar{X} \) uniformly on \([0, 1]\), it follows that there exists a null set on the complement of which \( \mu^n(dy \mid \bar{X}^n(t)) \rightharpoonup \mu(dy \mid \bar{X}(t)) \) for all \( t \in [0, 1] \). Theorem A.5.8 yields, on the complement of the same null set,

\[
\mu^n(dy \mid \bar{X}^n(t)) \otimes \lambda(dt) \rightharpoonup \mu(dy \mid \bar{X}(t)) \otimes \lambda(dt).
\]

As in the proof of (6.10), the lower limit (6.73) is now a consequence of part (f) of Lemma 1.4.3, the lower semicontinuity of \( R(\cdot \| \cdot) \), and Fatou’s Lemma. The proof of the Laplace principle upper bound in Theorem 6.7.5 is complete.

The Laplace principle lower bound in Theorem 6.3.3 is proved in Sections 6.4–6.6. Section 6.4 describes the strategy for the proof while the subsequent two sections complete it under Conditions 6.2.1 and 6.3.1 and Conditions 6.2.1 and 6.3.2, respectively. Let us first verify that under Conditions 6.6.2, 6.2.1, and 6.3.1, the calculations in Section 6.5 carry over to the present situation.

The main issue is to check the details of the Approximation Procedure 6.4.3. As in Section 6.5, we choose \( \mathcal{N} \) to be the class of functions \( \psi^* \in \mathcal{C}([0, 1] : \mathbb{R}^d) \) satisfying the following two conditions:

(a) \( \psi^*(t) \) is piecewise constant with only finitely many jumps in the interval \((0, 1)\).

(b) \( \psi^*(t) \in \Sigma \) for all \( t \in (0, 1) \).

Given \( \psi \in \mathcal{C}([0, 1] : \mathbb{R}^d) \) satisfying \( I_x(\psi) < \infty \) and any \( \delta > 0 \), Lemma 6.5.5 yields \( \psi^* \in \mathcal{N} \) satisfying \( \| \psi^* - \psi \|_\infty \leq \delta \) and \( I_x(\psi^*) \leq I_x(\psi) + \delta \). It remains to prove for any \( \psi^* \in \mathcal{N} \)

\[
\limsup_{n \to \infty} W^n(x) \leq I_x(\psi^*) + h(\psi^*), \tag{6.74}
\]

where \( W^n(x) \) is defined in equation (6.71). To prove this upper limit, we will use the facts proved in Lemma 6.5.2 that for each \( x \in \mathbb{R}^d \) the relative interior of the effective domain of \( L(x, \cdot) \) equals \( \Sigma \) and that \( L(x, \beta) \) is a continuous function of \((x, \beta) \in \mathbb{R}^d \times \Sigma \). By Condition 6.7.4 \( \Sigma \) is a nonempty open convex subset of \( \mathbb{R}^d \).

For \( n \in \mathbb{N} \) and \( x \) and \( \beta \) in \( \mathbb{R}^d \), consider

\[
L^n(x, \beta) = \sup_{\alpha \in \mathbb{R}^d} \| (\alpha, \beta) - H^n(x, \alpha) \|.
\]

We modify the construction given in Procedure 6.4.1 by replacing \( \mu(dy \mid x) \) with \( \mu^n(dy \mid x) \) and \( L(x, \beta) \) with \( L^n(x, \beta) \). With these modifications, we obtain as in that section an admissible control sequence \( \{ \nu^n_j, j = 0, 1, \ldots, c_n - 1 \} \) and controlled random vectors.
\( \{ \bar{X}^n_j, j = 0, 1, \ldots, c_n \} \), where \( \nu^n_j = \nu^n_j(\cdot | \xi) \) is a stochastic kernel on \( \mathbb{R}^d \) given \( (\mathbb{R}^d)^{j+1} \).

Fix \( \delta > 0 \). In terms of the stopping time
\[
\tau^n = \frac{1}{c_n} \left( \min \left\{ i \in \{0, 1, \ldots, c_n\} : \left\| \bar{X}^n_i - \psi^*(i/c_n) \right\| > \delta \right\} \land c_n \right),
\]
the admissible controls \( \nu^n_j(\cdot) = \nu^n_j(\cdot | \bar{X}^n_0, \bar{X}^n_1, \ldots, \bar{X}^n_j) \) have the following properties. For \( j \in \{0, 1, \ldots, c_n \tau^n - 1 \} \)
\[
R(\nu^n_j(\cdot) \| \mu^n(\cdot | \bar{X}^n_j)) = L^n(\bar{X}^n_j, \psi^*(j/c_n)) \quad \text{and} \quad \int_{\mathbb{R}^d} \nu^n_j(dy) = \psi^*(j/c_n).
\]
For \( j \in \{c_n \tau^n, c_n \tau^n + 1, \ldots, c_n - 1 \} \) \( \nu^n_j(\cdot) \) equals \( \mu^n(\cdot | \bar{X}^n_j) \) and thus \( R(\nu^n_j(\cdot) \| \mu^n(\cdot | \bar{X}^n_j)) \) equals 0.

The first item to check is that the sequence of running costs is bounded for all sufficiently large \( n \in \mathbb{N} \). As in formula (6.26), for each \( n \in \mathbb{N} \)
\[
\bar{E}_x \left\{ \frac{1}{c_n} \sum_{j=0}^{c_n \tau^n - 1} R(\nu^n_j(\cdot) \| \mu^n(\cdot | \bar{X}^n_j)) \right\} = \bar{E}_x \left\{ \frac{1}{c_n} \sum_{j=0}^{c_n \tau^n - 1} L^n(\bar{X}^n_j, \psi^*(j/c_n)) \right\}.
\]
We claim that the sequence \( \{L^n(x, \beta), n \in \mathbb{N}\} \) converges to \( L(x, \beta) \) uniformly for \( (x, \beta) \) in compact subsets of \( \mathbb{R}^d \times \Sigma \). Let us assume for a moment that this uniform convergence has been proved and finish the argument. \( \Sigma \) is a nonempty open subset of \( \mathbb{R}^d \) and \( L(x, \beta) \) is continuous on \( \mathbb{R}^d \times \Sigma \) [Lemma 6.5.2 (b)]. Furthermore, by definition of \( \mathcal{N} \psi^* \) is piecewise constant and \( \psi^*(t) \in \Sigma \) for all \( t \in (0, 1) \). Because of the definition of \( \tau^n \), it follows that the quantities \( L(\bar{X}^n_j, \psi^*(j/c_n)) \) are uniformly bounded for all sufficiently large \( n \in \mathbb{N} \) and all \( j \in \{0, 1, \ldots, c_n \tau^n - 1 \} \). The uniform convergence of \( L^n(x, \beta) \) to \( L(x, \beta) \) on compact subsets of \( \mathbb{R}^d \times \Sigma \) implies that the quantities \( L^n(\bar{X}^n_j, \psi^*(j/c_n)) \) are also uniformly bounded for all sufficiently large \( n \in \mathbb{N} \) and all \( j \in \{0, 1, \ldots, c_n \tau^n - 1 \} \). Except for the proof of the uniform convergence of \( L^n(x, \beta) \) to \( L(x, \beta) \), this completes the verification that the sequence of running costs is bounded for all sufficiently large \( n \in \mathbb{N} \).

As in Section 6.4, since \( \tau^n \in [0, 1] \), we can apply Theorem 5.3.5 and Proposition 5.3.8 to deduce convergence properties of subsequences of
\[
\{(\nu^n, \bar{X}^n, \mathcal{S}^n, \tau^n), n \in \mathbb{N}\}.
\]
Invoking the Skorohod Representation Theorem allows us to assume that any subsequence converging in distribution also converges w.p.1.

In order to prove that \( L^n(x, \beta) \to L(x, \beta) \) uniformly for \( (x, \beta) \) in compact subsets of \( \mathbb{R}^d \times \Sigma \), it suffices to show that if \( \{(\xi_n, \beta_n), n \in \mathbb{N}\} \) is any sequence in \( \mathbb{R}^d \times \Sigma \) converging to \( (\xi, \beta) \in \mathbb{R}^d \times \Sigma \), then
\[
\lim_{n \to \infty} L^n(\xi_n, \beta_n) = L(\xi, \beta).
\]
To carry this out, we define \( \alpha \in \mathbb{R}^d g_n(\alpha) = H^n(\xi_n, \alpha) \) and \( g(\alpha) = H(\xi, \alpha) \). These functions are all differentiable in \( \alpha \), and by part (b) of Condition 6.7.2, for any \( \alpha \in \mathbb{R}^d \)
\[
\lim_{n \to \infty} g_n(\alpha) = \lim_{n \to \infty} H^n(\xi_n, \alpha) = H(\xi, \alpha) = g(\alpha).
\]
Since the Legendre–Fenchel transforms \( g^*_n \) and \( g^* \) equal \( L^n(\xi_n, \cdot) \) and \( L(\xi, \cdot) \), respectively, the limit \( L^n(\xi_n, \beta_n) \to L(\xi, \beta) \) is a consequence of Lemma C.8.1.

The second item to be checked is that the display (6.27) can be adapted to the present situation. The replacement is the following:

\[
\limsup_{n \to \infty} W^n(x) = \limsup_{n \to \infty} V^n(x) \tag{6.75}
\]

\[
\leq \lim_{n \to \infty} E_x \left\{ \frac{1}{c_n} \sum_{j=0}^{c_n-1} R(\nu^n_j(\cdot)) \| \mu^n(\cdot | X^n_j) \| + h(\bar{X}^n) \right\}
\]

\[
= \lim_{n \to \infty} E_x \left\{ \frac{1}{c_n} \sum_{j=0}^{c_n-1} L^n(\bar{X}^n(j/c_n), \tilde{\psi}^*(j/c_n)) + h(\bar{X}^n) \right\}
\]

\[
= \int_0^1 L(\psi^*(t), \dot{\psi}^*(t)) \, dt + h(\psi^*)
\]

\[
= I_x(\psi^*) + h(\psi^*).
\]

The measures \( \mu(\cdot | X^n_j) \) in line two of (6.27) have been replaced by \( \mu^n(\cdot | X^n_j) \) because of the new definition of \( V^n(x) \) in (6.72). This leads to the replacement of \( L(\bar{X}^n(j/n), \tilde{\psi}^*(j/n)) \) in line three by \( L^n(\bar{X}^n(j/c_n), \tilde{\psi}^*(j/c_n)) \). In order to obtain line four of the display, we use the uniform convergence of \( L^n(x, \beta) \) to \( L(x, \beta) \) for \( (x, \beta) \) in compact subsets of \( \mathbb{R}^d \times \Sigma \), together with the facts, proved exactly as in Lemma 6.4.2, that w.p.1 \( \tau^n \to 1 \) and \( \bar{X}(t) = \psi^*(t) \) for all \( t \in [0, 1] \). The display (6.75) now yields the upper limit (6.74). This completes the proof of the Laplace principle lower bound in Theorem 6.7.5 under Conditions 6.7.2, 6.2.1, and 6.3.1.

Finally, we consider the Laplace principle lower bound under Conditions 6.7.2, 6.2.1, and 6.3.2, following the line of argument in Section 6.6. The main step was the verification of the upper limit (6.35), which states that

\[
\limsup_{n \to \infty} W^n(x) \leq I_x(\psi) + h(\psi) + \theta(\sigma) = \int_0^1 L(\psi(t), \dot{\psi}(t)) \, dt + h(\psi) + \theta(\sigma),
\]

where \( \theta(\sigma) \to 0 \) as \( \sigma \to 0 \). The proof of that limit in Section 6.6 was carried out via a perturbation argument that reduced the situation to the case considered in the previous section, Section 6.5, where Conditions 6.2.1 and 6.3.1 were assumed. We have just checked that the calculations in Section 6.5 carry over when Conditions 6.7.2, 6.2.1, and 6.3.1 are assumed. Using a virtually identical argument, we can prove that the analogue of the upper limit (6.35) in Section 6.6 continues to hold under the present conditions; namely, under Conditions 6.7.2, 6.2.1, and 6.3.2. This gives the Laplace principle lower bound in Theorem 6.7.5 under these three conditions. The proof of the theorem is complete.  

In the next chapter, we prove the Laplace principle for the random walk model in Section 4.3 with discontinuous statistics.
Chapter 7

Laplace Principle for the Random Walk Model with Discontinuous Statistics

7.1 Introduction

In the previous chapter, we proved the Laplace principle for the random walks $X^n$ defined in Section 4.3, assuming that the measures $\mu(dy|x)$ are continuous in $x$. Because of this continuity property of $\mu(dy|x)$, the random walk is said to have continuous statistics. In a variety of applications random processes arise naturally for which the assumption of continuous statistics is violated.

One such example is random motion in discontinuous media. This motion may be modeled by a Markov chain with state space $\mathbb{R}^d$ partitioned into finitely many sets $\{\Lambda^{(i)}, i = 1,\ldots,N\}$, in each of which the Markov chain has a different smooth transition mechanism. The transition mechanisms change discontinuously when the chain crosses the boundaries separating the sets $\Lambda^{(i)}$. The present chapter analyzes such a Markov chain with discontinuous statistics in the case $N = 2$. The simplest case of this is treated in [39]. There the transition mechanisms of the Markov chain are defined in terms of a given probability measure $\mu^{(1)}$ when the random walk is in the left halfspace $\Lambda^{(1)} = \{x \in \mathbb{R}^d : x_1 \leq 0\}$ and in terms of a different probability measure $\mu^{(2)}$ when the random walk is in the right halfspace $\Lambda^{(2)} = \{x \in \mathbb{R}^d : x_1 > 0\}$. Since $\mu^{(1)} \neq \mu^{(2)}$, the discontinuity occurs across the $(d-1)$-dimensional hyperplane $\partial = \{x \in \mathbb{R}^d : x_1 = 0\}$. When suitably scaled, the resulting Markov chain is given by the random vectors $X^n_j$ defined in equation (4.4), in which the distributions of the i.i.d. random vector fields $v_j(x)$ are $\mu^{(1)}$ for $x \in \Lambda^{(1)}$ and $\mu^{(2)}$ for $x \in \Lambda^{(2)}$. The main result in [39] is the large deviation principle for the sequence $\{X^n, n \in \mathbb{N}\}$, which represent the positions of the random walk at time $n$.

In the present chapter, we extend significantly the results of [39] by proving a process-level Laplace principle for the sequence of processes $\{X^n, n \in \mathbb{N}\}$ which are the piecewise linear interpolations of the random vectors $X^n_j$. We also allow the transition mechanisms inside each of the halfspaces to have a smooth spatial dependence. Although we
continue to deal with the case of two halfspaces \( \Lambda^{(1)} \) and \( \Lambda^{(2)} \) of smooth statistical behavior separated by the hyperplane \( \partial \), the problem where \( \partial \) is replaced by a smooth \((d - 1)\)-dimensional manifold may be reduced to the case considered here using standard techniques [2].

As we describe in Section 7.2, the random walk model with discontinuous statistics is defined by two stochastic kernels \( \mu^{(1)}(dy|x) \) and \( \mu^{(2)}(dy|x) \), which govern the transitions in the respective halfspaces \( \Lambda^{(1)} \) and \( \Lambda^{(2)} \). We assume that these measures are appropriately bounded and that for each \( i = 1, 2 \) the functions mapping \( x \in \mathbb{R}^d \rightarrow \mu^{(i)}(dy|x) \in \mathcal{P}(\mathbb{R}^d) \) are continuous [Condition 7.2.1]. After defining the model in Section 7.2, we formulate the Laplace principle in Theorem 7.2.3. The rate function is given by an explicit formula that is influenced strongly by the presence of the boundary \( \partial \). Through equation (7.7) the form of the rate function incorporates a necessary and sufficient condition for the stability of the one-component of a class of controlled random walks that will arise in the proof of the Laplace principle. In Section 7.3, under the assumption that \( \mu^{(1)}(dy|x) \) and \( \mu^{(2)}(dy|x) \) are both independent of \( x \), we use the contraction principle to derive the Laplace principle for the sequence of final position vectors of the random walk and give an explicit formula for the rate function. We then illustrate the form of the rate function by means of some one-dimensional examples. In Section 7.4 we prove the Laplace principle upper bound. The lower bound is proved in Section 7.5 and the compactness of the level sets of the rate function in Section 7.6. The results in the present chapter were given in preliminary form in [40]. For a special case of the processes treated here, the large deviation principle is derived in [10]. The papers [63, 64] prove the large deviation principle for a restricted class of diffusion processes with discontinuous drifts, using continuous mapping techniques that are found in the literature. The large deviation principle for a jump Markov process analogous to our random walk is proved in [1] under a uniform absolute continuity condition on the jump measures in the two halfspaces.

A more ambitious problem than the one analyzed in the present chapter concerns Markov processes in \( \mathbb{R}^d \) having three or more regions of smooth statistical behavior separated by an arbitrary number of intersecting smooth \((d - 1)\)-dimensional manifolds across which the statistical behavior may change discontinuously. Aside from some interesting special cases, giving an explicit formula for the rate function along intersections of the manifolds of codimension 2 or higher involves the extremely difficult problem of determining necessary and sufficient conditions for stability that are analogous to formula (7.7) and are valid for this more complicated geometry. Prototypes for such processes are a class of jump Markov processes modeling queueing systems. These processes are widely used in the study of computer networks and communication networks as well as in other important applications. To date, [41] is the only result in the literature that proves the large deviation principle for such multiboundary problems in a general setting. The paper uses continuous-time analogues of the representation formulas introduced in this book.

The large deviation analysis of Markov processes with discontinuous statistics is an area rich in mathematical problems and in potential applications. The methods that are developed in the present chapter are sophisticated enough to handle other challenging large deviation problems such as the diffusion analogues of the models considered in this chapter. We now turn to the random walk model, which must be thoroughly understood
before the analysis of more complicated examples can be undertaken.

7.2 Statement of the Laplace Principle

We use the notation of Section 4.3. The reader will recall that the random walk model is defined in terms of a stochastic kernel \( \mu(dy|x) \) on \( \mathbb{R}^d \) given \( \mathbb{R}^d \). This stochastic kernel gives the distributions of the i.i.d. random vector fields \( v_j(x) \) by the formula \( P\{v_j(x) \in dy\} = \mu(dy|x) \). For \( x \in \mathbb{R}^d, n \in \mathbb{N}, \) and \( j \in \{0,1,\ldots,n-1\} \) we define random vectors \( X^n_0 = x \) and

\[
X^n_{j+1} = X^n_j + \frac{1}{n} v_j(X^n_j)
\]

and the random process

\[
X^n(t) \doteq X^n_j + \left( t - \frac{j}{n} \right) v_j(X^n_j) \quad \text{for} \quad t \in [j/n, (j+1)/n], \quad j = 0,1,\ldots,n-1.
\]

This process is the piecewise linear interpolation of the random vectors \( \{X^n_j, j = 0,1,\ldots,n\} \). The purpose of the present section is to formulate the Laplace principle for the sequence \( \{X^n, n \in \mathbb{N}\} \). The new ingredient, absent in the previous chapter, is to allow the stochastic kernel \( \mu(dy|x) \) to depend discontinuously upon the parameter \( x \) as \( x \) crosses a \((d-1)\)-dimensional boundary in \( \mathbb{R}^d \).

We define

\[
\Lambda^{(1)} \doteq \left\{ x \in \mathbb{R}^d : x_1 \leq 0 \right\}, \quad \Lambda^{(2)} \doteq \left\{ x \in \mathbb{R}^d : x_1 > 0 \right\}, \quad \partial \doteq \left\{ x \in \mathbb{R}^d : x_1 = 0 \right\}.
\]

An asymmetry in the definitions and notation arises from our having included the hyperplane \( \partial \) together with the open left halfspace, giving the closed left halfspace \( \Lambda^{(1)} \). However, under the assumptions of the Laplace principle formulated in Theorem 7.2.3, the large deviation behavior is insensitive to this choice. Indeed, the same Laplace principle would hold if the hyperplane \( \partial \) were included with the open right halfspace.

Let \( \mu^{(1)}(dy|x) \) and \( \mu^{(2)}(dy|x) \) be stochastic kernels on \( \mathbb{R}^d \) given \( \mathbb{R}^d \) having finite moment generating functions for each \( x \in \mathbb{R}^d \). In terms of them we define

\[
\mu(dy|x) \doteq \begin{cases} 
\mu^{(1)}(dy|x) & \text{if } x \in \Lambda^{(1)} \\
\mu^{(2)}(dy|x) & \text{if } x \in \Lambda^{(2)}.
\end{cases}
\]

For \( i = 1,2 \) and \( x \) and \( \alpha \) in \( \mathbb{R}^d \), we then define the cumulant generating functions

\[
H^{(i)}(x,\alpha) \doteq \log \int_{\mathbb{R}^d} \exp(\alpha,y) \mu^{(i)}(dy|x)
\]

and, for \( \beta \in \mathbb{R}^d \), the Legendre–Fenchel transforms

\[
L^{(i)}(x,\beta) \doteq \sup_{\alpha \in \mathbb{R}^d} \left\{ \langle \alpha,\beta \rangle - H^{(i)}(x,\alpha) \right\}, \quad i = 1,2.
\]

We also set \( H(x,\alpha) \) equal to \( H^{(1)}(x,\alpha) \) if \( x \in \Lambda^{(1)} \) and equal to \( H^{(2)}(x,\alpha) \) if \( x \in \Lambda^{(2)} \).
The next condition will be needed in the proofs of both the Laplace principle upper and lower bounds. Part (a) of the condition gives Condition 5.3.1, which is needed in order to apply the convergence results stated in Theorem 5.3.5. Although in part (b) we assume that both \( \mu^{(1)}(dy|x) \) and \( \mu^{(2)}(dy|x) \) are continuous functions of \( x \in \mathbb{R}^d \), in general \( \mu(dy|x) \) is not continuous at \( \partial \).

**Condition 7.2.1.**

(a) For each \( \alpha \in \mathbb{R}^d \), \( \max_{i=1,2} \left( \sup_{x \in \mathbb{R}^d} H^{(i)}(x, \alpha) \right) < \infty \).

(b) For each \( i = 1, 2 \), the function mapping \( x \in \mathbb{R}^d \) \( \longrightarrow \mu^{(i)}(dy|x) \in \mathcal{P}(\mathbb{R}^d) \) is continuous in the topology of weak convergence on \( \mathcal{P}(\mathbb{R}^d) \).

With the definition (7.3) of \( \mu(dy|x) \), the sequence \( \{X^n_j, j = 0, 1, \ldots, n\} \) forms a Markov chain with transition probability function

\[
P\{X_{j+1}^n \in dy | X_j^n = x\} = 1_{\Lambda^{(1)}}(x) \mu^{(1)}(n(dy - x)|x) + 1_{\Lambda^{(2)}}(x) \mu^{(2)}(n(dy - x)|x).
\]

This transition probability function satisfies none of the usual absolute continuity, irreducibility, or smoothness conditions assumed in other large deviation studies. As a consequence, one is not able to analyze the random walks \( X^n \) by any of the usual methods.

The following assumption on the supports of \( \mu^{(i)}(dy|x) \) will be needed to prove the Laplace principle lower bound. This assumption is reminiscent of Condition 6.3.1, which was needed to prove the lower bound for the random walk model with continuous statistics. The first half of part (b) is a technical hypothesis that will be used in the proof of Lemma 7.5.4. Condition 7.2.2 is satisfied if, for example, the support of \( \mu^{(i)}(dy|x) \) is all of \( \mathbb{R}^d \) for each \( x \in \mathbb{R}^d \). Instead of Condition 7.2.2, we could also introduce an alternative smoothness assumption, analogous to Condition 6.3.2, under which the lower bound could be proved. However, we will omit this.

**Condition 7.2.2.**

(a) The sets \( \text{ri}(\text{conv} \ S_{\mu^{(i)}(\cdot|x)}) \) are independent of \( x \in \mathbb{R}^d \) and \( i = 1, 2 \).

(b) \( 0 \in \Sigma = \text{ri}(\text{conv} \ S_{\mu^{(i)}(\cdot|x)}) \) and \( \Sigma \) is not a subset of \( \partial = \{x \in \mathbb{R}^d : x_1 = 0\} \).

Part (b) implies that \( \Sigma \cap \text{int} \Lambda^{(i)} \neq \emptyset \) for \( i = 1, 2 \).

Let \( \mu^{(1)} \) and \( \mu^{(2)} \) be probability measures on \( \mathbb{R}^d \) having finite moment generating functions. In [39] we proved the large deviation principle for the sequence \( \{X^n_n, n \in \mathbb{N}\} \), assuming that

\[
\mu(dy|x) = \begin{cases} 
\mu^{(1)}(dy) & \text{if } x \in \Lambda^{(1)} \\
\mu^{(2)}(dy) & \text{if } x \in \Lambda^{(2)} 
\end{cases}
\]

and that the supports \( S_{\mu^{(1)}} \) and \( S_{\mu^{(2)}} \) satisfy

\[
S_{\mu^{(i)}} \cap \text{int} \Lambda^{(3-i)} \neq \emptyset \quad (7.4)
\]

for \( i = 1, 2 \). In general \( S_{\mu^{(1)}} \) and \( S_{\mu^{(2)}} \) need not be equal. The sequence \( \{X^n_n\} \) is the sequence of final position vectors of the random walk at time 1. Condition (7.4) allows
7.2. **STATEMENT OF THE LAPLACE PRINCIPLE**

the interiors of the two half-spaces \( \Lambda^{(1)} \) and \( \Lambda^{(2)} \) to communicate in the sense that each of the two transition mechanisms of the random walk in the two halfspaces has a positive probability of being activated for arbitrarily large times. Thus the condition is natural. In the present chapter we prove the Laplace principle for the sequence \( \{X^n\} \), assuming for simplicity that the sets \( ri(\text{conv } S_{\mu^0}(x)) \) are independent of \( x \in \mathbb{R}^d \) and \( i = 1, 2 \). With some additional work this hypothesis can be significantly weakened.

The rate function for the random walk model with discontinuous statistics has an explicit but complicated form. In order to specify it, we need some additional notation. For \( x \) and \( \beta \) in \( \mathbb{R}^d \) we define

\[
L^{(0)}(x, \beta) \doteq \inf \left\{ \rho^{(1)} L^{(1)}(x, \beta^{(1)}) + \rho^{(2)} L^{(2)}(x, \beta^{(2)}) \right\},
\]

where the infimum is taken over all \( \rho^{(1)} \in \mathbb{R}, \rho^{(2)} \in \mathbb{R}, \beta^{(1)} \in \mathbb{R}^d \), and \( \beta^{(2)} \in \mathbb{R}^d \) satisfying

\[
\rho^{(1)} \geq 0, \rho^{(2)} \geq 0, \rho^{(1)} + \rho^{(2)} = 1,
\]

\[
(\beta^{(1)})_1 \geq 0, (\beta^{(2)})_1 \leq 0,
\]

\[
\rho^{(1)} \beta^{(1)} + \rho^{(2)} \beta^{(2)} = \beta.
\]

The subscripts 1 denote the one–components of the vectors. We then define for \( x \) and \( \beta \) in \( \mathbb{R}^d \)

\[
\tilde{L}(x, \beta) \doteq \begin{cases} 
L^{(1)}(x, \beta) & \text{if } x_1 < 0 \\
L^{(0)}(x, \beta) & \text{if } x_1 = 0 \\
L^{(2)}(x, \beta) & \text{if } x_1 > 0.
\end{cases}
\]

We would like to illuminate the complicated definition of \( \tilde{L}(x, \beta) \) by commenting first on equation (7.7). This expresses a necessary and sufficient condition for the stability of the one–component of a class of controlled random walks. In order to see the role played by this equation, let us consider, for example, the problem of estimating the probability that \( \sup_{t \in [0,1]} \|X^n(t) - \beta t\| < \varepsilon \), where \( \varepsilon > 0 \) and \( \beta \) is a vector in \( \mathbb{R}^d \) satisfying \( \beta_1 = 0 \). When using the representation formula, one must consider controls that shift the drifts of the processes in the halfspaces \( \Lambda^{(1)} \) and \( \Lambda^{(2)} \) to \( \beta^{(1)} \) and \( \beta^{(2)} \), respectively, in such a way that the event of interest has a probability of order 1 as \( n \to \infty \). Unless the constraints in formula (7.7) are satisfied, the probability of the event is of order 0 as \( n \to \infty \). It should be stressed that these constraints do not express a stability condition on the original processes but rather on the controlled processes appearing in the representation formula.

We now state the Laplace principle.

**Theorem 7.2.3.** We assume Conditions 7.2.1 and 7.2.2. For any \( x \in \mathbb{R}^d \), consider the piecewise–linearly interpolated processes \( \{X^n, n \in \mathbb{N}\} \) defined in equations (7.1) and (7.2) and satisfying \( X^n(0) = x \). For absolutely continuous functions \( \varphi \in C([0,1] : \mathbb{R}^d) \) satisfying \( \varphi(0) = x \), we define

\[
I_x(\varphi) \doteq \int_0^1 \tilde{L}(\varphi(t), \dot{\varphi}(t)) \, dt,
\]

where \( \tilde{L} \) is defined in equations (7.5)–(7.9). For all other \( \varphi \in C([0,1] : \mathbb{R}^d) \), we set \( I_x(\varphi) \doteq \infty \). Then the sequence \( \{X^n\} \) satisfies the Laplace principle on \( C([0,1] : \mathbb{R}^d) \) with rate function \( I_x \). In fact, the Laplace principle holds uniformly on compacts.
We will focus on the proof of the nonuniform Laplace principle, omitting, as in Chapter 6, the routine modifications needed to prove the uniform version. The Laplace principle upper bound and the fact that $I_x$ has compact level sets will be proved in Section 7.4 under Condition 7.2.1. The lower bound will be proved in Section 7.5 under Conditions 7.2.1 and 7.2.2.

Before turning to these technical proofs, in the next section we use the contraction principle to derive from Theorem 7.2.3 the Laplace principle for the sequence $\{X^n_n, n \in \mathbb{N}\}$ of final position vectors of the random walk under Conditions 7.2.1 and 7.2.2. When $\mu^{(1)}(dy|x)$ and $\mu^{(2)}(dy|x)$ are each independent of $x \in \mathbb{R}^d$, we evaluate explicitly the contracted rate function, giving a new proof of the main result, Theorem 2.1, in [39]. We end the section by giving some one-dimensional examples of rate functions for the sequence of final position vectors of the random walk.

### 7.3 Laplace Principle for the Final Position Vectors and One-Dimensional Examples

Let $f$ denote the continuous function mapping $C([0,1] : \mathbb{R}^d)$ into $\mathbb{R}^d$ which is given by $f(\varphi) \equiv \varphi(1)$. Since $f(X^n) = X^n_n$, Theorem 7.2.3 and the contraction principle [Theorem 1.3.2] yield the Laplace principle for the sequence $\{X^n_n, n \in \mathbb{N}\}$ with rate function

$$\Upsilon_x(\beta) \equiv \inf\{I_x(\varphi) : \varphi \in f^{-1}(\beta)\}.$$

In order to have an intuitive feel for the form of the rate function in Theorem 7.2.3, it is useful to understand the large deviation properties of an increment of the process. If the time interval is small, then up to a linear time change, such an increment is well approximated by the final position vector $X^n_n$ of a model for which the stochastic kernels $\mu^{(1)}(dy|x)$ and $\mu^{(2)}(dy|x)$ are each independent of $x \in \mathbb{R}^d$. In this case the infimization in the last display can be carried out explicitly. We will give the formula assuming that $X^n_n(0) = 0$. The case where $X^n_n(0) = x$, $x \neq 0$, can be handled with minor modifications.

Let $\mu^{(1)}$ and $\mu^{(2)}$ be probability measures on $\mathbb{R}^d$ having finite moment generating functions and consider the random walk model defined in terms of

$$\mu(dy|x) \equiv \begin{cases} 
\mu^{(1)}(dy) & \text{if } x \in \Lambda^{(1)} \\
\mu^{(2)}(dy) & \text{if } x \in \Lambda^{(2)}.
\end{cases}$$

For $i = 1, 2$ and $\alpha$ and $\beta$ in $\mathbb{R}^d$ we define

$$H^{(i)}(\alpha) \equiv \log \int_{\mathbb{R}^d} \exp\langle \alpha, y \rangle \mu^{(i)}(dy) \quad \text{and} \quad L^{(i)}(\beta) \equiv \sup_{\alpha \in \mathbb{R}^d} \{\langle \alpha, \beta \rangle - H^{(i)}(\alpha)\}.$$

We make the following assumption, which corresponds to part (a) of Condition 7.2.1 and to Condition 7.2.2. In the present setting part (b) of Condition 7.2.1 is automatic.

**Condition 7.3.1.**

(a) For $i = 1, 2$ and each $\alpha \in \mathbb{R}^d$, $H^{(i)}(\alpha) < \infty$.

(b) The set $\text{ri}(\text{conv } S_{\mu^{(i)}})$ equals $\text{ri}(\text{conv } S_{\mu^{(i)}})$.

(c) $0 \in \Sigma \equiv \text{ri}(\text{conv } S_{\mu^{(i)}})$ and $\Sigma$ is not a subset of $\partial \equiv \{x \in \mathbb{R}^d : x_1 = 0\}$.


7.3. ANOTHER LAPLACE PRINCIPLE AND EXAMPLES

We must introduce several other functions. As in formula (7.5) define for \( \beta \in \mathbb{R}^d \)

\[
L^{(0)}(\beta) \doteq \inf \left\{ \rho^{(1)} L^{(1)}(\beta^{(1)}) + \rho^{(2)} L^{(2)}(\beta^{(2)}) \right\},
\]

where the infimum is taken over all \( \rho^{(1)},\rho^{(2)},\beta^{(1)}, \) and \( \beta^{(2)} \) satisfying

\[
\rho^{(1)} \geq 0, \rho^{(2)} \geq 0, \rho^{(1)} + \rho^{(2)} = 1,
\]

\[
(\beta^{(1)})_1 \geq 0, (\beta^{(2)})_1 \leq 0, \rho^{(1)} \beta^{(1)} + \rho^{(2)} \beta^{(2)} = \beta.
\]

For \( i = 1,2 \) and \( \beta \in \mathbb{R}^d \) we also define

\[
\tilde{L}^{(i)}(\beta) \doteq \inf \left\{ \gamma^{(0)} L^{(0)}(\beta^{(0)}) + \gamma^{(i)} L^{(i)}(\beta^{(i)}) \right\},
\]

where the infimum is taken over all \( \gamma^{(0)} \in \mathbb{R}, \gamma^{(i)} \in \mathbb{R}, \beta^{(0)} \in \mathbb{R}^d, \beta^{(i)} \in \mathbb{R}^d \) satisfying

\[
\gamma^{(0)} \geq 0, \gamma^{(i)} \geq 0, \gamma^{(0)} + \gamma^{(i)} = 1,
\]

\[
(\beta^{(0)})_1 = 0, \gamma^{(0)} \beta^{(0)} + \gamma^{(i)} \beta^{(i)} = \beta.
\]

Finally, we define

\[
M(\beta) \doteq \begin{cases} 
L^{(0)}(\beta) & \text{if } \beta_1 = 0 \\
\tilde{L}^{(1)}(\beta) & \text{if } \beta_1 < 0 \\
\tilde{L}^{(2)}(\beta) & \text{if } \beta_1 > 0.
\end{cases}
\]  

The next theorem states that \( M(\beta) \) is the rate function in the Laplace principle for the sequence \( \{X_n, n \in \mathbb{N}\} \). Indeed, when \( X^n(0) = 0 \), then as we mentioned at the start of this section, Theorem 7.2.3 and the contraction principle yield the Laplace principle for this sequence with rate function

\[
\Upsilon_0(\beta) \doteq \inf \{ I_0(\varphi) : \varphi \in C([0,1] : \mathbb{R}^d), \varphi(1) = \beta \}.
\]

Let \( \varphi \) be any path in \( C([0,1] : \mathbb{R}^d) \) satisfying \( \varphi(1) = \beta \). The key to the proof that \( \Upsilon_0(\beta) = M(\beta) \) is to partition the time interval \([0,1]\) according to the last time \( t \) that \( (\varphi(t))_1 = 0 \) and then to consider the infimization problem on the intervals \([0,t]\) and \([t,1]\) separately. This procedure gives a new proof of the main result, Theorem 2.1, in [39] when the sets ri(conv \( S_{\mu(\cdot)} \)) coincide. When the measures \( \mu^{(1)} \) and \( \mu^{(2)} \) are equal, that theorem reduces to Cramér’s Theorem [Theorem 3.5.1]. However, in contrast to the rate function in Cramér’s Theorem, which is always convex on \( \mathbb{R}^d \), in general \( M(\beta) \) does not have this convexity property. This will be illustrated by means of some one-dimensional examples later in the section. Other properties of \( M(\beta) \) are given in part (b) of the next theorem, which is proved in [39, Prop. 4.3].

**Theorem 7.3.2.** We assume Condition 7.3.1. Setting \( X_0^n = 0 \), we consider the sequence \( \{X_n, n \in \mathbb{N}\} \) defined in equation (7.1). The following conclusions hold.

(a) The sequence satisfies the Laplace principle on \( \mathbb{R}^d \) with rate function \( M(\beta) \) defined in equation (7.11).

(b) \( M(\beta) \) is lower semicontinuous and superlinear and is convex on each halfspace \( \{\beta \in \mathbb{R}^d : (-1)^i \beta_1 \geq 0\}, i = 1,2 \). However, \( M(\beta) \) is in general not convex on \( \mathbb{R}^d \).
We spend the remainder of this section interpreting the rate function $M(\beta)$ in this theorem and giving examples. For simplicity we will assume that the supports of $\mu^{(1)}$ and $\mu^{(2)}$ are all of $\mathbb{R}^d$. Then both $L^{(1)}$ and $L^{(2)}$ are finite on all of $\mathbb{R}^d$, as are $L^{(0)}$, $\bar{L}^{(1)}$, $\bar{L}^{(2)}$, and $M$. For all $\beta \in \mathbb{R}^d$ the Laplace principle yields the large deviation limit

$$
\lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log P\{\|X_n - \beta\| < \varepsilon\} = -M(\beta).
$$

Let us consider the case where the one-component $\beta_1$ of $\beta$ equals 0, so that $M(\beta)$ equals $L^{(0)}(\beta)$. Then the large deviation limit allows us to interpret $L^{(0)}(\beta)$, whenever it is positive, as the positive cost associated with the atypical event that the final position vectors $X_n$ of the random walk “track” $\beta$. If the infimum in the definition of $L^{(0)}(\beta)$ is attained at $\rho^{(1)}$, $\rho^{(2)}$, $\beta^{(1)}$, $\beta^{(2)}$, then the most likely way for the final position vectors to track $\beta$ is for the increments of the random walk to track $\beta^{(1)}$ in the left halfspace and $\beta^{(2)}$ in the right halfspace. When this occurs, $\rho^{(1)}$ and $\rho^{(2)}$ are the asymptotic fractions of time spent in the respective halfspaces. $\bar{L}^{(1)}$ and $\bar{L}^{(2)}$ may be interpreted similarly.

We now consider the form of $M(\beta)$ in the case $d = 1$. When $d = 1$, equation (7.10) for $\bar{L}^{(i)}(\beta)$ takes the form

$$
\bar{L}^{(i)}(\beta) = \inf \left\{ (1 - \gamma)L^{(0)}(0) + \gamma L^{(i)}(\beta/\gamma) : 0 < \gamma \leq 1 \right\} \quad \text{for } i = 1, 2. \quad (7.12)
$$

According to Lemma 3.1 in [39], when $(-1)^i \beta > 0$ the infimum is uniquely attained. If it is attained at $\gamma \in (0, 1)$, then in tracking $\beta$ the random walk, with high probability, spends at the origin an asymptotically positive fraction $(1 - \gamma)$ of the total time $n$, paying a cost $(1 - \gamma)\bar{L}^{(0)}(0)$ before entering the halfline where $\beta$ lies. If the infimum in (7.12) is attained at $\gamma = 1$, then in tracking $\beta$ the random walk, with high probability, spends at the origin an asymptotically zero fraction of the total time $n$ before entering the halfline where $\beta$ lies.

The following three examples of rate functions $M(\beta)$ in the case $d = 1$ are taken from Section 3 of [39], where complete proofs can be found. The calculations depend on determining where in equation (7.12) the infimum is attained, and in each example this can be done precisely. For $i = 1, 2$ $\bar{\beta}^{(i)}$ denotes the mean $\int_{\mathbb{R}} y \mu^{(i)}(dy)$, which is the unique minimum point and the unique zero of $L^{(i)}$.

**Example 1:** Stable Boundary. A stable boundary is characterized by $\bar{\beta}^{(1)} > 0$ and $\bar{\beta}^{(2)} < 0$. Thus in each open halfline the random walk tends to move toward the origin. In this case there exist unique points $b^{(1)} < 0$ and $b^{(2)} > 0$ such that the following hold. For $\beta < b^{(1)} \bar{L}^{(1)}(\beta)$ equals $L^{(1)}(\beta)$; for $\beta \in [b^{(1)}, 0)$ $\bar{L}^{(1)}(\beta)$ equals the linear function $\beta L^{(1)}(b^{(1)})/b^{(1)}$. Similarly, for $\beta > b^{(2)} \bar{L}^{(2)}(\beta)$ equals $L^{(2)}(\beta)$; for $\beta \in (0, b^{(2)}) \bar{L}^{(2)}(\beta)$ equals the linear function $\beta L^{(2)}(b^{(2)})/b^{(2)}$. The rate function $M(\beta)$ is shown in Figure 1. It is convex but not strictly convex, has a unique minimum point at the origin, and is not differentiable at the origin.

Figure 1. Rate function for $d = 1$: stable boundary.
Example 2: Unstable Boundary. An unstable boundary is characterized by $\tilde{\beta}(1) < 0$ and $\tilde{\beta}(2) > 0$. Thus in each open halfline the random walk tends to move away from the origin. Since $\tilde{\beta}(1) < 0$ and $\tilde{\beta}(2) > 0$, both $L^{(1)}(0)$ and $L^{(2)}(0)$ must be positive. In the case where $L^{(1)}(0) = L^{(2)}(0)$, then for $\beta < 0 \tilde{L}^{(1)}(\beta)$ equals $L^{(1)}(\beta)$ and for $\beta > 0 \tilde{L}^{(2)}(\beta)$ equals $L^{(2)}(\beta)$. The rate function $M(\beta)$ is shown in Figure 2. It is continuous but not convex, has minimum points at $\tilde{\beta}(1)$ and $\tilde{\beta}(2)$, and is not differentiable at the origin. $L(\beta)$ is real analytic for $\beta \in (-\infty, 0) \cup (0, \infty)$. If $L^{(1)}(0) \neq L^{(2)}(0)$, then $M(\beta)$ has a slightly more complicated form.

Figure 2. Rate function for $d = 1$: unstable boundary with $L^{(1)}(0) = L^{(2)}(0)$.

Example 3: One-Sided Unstable Boundary. A one-sided unstable boundary is characterized by $\tilde{\beta}(1) = 0$ and $\tilde{\beta}(2) > 0$. In this case, for $\beta < 0 \tilde{L}^{(1)}(\beta)$ equals $L^{(1)}(\beta)$. For $\beta > \tilde{\beta}(2) \tilde{L}^{(2)}(\beta)$ equals $L^{(2)}(\beta)$ and for $\beta \in (0, \tilde{\beta}(2)] \tilde{L}^{(2)}(\beta)$ equals 0. The rate function $M(\beta)$ is shown in Figure 3. It is differentiable and convex but not strictly convex, and it attains its minimum at all points in the interval $[0, \tilde{\beta}(2)]$.

Figure 3. Rate function for $d = 1$: one-sided unstable boundary.

This completes our discussion of the Laplace principle for the sequence of final position vectors of the random walk and our presentation of one-dimensional examples. In the next section we prove the Laplace principle upper bound for the random walk model with discontinuous statistics.

### 7.4 Proof of the Laplace Principle Upper Bound

In this section the following proposition will be proved.

**Proposition 7.4.1.** We assume Condition 7.2.1. For $x \in \mathbb{R}^d$ we define $I_x$ as in Theorem 7.2.3. Then for all bounded continuous functions $h$ mapping $\mathcal{C}([0, 1]: \mathbb{R}^d)$ into $\mathbb{R}$ we have the Laplace principle upper bound

$$\limsup_{n \to \infty} \frac{1}{n} \log E_x \{\exp[-n h(X^n)]\} \leq - \inf_{\varphi \in \mathcal{C}([0, 1]: \mathbb{R}^d)} \{I_x(\varphi) + h(\varphi)\}.$$ 

The Laplace principle upper bound is equivalent to the lower limit

$$\liminf_{n \to \infty} W^n(x) \geq \inf_{\varphi \in \mathcal{C}([0, 1]: \mathbb{R}^d)} \{I_x(\varphi) + h(\varphi)\},$$ \quad (7.13)
where
\[ W^n(x) = -\frac{1}{n} \log E_x \{ \exp[-n h(X^n)] \}. \]

As in Section 6.2, the key to the proof is to use the stochastic control representation formula for \( W^n(x) \) given in Theorem 4.2.2 and Corollary 5.2.1. This representation formula states that
\[
W^n(x) = V^n(x) \quad (7.14)
\]
\[
\begin{aligned}
\quad &\quad = \inf_{\{\nu^n_j\}} \bar{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot)||\mu(\cdot|\bar{X}^n_j)) + h(\bar{X}^n) \right\} \\
&\quad = \inf_{\{\nu^n_j\}} \bar{E}_x \left\{ \int_0^1 R(\nu^n(\cdot|t)||\mu(\cdot|\bar{X}^n(t))) dt + h(\bar{X}^n) \right\}.
\end{aligned}
\]

In order to help the reader, let us recall the definitions. The infima in the last two lines of this display are taken over all admissible control sequences \( \{\nu^n_j\} \), each \( \nu^n_j \) being a stochastic kernel on \( \mathbb{R}^d \) given \((\mathbb{R}^d)^{j+1}\). Given an admissible control sequence we defined a sequence of controlled random vectors by \( \bar{X}^n_0 = X^0_0 = x \) and
\[
\bar{X}^n_{j+1} = \bar{X}^n_j + \frac{1}{n} Y^0_j,
\]
where
\[
\bar{P}_x \{ \bar{Y}^n_j \in dy|\bar{X}^n_0, \bar{X}^n_1, \ldots, \bar{X}^n_j \} = \nu^n_j(dy|\bar{X}^n_0, \bar{X}^n_1, \ldots, \bar{X}^n_j).
\]
The process \( \bar{X}^n = \{ \bar{X}^n(t), t \in [0, 1] \} \) defined in formula (5.6), is the piecewise linear interpolation of these controlled random vectors. \( \bar{E}_x \) denotes expectation with respect to \( \bar{P}_x \), and the quantity \( \nu^n_j(\cdot) \) appearing as the first argument of the relative entropy equals \( \nu^n_j(\cdot|\bar{X}^n_0, \bar{X}^n_1, \ldots, \bar{X}^n_j) \). In the third line of (7.14), the stochastic kernel \( \nu^n(dy|t) \) is defined in terms of an admissible control sequence \( \{\nu^n_j\} \) by
\[
\nu^n(dy|t) = \begin{cases} 
\nu^n_j(dy|\bar{X}^n_0, \bar{X}^n_1, \ldots, \bar{X}^n_j) & \text{for } t \in [j/n, (j+1)/n), j = 0, 1, \ldots, n-2 \\
\nu^n_{n-1}(dy|\bar{X}^n_0, \bar{X}^n_1, \ldots, \bar{X}^n_n) & \text{for } t \in [(n-1)/n, 1].
\end{cases}
\]

We then define the admissible control measure \( \nu^n \) by
\[
\nu^n(A \times B) = \int_B \nu^n(A|t) \, dt \quad (7.16)
\]
for Borel subsets \( A \) of \( \mathbb{R}^d \) and \( B \) of \( [0, 1] \). Finally, the process \( \bar{X}^n = \{ \bar{X}^n(t), t \in [0, 1] \} \), defined in formula (5.5), is the piecewise constant interpolation of the controlled random vectors \( \bar{X}^n_j \).

It is sufficient to prove the lower limit (7.13) when \( n \) is replaced by any subsequence along which the functions \( W^n(x) \) converge. Such a subsequence exists since \( |W^n(x)| \leq \|h\|_\infty \). We will work with a fixed such subsequence for the remainder of the proof, indexing it, as usual, by \( n \in \mathbb{N} \).
7.4. Proof of Upper Bound

Given \( \varepsilon > 0 \), let \( \{ \nu^n, n \in \mathbb{N} \} \) be a sequence of admissible control measures satisfying for each \( n \)
\[
V^n(x) + \varepsilon \geq \bar{E}_x \left\{ \int_0^1 R \left( \nu^n(\cdot | t) \| \mu(\cdot | X^n(t)) \right) dt + h(X^n) \right\}.
\]
Until this point we have proceeded in precisely the same way as in the proof of the Laplace principle upper bound for the random walk model with continuous statistics, which is given in Section 6.2. However, because of the discontinuity of the function mapping \( x \mapsto \mu(dy|x) \) in the present model, we must now work harder. Specifically, it will be essential to analyze the controls used in each of the halfspaces \( \Lambda^{(1)} \) and \( \Lambda^{(2)} \) and their relations to the asymptotic fractions of time that the processes \( X^n \) spend in each of these halfspaces. This necessitates the introduction of additional processes and measures.

For Borel subsets \( A \) of \( \mathbb{R}^d \) and \( B \) of \([0, 1]\) we define
\[
\nu^{(1),n}(A \times B) = \int_B 1 \{ s \in [0, 1] \mid (\bar{X}^n(s))_1 \leq 0 \} \nu^n(A | t) dt, \tag{7.17}
\]
\[
\nu^{(2),n}(A \times B) = \int_B 1 \{ s \in [0, 1] \mid (\bar{X}^n(s))_1 > 0 \} \nu^n(A | t) dt. \tag{7.18}
\]
These quantities take values in the space \( \mathcal{M}(\mathbb{R}^d \times [0, 1]) \) of subprobability measures on \( \mathbb{R}^d \times [0, 1] \). We also define \( \gamma^{(1),n} \) and \( \gamma^{(2),n} \) to be the respective second marginals of \( \nu^{(1),n} \) and \( \nu^{(2),n} \). Thus
\[
\gamma^{(1),n}(B) = \nu^{(1),n}(\mathbb{R}^d \times B) = \int_B 1 \{ s \in [0, 1] \mid (\bar{X}^n(s))_1 \leq 0 \} \nu^n(A | t) dt \tag{7.19}
\]
and
\[
\gamma^{(2),n}(B) = \nu^{(2),n}(\mathbb{R}^d \times B) = \int_B 1 \{ s \in [0, 1] \mid (\bar{X}^n(s))_1 > 0 \} \nu^n(A | t) dt. \tag{7.20}
\]
These quantities are the Lebesgue measures of the sets of times \( t \in B \) at which \( \bar{X}^n(t) \) lies in the respective halfspaces; \( \gamma^{(1),n} \) and \( \gamma^{(2),n} \) take values in \( \mathcal{M}([0, 1]) \). For \( i = 1, 2 \)
\[
\nu^{(i),n}(A \times B) = \int_B \nu^n(A | t) \gamma^{(i),n}(dt),
\]
which is summarized as \( \nu^{(i),n}(dy \times dt) = \nu^n(dy | t) \otimes \gamma^{(i),n}(dt) \). Since \( \nu^n = \nu^{(1),n} + \nu^{(2),n} \), it follows that for \( i = 1, 2 \) and any nonnegative measurable function \( g \) mapping \( \mathbb{R}^d \times [0, 1] \) into \([0, \infty]\)
\[
0 \leq \int_{\mathbb{R}^d \times [0, 1]} g(y, t) \nu^{(i),n}(dy \times dt) \leq \int_{\mathbb{R}^d \times [0, 1]} g(y, t) \nu^n(dy \times dt). \tag{7.21}
\]

We will make use of this observation in a moment when we prove the following extension of Proposition 5.3.2. This extension gives the tightness of the sequence \( \{ (\nu^n, \nu^{(1),n}, \nu^{(2),n}, \gamma^{(1),n}, \gamma^{(2),n}), n \in \mathbb{N} \} \) as well as a uniform integrability property of the individual sequences \( \{ \nu^n \}, \{ \nu^{(1),n} \}, \text{and} \{ \nu^{(2),n} \} \). Inequality (7.22) is a bound on the sequence of running costs associated with the admissible control sequences \( \{ \nu^n \} \), and it coincides with the bound assumed in Proposition 5.3.2 and Theorem 5.3.5. This bound will be automatically satisfied by the admissible control sequences that arise in the proof of the Laplace principle.
Proposition 7.4.2. In the random walk model we assume part (a) of Condition 7.2.1. For each \( n \in \mathbb{N} \) and \( x \in \mathbb{R}^d \), consider any admissible control sequence \( \{\nu^n_j\} \) such that

\[
\sup_{n \in \mathbb{N}} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R \left( \nu^n_j (\cdot) \| \mu (\cdot | \tilde{X}^n_j) \right) \right\} \equiv \Delta < \infty,
\]

where \( \nu^n_j (dy) = \nu^n_j (dy | \tilde{X}^n_0, \tilde{X}^n_1, \ldots, \tilde{X}^n_j) \). Let the sequence of admissible control measures \( \{\nu^n, n \in \mathbb{N}\} \) be defined by (7.15) and (7.16) and for \( i = 1, 2 \) let the sequences \( \{\nu^{(i),n}, n \in \mathbb{N}\} \) and \( \{\gamma^{(i),n}, n \in \mathbb{N}\} \) be defined by (7.17)–(7.20). The following conclusions hold.

(a) The sequence \( \{(\nu^n, \nu^{(1),n}, \nu^{(2),n}, \gamma^{(1),n}, \gamma^{(2),n})\} \) is tight.

(b) The sequences \( \{\nu^n\} \) and \( \{\nu^{(i),n}\}, i = 1, 2 \), have the uniform integrability properties

\[
\lim_{C \to \infty} \sup_{n \in \mathbb{N}} E_x \left\{ \int_{\|y\| > C} \|y\| \nu^n(dy \times dt) \right\} = 0
\]

and

\[
\lim_{C \to \infty} \sup_{n \in \mathbb{N}} E_x \left\{ \int_{\|y\| > C} \|y\| \nu^{(i),n}(dy \times dt) \right\} = 0.
\]

Proof. Since part (a) of Condition 7.2.1 implies Condition 5.3.1, we can appeal to Proposition 5.3.2, which proves the first limit in part (b). Since (7.21) holds for any nonnegative measurable function \( g \), the first limit in part (b) implies the second limit. As in the proof of Proposition 5.3.2, these limits together with Theorem A.3.17 imply that the individual sequences \( \{\nu^n\}, \{\nu^{(1),n}\}, \) and \( \{\nu^{(2),n}\} \) are tight. The random measures \( \gamma^{(1),n} \) and \( \gamma^{(2),n} \) take values in \( \mathcal{M}([0, 1]) \), which is compact since \([0, 1]\) is compact [Corollary A.3.16]. We conclude that \( \{(\nu^n, \nu^{(1),n}, \nu^{(2),n}, \gamma^{(1),n}, \gamma^{(2),n})\} \) is tight.

Before we continue, it is convenient to extend the concept of a stochastic kernel to suitable collections of subprobability measures. Let \((\mathcal{V}, \mathcal{A})\) be a measurable space and \(\mathcal{Y}\) a Polish space and let \(\tau(dy|x)\) be a family of subprobability measures on \(\mathcal{Y}\) parametrized by \(x \in \mathcal{V}\). We call \(\tau(dy|x)\) a \textbf{stochastic kernel} on \(\mathcal{Y}\) given \(\mathcal{V}\) if for every Borel subset \(E\) of \(\mathcal{Y}\) the function mapping \(x \in \mathcal{V} \mapsto \tau(E|x) \in [0, 1]\) is measurable. As in the case of stochastic kernels, a family \(\tau(dy|x)\) of subprobability measures on \(\mathcal{Y}\) parametrized by \(x \in \mathcal{V}\) is a stochastic kernel if and only if the function mapping \(x \in \mathcal{V} \mapsto \tau(\cdot|x) \in \mathcal{M}(\mathcal{V})\) is measurable. This is proved exactly like Theorem A.5.2.

In addition to the compactness of the random measures as expressed in Proposition 7.4.2, we will need to understand the convergence properties of these random measures and of the associated controlled processes. This will be done in Theorem 7.4.4. The next lemma gives decompositions of quantities arising in the theorem.

Lemma 7.4.3. Let \((\nu, \nu^{(1)}, \nu^{(2)}, \gamma^{(1)}, \gamma^{(2)})\) be the limit in distribution of a convergent subsequence of \(\{(\nu^n, \nu^{(1),n}, \nu^{(2),n}, \gamma^{(1),n}, \gamma^{(2),n}), n \in \mathbb{N}\}\). Then there exists a probability space \((\Omega, \mathcal{F}, P_\nu)\) such that the following hold.
7.4. PROOF OF UPPER BOUND

(a) The limiting quantity $\nu$ is a stochastic kernel $\nu(dy \times dt|\omega)$ on $\mathbb{R}^d \times [0, 1]$ given $\bar{\Omega}$. There exists a stochastic kernel $\nu(dy|t, \omega)$ on $\mathbb{R}^d$ given $[0, 1] \times \bar{\Omega}$ such that $\bar{P}_x$-a.s. for $\omega \in \bar{\Omega}$

$$\nu(A \times B|\omega) = \int_B \nu(A|t, \omega) \, dt$$

for all Borel subsets $A$ of $\mathbb{R}^d$ and $B$ of $[0, 1]$.

(b) For $i = 1, 2$ $\nu^{(i)}$ is a substochastic kernel $\nu^{(i)}(dy \times dt|\omega)$ on $\mathbb{R}^d \times [0, 1]$ given $\bar{\Omega}$. There exists a substochastic kernel $\nu^{(i)}(dy|t, \omega)$ on $\mathbb{R}^d$ given $[0, 1] \times \bar{\Omega}$ such that $\bar{P}_x$-a.s. for $\omega \in \bar{\Omega}$

$$\nu^{(i)}(A \times B|\omega) = \int_B \nu^{(i)}(A|t, \omega) \, dt$$

for all Borel subsets $A$ of $\mathbb{R}^d$ and $B$ of $[0, 1]$. With probability 1, we have a.s. (with respect to Lebesgue measure) for $t \in [0, 1]$

$$\nu^{(1)}(dy|t, \omega) + \nu^{(2)}(dy|t, \omega) = \nu(dy|t, \omega).$$

(c) For $i = 1, 2$ $\gamma^{(i)}$ is a substochastic kernel $\gamma^{(i)}(dt|\omega)$ on $[0, 1]$ given $\bar{\Omega}$. There exists a measurable function $\hat{\gamma}^{(i)}(t, \omega)$ mapping $[0, 1] \times \bar{\Omega}$ into $[0, 1]$ such that $\bar{P}_x$-a.s. for $\omega \in \bar{\Omega}$

$$\gamma^{(i)}(B|\omega) = \int_B \hat{\gamma}^{(i)}(t, \omega) \, dt$$

for all Borel subsets $B$ of $[0, 1]$. In addition, for each $\omega \in \bar{\Omega}$ and each $t \in [0, 1]$ $\hat{\gamma}^{(i)}(t, \omega) = \nu^{(i)}(\mathbb{R}^d|t, \omega)$, so that if for some $t \in [0, 1]$ $\hat{\gamma}^{(i)}(t, \omega) = 0$, then the substochastic kernel $\nu^{(i)}(dy|t, \omega)$ equals 0.

In the sequel $\omega$ will be suppressed in the notations for $\nu(dy \times dt), \nu(dy|t), \nu^{(i)}(dy \times dt), \nu^{(i)}(dy|t), \gamma^{(i)}(dt)$, and $\hat{\gamma}^{(i)}(t)$, and the displays will be summarized as

$$\nu(dy \times dt) = \nu(dy|t) \otimes dt, \nu^{(i)}(dy \times dt) = \nu^{(i)}(dy|t) \otimes dt, \text{ and } \gamma^{(i)}(dt) = \hat{\gamma}^{(i)}(t) \, dt.$$ 

**Proof.** By the Skorohod Representation Theorem, for $i = 1, 2$ $\nu^n \Longrightarrow \nu$, $\nu^{(i),n} \Longrightarrow \nu^{(i)}$, and $\gamma^{(i),n} \Longrightarrow \gamma^{(i)}$ w.p.1 on $(\Omega, \mathcal{F}, \bar{P}_x)$. We omit the proof of the decomposition of $\nu(dy \times dt|\omega)$ given in part (a) since it is carried out exactly like the proof of Lemma 3.3.1. Since $\nu^n = \nu^{(1),n} + \nu^{(2),n}$, w.p.1 we have $\nu = \nu^{(1)} + \nu^{(2)}$.

We now derive the decomposition of $\nu^{(i)}(dy \times dt|\omega)$ given in part (b), denoting by $(\nu^{(i)}_2(dt|\omega))$ the second marginal of this substochastic kernel. Since for $n \in \mathbb{N}$ and all $\omega \in \bar{\Omega}$ the second marginal of $\nu^n(\cdot|\omega)$ equals Lebesgue measure $\lambda$ on $[0, 1]$, w.p.1 the second marginal of $\nu(\cdot|\omega)$ equals $\lambda$. The probability-1 equality $\nu = \nu^{(1)} + \nu^{(2)}$ now implies that w.p.1 $(\nu^{(i)}_2 \ll \lambda$. By normalizing we can apply Theorems A.5.6 and A.5.7 to obtain information concerning the substochastic kernels $\nu^{(i)}(dy \times dt|\omega)$ and $(\nu^{(i)}_2(dt|\omega))$. The second of these theorems guarantees that w.p.1 there exists a nonnegative measurable version of the Radon–Nikodym derivative

$$\hat{\gamma}^{(i)}(t, \omega) = \frac{d(\nu^{(i)}_2(\cdot|\omega))}{d\lambda(\cdot)}(t).$$
Chapter 7. Random Walk with Discontinuous Statistics

With probability 1, \( \hat{\gamma}(t, \omega) \) is in \([0, 1]\) \( \lambda \)-a.s. for \( t \in [0, 1] \). In turn, Theorem A.5.6 yields the existence of a stochastic kernel \( \tau_i(dy|t, \omega) \) such that w.p.1

\[
\nu_i(A \times B|\omega) = \int_B \tau_i(A|t, \omega) (\nu_i)_2(dt|\omega) = \int_B \hat{\gamma}_i(t, \omega) \tau_i(A|t, \omega) dt
\]

for all Borel subsets \( A \) of \( \mathbb{R}^d \) and \( B \) of \([0, 1]\). This gives the first display in part (b) of the proposition with

\[
\nu_i(dy|t, \omega) \equiv \hat{\gamma}_i(t, \omega) \tau_i(dy|t, \omega).
\]

(7.23)

The probability-1 equality \( \nu = \nu(1) + \nu(2) \) yields w.p.1 the second display in part (b) \( \lambda \)-a.s. for \( t \in [0, 1] \).

For \( n \in \mathbb{N} \), \( i = 1, 2 \), and all \( \omega \in \tilde{\Omega} \), \( \gamma_i^n(\cdot|\omega) \) equals the second marginal of \( \nu_i^n(\cdot|\omega) \). Hence w.p.1 \( \gamma_i(\cdot|\omega) \) equals the second marginal of \( \nu_i(\cdot|\omega) \). It follows that w.p.1

\[
\gamma_i(B|\omega) = (\nu_i)_2(B|\omega) = \int_B \hat{\gamma}_i(t, \omega) dt
\]

for every Borel subset \( B \) of \([0, 1]\). This gives the decomposition in part (c). Since \( \tau_i(dy|t, \omega) \) is a stochastic kernel on \( \mathbb{R}^d \) given \([0, 1] \times \tilde{\Omega} \), it follows that \( \tau_i(\mathbb{R}^d|t, \omega) = 1 \) and thus from equation (7.23) that \( \hat{\gamma}_i(t, \omega) = \nu_i(\mathbb{R}^d|t, \omega) \). The proof of the lemma is complete.

We return to the proof of the Laplace principle upper bound. Given \( \varepsilon > 0 \) we have selected a sequence \( \{\nu^n, n \in \mathbb{N}\} \) of admissible control measures satisfying for each \( n \in \mathbb{N} \)

\[
V^n(x) + \varepsilon \geq E_x \left\{ \int_0^1 R(\nu^n(\cdot|t)||\mu(\cdot|\bar{X}^n(t))) dt + h(\bar{X}^n) \right\}.
\]

(7.24)

In order to relate limit points of the sequence \( \{(\nu^n, \nu^{(1), n}, \nu^{(2), n}, \gamma_1^n, \gamma_2^n)\} \) and of the sequence of controlled processes \( \{(\bar{X}^n, \bar{X}^n)\} \), we would like to apply the convergence and compactness results stated in Theorem 5.3.5 and Proposition 7.4.2. The following theorem derives several key properties of the limiting quantities \( \nu, \nu_1, \nu_2, \gamma_1, \gamma_2, \) and \( \bar{X} \). The theorem is an adaptation of a result in [38]. The bound (7.25) on the sequence of running costs will be automatically satisfied in the proof of the Laplace principle.

**Theorem 7.4.4.** In the random walk model we assume part (a) of Condition 7.2.1. For each \( n \in \mathbb{N} \) and \( x \in \mathbb{R}^d \), consider any admissible control sequence \( \{\nu^n\} \) such that

\[
\sup_{n \in \mathbb{N}} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot)||\mu(\cdot|\bar{X}^n_j)) \right\} < \infty,
\]

(7.25)

where \( \nu^n_j(dy) = \nu^n_j(dy|\bar{X}^n_j, \bar{X}^n, \ldots, \bar{X}^n) \). The following conclusions hold.

(a) Given any subsequence of \( \{(\nu^n, \nu^{(1), n}, \nu^{(2), n}, \gamma_1^n, \gamma_2^n, \bar{X}^n, \bar{X}^n)\} \), there exist a subsequence, a probability space \( (\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}) \), a stochastic kernel \( \nu \) on \( \mathbb{R}^d \times [0, 1] \) given \( \tilde{\Omega} \), substochastic kernels \( \nu_1 \) and \( \nu_2 \) on \( \mathbb{R}^d \times [0, 1] \) given \( \tilde{\Omega} \), substochastic kernels \( \gamma_1 \) and \( \gamma_2 \) on \([0, 1]\) given \( \tilde{\Omega} \), and a random variable \( \bar{X} \) mapping \( \tilde{\Omega} \) into \( \mathcal{C}([0, 1]: \mathbb{R}^d) \) such
that the subsequence converges in distribution to \((\nu, \nu^{(1)}, \nu^{(2)}, \gamma^{(1)}, \gamma^{(2)}, \bar{X}, \bar{X})\). These (sub)stochastic kernels have the decompositions given in Lemma 7.4.3.

(b) For \(i = 1, 2\)

\[
\int_{\mathbb{R}^d \times [0, 1]} y \nu^{(i,n)}(dy \times dt) \overset{D}{\to} \int_{\mathbb{R}^d \times [0, 1]} y \nu^{(i)}(dy \times dt).
\]

(c) With probability 1, we have a.s. (with respect to Lebesgue measure) for \(t \in [0, 1]\)

whenever \((\bar{X}(t))_1 < 0, \gamma^{(1)}(t) = \nu^{(1)}(\mathbb{R}^d|t) = 1\) and \(\gamma^{(2)}(t) = \nu^{(2)}(\mathbb{R}^d|t) = 0,\)

whenever \((\bar{X}(t))_1 > 0, \gamma^{(2)}(t) = \nu^{(2)}(\mathbb{R}^d|t) = 1\) and \(\gamma^{(1)}(t) = \nu^{(1)}(\mathbb{R}^d|t) = 0,\)

and for any value of \((\bar{X}(t))_1\)

\[
\gamma^{(1)}(t) + \gamma^{(2)}(t) = \nu^{(1)}(\mathbb{R}^d|t) + \nu^{(2)}(\mathbb{R}^d|t) = 1.
\]

(d) With probability 1, we have a.s. for \(t \in [0, 1]\) that whenever \((\bar{X}(t))_1 = 0\)

\[
\left(\int_{\mathbb{R}^d} y \nu^{(1)}(dy|t)\right)_1 \geq 0
\]

and

\[
\left(\int_{\mathbb{R}^d} y \nu^{(2)}(dy|t)\right)_1 \leq 0.
\]

(e) With probability 1, for every \(t \in [0, 1]\)

\[\bar{X}(t) = x + \int_{\mathbb{R}^d \times [0,t]} y \nu(dy \times ds) = x + \int_0^t \left(\int_{\mathbb{R}^d} y \nu(dy|s)\right) ds\]

and \(\bar{X}(t)\) is an absolutely continuous function of \(t \in [0, 1]\). Therefore, a.s. for \(t \in [0, 1]\)

the derivative of \(\bar{X}(t)\) is given by

\[
\dot{\bar{X}}(t) = \int_{\mathbb{R}^d} y \nu(dy|t) = \int_{\mathbb{R}^d} y \nu^{(1)}(dy|t) + \int_{\mathbb{R}^d} y \nu^{(2)}(dy|t).
\]

**Proof.** We will prove these properties in the order (a), (b), (c), (e), and (d). Except for part (d), the proof of which is quite technical, the other parts of the theorem are not difficult to prove. Part (a) of Condition 7.2.1 implies Condition 5.3.1, which is needed to apply Proposition 5.3.2 and Theorem 5.3.5.

(a) The convergence in distribution asserted in part (a) is a consequence of part (a) of Theorem 5.3.5 and part (a) of Proposition 7.4.2. The identification of the limiting quantities is carried out in Theorem 5.3.5 and Lemma 7.4.3.

(b) These limits are due to the uniform integrability given in part (b) of Proposition 7.4.2. They are proved exactly like the limit \(S^n \overset{D}{\to} \bar{X}\) in Step 2 in the proof of Theorem 5.3.5, which asserts that

\[
\int_{\mathbb{R}^d \times [0,1]} y \nu^n(dy \times dt) \overset{D}{\to} \int_{\mathbb{R}^d \times [0,1]} y \nu(dy \times dt).
\]
In order to complete the proof of the theorem, we invoke the Skorohod Representation Theorem, which allows us to assume that the convergence asserted in parts (a) and (b) occurs w.p.1.

(c) We denote Lebesgue measure on $[0,1]$ by $\lambda$. Since for $n \in \mathbb{N}$ $\gamma^{(1)} n + \gamma^{(2)} n = \lambda$, we have w.p.1 $\gamma^{(1)} + \gamma^{(2)} = \lambda$. Hence w.p.1, by part (c) of Lemma 7.4.3 the respective densities satisfy a.s. for $t \in [0,1]$

$$\hat{\gamma}^{(1)}(t) + \hat{\gamma}^{(2)}(t) = \nu^{(1)}(\mathbb{R}^d|t) + \nu^{(2)}(\mathbb{R}^d|t) = 1.$$

This gives (7.27).

We next prove the first line of formula (7.26). With probability 1, since $\tilde{X}(t)$ is a continuous function of $t \in [0,1]$, there exist random variables $a_i$ and $b_i$ satisfying $0 \leq a_i \leq b_i \leq 1$ and indexed by $i \in \mathbb{N}$ such that

$$\{t \in (0, 1) : (\tilde{X}(t))_i < 0\} = \cup_{i \in \mathbb{N}}(a_i, b_i).$$

Choose any $\omega \in \Omega$ such that $(\tilde{X}(t, \omega))_i < 0$ for all $t \in (a_i(\omega), b_i(\omega))$. Then given any $\delta > 0$, there exists $N \in \mathbb{N}$ such that $(\tilde{X}^n(s))_i < 0$ for all $n \geq N$ and all $s \in (a_i + \delta, b_i - \delta)$. Hence

$$\nu^{(2), n}(\mathbb{R}^d \times (a_i + \delta, b_i - \delta)) = \int_{\mathbb{R}^d \times [0,1]} 1_{(a_i + \delta, b_i - \delta)}(s) \nu^{(2), n}(dy \times ds) = 0.$$

With probability 1, since $\nu^{(2), n} \rightarrow \nu^{(2)}$, it follows that

$$\nu^{(2)}(\mathbb{R}^d \times (a_i + \delta, b_i - \delta)) = \int_{a_i + \delta}^{b_i - \delta} \nu^{(2)}(\mathbb{R}^d|s) ds = 0.$$

Sending $\delta \to 0$, we obtain w.p.1

$$\int_{a_i}^{b_i} \nu^{(2)}(\mathbb{R}^d|s) ds = 0.$$

With probability 1 this implies that $\lambda$-a.s. for $s \in (a_i, b_i)$ $\hat{\gamma}^{(2)}(s) = \nu^{(2)}(\mathbb{R}^d|s) = 0$ and thus that $\nu^{(1)}(\mathbb{R}^d|s) = \hat{\gamma}^{(1)}(s) = 1$ [Lemma 7.4.3 (c)]. This completes the proof of the first line of formula (7.26). The second line is proved similarly.

(e) This is a consequence of part (b) of Theorem 5.3.5 and parts (a) and (b) of Lemma 7.4.3.

(d) The proof of (7.28) and (7.29) under the assumption that $(\tilde{X}(t))_i = 0$ is lengthy. Let us temporarily fix $a > 0$. The first step is to build an approximation to the function $G$ mapping $\mathbb{R}$ into $\mathbb{R}$ and defined by

$$G(z) = \begin{cases} 
|z| & \text{if } |z| \leq a \\
 0 & \text{if } |z| > a.
\end{cases}$$
7.4. PROOF OF UPPER BOUND

For each \(\kappa > 0\) let \(G_{\kappa}(z)\) be a twice continuously differentiable function of \(z \in \mathbb{R}\setminus\{0\}\) such that

\[
\begin{align*}
|G_{\kappa}(z)| & \leq 2a \quad \text{for all } z \in \mathbb{R}, \\
G_{\kappa}(z) & = G(z) \quad \text{for } |z| \leq a, \\
|d^2G_{\kappa}(z)/dz^2| & < B \quad \text{for } |z| > a/4,
\end{align*}
\]

where \(B < \infty\) depends on \(\kappa > 0\), and define

\[
g_{\kappa}(z) = \begin{cases} 
\frac{dG_{\kappa}(z)}{dz} & \text{if } z \neq 0 \\
-1 & \text{if } z = 0.
\end{cases}
\]

By suitably defining \(G_{\kappa}(z)\) for \(|z| > a\), we can assume that for \(|z| > a\) \(g_{\kappa}(z) \to 0\) as \(\kappa \to 0\) and that \(g_{\kappa}(z)\) is uniformly bounded for \(\kappa > 0\) and \(z \in \mathbb{R}\). Although \(g_{\kappa}\) is not continuous on all of \(\mathbb{R}\), the restrictions of \(g_{\kappa}\) both to \((-\infty, 0]\) and to \((0, \infty)\) are bounded and continuous, and \(g_{\kappa}(z)\) is Lipschitz continuous for \(|z| \geq a/4\) with constant \(B\). Furthermore

\[
\lim_{\kappa \to 0} g_{\kappa}(z) = \begin{cases} 
-1_{[-a,0]}(z) & \text{if } z \leq 0 \\
1_{(0,a]}(z) & \text{if } z > 0.
\end{cases}
\] (7.31)

For any two points \(x\) and \(y\) in \(\mathbb{R}\), these definitions and the mean value theorem imply the following properties of \(G_{\kappa}\):

- If \(|x - y| \leq a/4\) and \(|x| \vee |y| \leq a/2\), then

\[
G_{\kappa}(y) - G_{\kappa}(x) \geq g_{\kappa}(x) \cdot (y - x).
\]

- If \(|x - y| \leq a/4\) and \(|x| \vee |y| > a/2\), then for some point \(z\) between \(x\) and \(y\)

\[
G_{\kappa}(y) - G_{\kappa}(x) = g_{\kappa}(z) \cdot (y - x).
\]

- If \(|x - y| > a/4\), then

\[
G_{\kappa}(y) - G_{\kappa}(x) \geq -\|g_{\kappa}\|_{\infty} \cdot |y - x|.
\]

Thus for any two points \(x\) and \(y\) in \(\mathbb{R}\)

\[
G_{\kappa}(y) - G_{\kappa}(x) \\
\geq 1_{\{|x-y| \leq a/4, |x| \vee |y| \leq a/2\}} g_{\kappa}(x) \cdot (y - x) \\
+ 1_{\{|x-y| \leq a/4, |x| \vee |y| > a/2\}} g_{\kappa}(z) \cdot (y - x) \\
- 1_{\{|x-y| > a/4\}} \|g_{\kappa}\|_{\infty} \cdot |y - x| \\
\geq 1_{\{|x-y| \leq a/4\}} g_{\kappa}(x) \cdot (y - x) \\
- 1_{\{|x-y| \leq a/4, |x| \vee |y| > a/2\}} |g_{\kappa}(z) - g_{\kappa}(x)| \cdot |y - x| \\
- 1_{\{|x-y| > a/4\}} \|g_{\kappa}\|_{\infty} \cdot |y - x|.
\]
Substituting \( x = (\bar{X}^n_j)_1 \) and \( y = (\bar{X}^n_{j+1})_1 \) and summing over \( j \in \{0, 1, \ldots, n - 1\} \), we see that

\[
4a \geq \sum_{j=0}^{n-1} \left[ G_\kappa \left((\bar{X}^n_{j+1})_1\right) - G_\kappa \left((\bar{X}^n_j)_1\right) \right] \\
\geq \sum_{j=0}^{n-1} 1_{\{(\bar{X}^n_{j+1})_1 - (\bar{X}^n_j)_1 \leq a/4\}} g_\kappa \left((\bar{X}^n_j)_1\right) \cdot \left((\bar{X}^n_{j+1})_1 - (\bar{X}^n_j)_1\right) \\
- \sum_{j=0}^{n-1} 1_{\{(\bar{X}^n_{j+1})_1 - (\bar{X}^n_j)_1 > a/4\} \lor \{|(\bar{X}^n_j)_1| > a/2\}} |g_\kappa(Z^n_j)| - g_\kappa((\bar{X}^n_j)_1)| \cdot |(\bar{X}^n_{j+1})_1 - (\bar{X}^n_j)_1| \\
+ \|g_\kappa\|_\infty \sum_{j=0}^{n-1} 1_{\{(\bar{X}^n_{j+1})_1 - (\bar{X}^n_j)_1 > a/4\}} |(\bar{X}^n_{j+1})_1 - (\bar{X}^n_j)_1| \\
\equiv \Phi_1(n, \kappa) + \Phi_2(n, \kappa) + \Phi_3(n, \kappa).
\]

\( \Phi_1(n, \kappa), \Phi_2(n, \kappa), \) and \( \Phi_3(n, \kappa) \) denote the terms appearing on lines two, three, and four of this display. In \( \Phi_2(n, \kappa) \rightarrow 0 \) is an appropriately selected point between \( (\bar{X}^n_j)_1 \) and \( (\bar{X}^n_{j+1})_1 \).

Fixing an arbitrary \( \kappa > 0 \), we consider the limits as \( n \to \infty \) of \( \Phi_1(n, \kappa), \Phi_2(n, \kappa), \) and \( \Phi_3(n, \kappa) \). We first prove that as \( n \to \infty \) \( \Phi_2(n, \kappa) \xrightarrow{P} 0 \). In order to carry this out, we need the fact that

\[
\lim_{C \to \infty} \sup_{n \in \mathbb{N}} E_x \left\{ \left\{ \sum_{j=0}^{n-1} 1_{\{(\bar{X}^n_{j+1})_1 - (\bar{X}^n_j)_1 \leq C\}} \right\} \right\} = 0,
\]

which expresses the uniform integrability of the sequence \( \{\sum_{j=0}^{n-1} |(\bar{X}^n_{j+1})_1 - (\bar{X}^n_j)_1|, n \in \mathbb{N}\} \). This limit will be proved in the next paragraph. We have

\[
0 \leq -\Phi_2(n, \kappa) \leq A(n, \kappa) \sum_{j=0}^{n-1} |(\bar{X}^n_{j+1})_1 - (\bar{X}^n_j)_1|,
\]

where

\[
A(n, \kappa) \doteq \max_{j=0, 1, \ldots, n-1} 1_{\{(\bar{X}^n_{j+1})_1 - (\bar{X}^n_j)_1 \leq a/4\} \lor \{|(\bar{X}^n_j)_1| > a/2\}} |g_\kappa(Z^n_j)| - g_\kappa((\bar{X}^n_j)_1)|.
\]

Since w.p.1 the sequence \( \{\bar{X}^n, n \in \mathbb{N}\} \) converges uniformly on \([0, 1]\) to the continuous process \( \bar{X} \), we have w.p.1

\[
\max_{j=0, 1, \ldots, n-1} |(\bar{X}^n_{j+1})_1 - (\bar{X}^n_j)_1| = \max_{j=0, 1, \ldots, n-1} \left| \bar{X}^n((j+1)/n)_1 - (\bar{X}^n(j/n))_1 \right| \to 0
\]

and thus, by the Lipschitz continuity of \( g_\kappa(z) \) for \(|z| \geq a/4\), \( A(n, \kappa) \to 0 \) w.p.1. It follows from (7.33) and Chebyshev’s Inequality that

\[
A(n, \kappa) \sum_{j=0}^{n-1} |(\bar{X}^n_{j+1})_1 - (\bar{X}^n_j)_1| \xrightarrow{P} 0.
\]
7.4. PROOF OF UPPER BOUND

We conclude that $\Phi_2(n, \kappa) \xrightarrow{P} 0$, as claimed.

For $j \in \{0, 1, \ldots, n - 1\}$ let $b_j$ denote $|(\overline{X}_j^n) - (\overline{X}_j^n)|$ and take any $C < \infty$. We now prove (7.33) using

$$1 \{ \sum_{j=0}^{n-1} b_j > C \} \sum_{j=0}^{n-1} b_j \leq 2 \sum_{j=0}^{n-1} 1 \{ n b_j > C/2 \} b_j.$$

To show the latter, assume that $\sum_{j=0}^{n-1} b_j > C$. Then, since

$$C < \sum_{j=0}^{n-1} b_j = \sum_{j=0}^{n-1} 1 \{ n b_j > C/2 \} b_j + \sum_{j=0}^{n-1} 1 \{ n b_j \leq C/2 \} b_j \leq \sum_{j=0}^{n-1} 1 \{ n b_j > C/2 \} b_j + C/2,$$

we have

$$C/2 \leq \sum_{j=0}^{n-1} 1 \{ n b_j > C/2 \} b_j,$$

and so

$$1 \{ \sum_{j=0}^{n-1} b_j > C \} \sum_{j=0}^{n-1} b_j \leq \sum_{j=0}^{n-1} 1 \{ n b_j > C/2 \} b_j + C/2 \leq 2 \sum_{j=0}^{n-1} 1 \{ n b_j > C/2 \} b_j,$$

as claimed. Since $\nu_j^n(dy|\overline{X}_0^n, \overline{X}_1^n, \ldots, \overline{X}_j^n)$ is a regular conditional distribution for $n(\overline{X}_{j+1}^n - \overline{X}_j^n)$ given $\overline{X}_0^n, \overline{X}_1^n, \ldots, \overline{X}_j^n$, it follows that

$$E_x \left\{ 1 \{ \sum_{j=0}^{n-1} |(\overline{X}_j^n)| > C \} \sum_{j=0}^{n-1} |(\overline{X}_{j+1}^n) - (\overline{X}_j^n)| \right\}$$

$$\leq 2 E_x \left\{ \sum_{j=0}^{n-1} 1 \{ n |(\overline{X}_j^n)| > C/2 \} |(\overline{X}_{j+1}^n) - (\overline{X}_j^n)| \right\}$$

$$= 2 E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \int_{\{y \in \mathbb{R}^d : |y_1| > C/2\}} |y_1| \nu_j^n(dy) \right\}$$

$$\leq 2 E_x \left\{ \int \int_{\{y \in \mathbb{R}^d : |y| > C/2 \} \times [0, 1]} \|y\| \nu^n(dy \times ds) \right\}.$$

Hence by Proposition 5.3.2

$$\lim_{n \to \infty} \sup_{n \in \mathbb{N}} E_x \left\{ 1 \{ \sum_{j=0}^{n-1} |(\overline{X}_j^n)| > C \} \sum_{j=0}^{n-1} |(\overline{X}_{j+1}^n) - (\overline{X}_j^n)| \right\}$$

$$\leq 2 \lim_{n \to \infty} \sup_{n \in \mathbb{N}} E_x \left\{ \int \int_{\{y \in \mathbb{R}^d : |y| > C/2 \} \times [0, 1]} \|y\| \nu^n(dy \times ds) \right\} = 0.$$

This proves the uniform integrability of $\{ \sum_{j=0}^{n-1} |(\overline{X}_j^n)|, n \in \mathbb{N} \}$ and completes the proof that $\Phi_2(n, \kappa) \xrightarrow{P} 0$. 


We next prove that in (7.32) \( \Phi_3(n, \kappa) \xrightarrow{P} 0 \). Indeed

\[
E_x \{-\Phi_3(n, \kappa)\} = \|g_\kappa\|_\infty E_x \left\{ \sum_{j=0}^{n-1} 1_{\{|(\tilde{X}_{j+1}^n) - (\tilde{X}_j^n)| > a/4\}} |(\tilde{X}_{j+1}^n) - (\tilde{X}_j^n)| \right\} = \|g_\kappa\|_\infty E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \int_{\{y \in \mathbb{R}^d : |y| > na/4\}} |y_1| \nu^n_j(dy) \right\}.
\]

Again we may apply Proposition 5.3.2 and conclude that

\[
\lim_{n \to \infty} E_x \{-\Phi_3(n, \kappa)\} = 0.
\]

Since \(-\Phi_3(n, \kappa) > 0\), Chebyshev’s Inequality implies that \( \Phi_3(n, \kappa) \xrightarrow{P} 0 \).

This leaves \( \Phi_1(n, \kappa) \). We rewrite it as

\[
\Phi_1(n, \kappa) = \sum_{j=0}^{n-1} 1_{\{(\tilde{X}_{j+1}^n) - (\tilde{X}_j^n) \leq a/4\}} g_\kappa \left( (\tilde{X}_j^n) \right) \cdot \left( (\tilde{X}_{j+1}^n) - (\tilde{X}_j^n) \right)
\]

\[
= \frac{1}{n} \sum_{j=0}^{n-1} g_\kappa \left( (\tilde{X}_j^n) \right) \int_{\mathbb{R}^d} y_1 \nu^n_j(dy) + \Phi_{1,1}(n, \kappa) + \Phi_{1,2}(n, \kappa),
\]

where

\[
\Phi_{1,1}(n, \kappa) = \sum_{j=0}^{n-1} g_\kappa \left( (\tilde{X}_j^n) \right) \left[ \left( (\tilde{X}_{j+1}^n) - (\tilde{X}_j^n) \right) - \frac{1}{n} \int_{\mathbb{R}^d} y_1 \nu^n_j(dy) \right]
\]

and

\[
\Phi_{1,2}(n, \kappa) = \sum_{j=0}^{n-1} 1_{\{(\tilde{X}_{j+1}^n) - (\tilde{X}_j^n) > a/4\}} g_\kappa \left( (\tilde{X}_j^n) \right) \cdot \left( (\tilde{X}_{j+1}^n) - (\tilde{X}_j^n) \right).
\]

The same truncation and martingale argument used in the proof of Lemma 5.3.7 shows that \( \Phi_{1,1}(n, \kappa) \xrightarrow{P} 0 \). In addition, since \( |\Phi_{1,2}(n, \kappa)| \leq |\Phi_3(n, \kappa)| \), the convergence \( \Phi_3(n, \kappa) \xrightarrow{P} 0 \) to 0 implies that \( \Phi_{1,2}(n, \kappa) \xrightarrow{P} 0 \).

Up to this point in the proof we have shown that

\[
4a \geq \sum_{j=0}^{n-1} \left[ G_\kappa \left( (\tilde{X}_{j+1}^n) \right) - G_\kappa \left( (\tilde{X}_j^n) \right) \right] \geq \frac{1}{n} \sum_{j=0}^{n-1} g_\kappa \left( (\tilde{X}_j^n) \right) \int_{\mathbb{R}^d} y_1 \nu^n_j(dy) + \Phi_{1,1}(n, \kappa) + \Phi_{1,2}(n, \kappa) + \Phi_2(n, \kappa) + \Phi_3(n, \kappa)
\]

\[
= \int_{\mathbb{R}^d \times [0, 1]} y_1 g_\kappa \left( (\tilde{X}_j^n(t)) \right) \nu^n(dy \times dt) + \Phi_{1,1}(n, \kappa) + \Phi_{1,2}(n, \kappa) + \Phi_2(n, \kappa) + \Phi_3(n, \kappa).
\]

(7.34)
7.4. PROOF OF UPPER BOUND

For each $\kappa > 0$ the four terms $\Phi_{1,1}(n, \kappa), \Phi_{1,2}(n, \kappa), \Phi_{2}(n, \kappa),$ and $\Phi_{3}(n, \kappa)$ all converge to 0 in probability as $n \to \infty$. The Borel–Cantelli Lemma guarantees that as $n \to \infty$ along a suitable subsequence these four terms all converge to 0 w.p.1.

We would like to send $n \to \infty$ in the last display. However, care must be exercised because $g_\kappa$ is not continuous on all of $\mathcal{R}$. From the definitions $\nu^n = \nu^{(1),n} + \nu^{(2),n}$, and on the support of $\nu^{(i),n} (\bar{X}^n(t))_1$ is nonpositive for $i = 1$ and is positive for $i = 2$. Furthermore, $g_\kappa$ is bounded and continuous on the closed interval $(-\infty, 0]$ and on the open interval $(0, \infty)$. We define $g^{(1)}_\kappa$ to be a bounded continuous extension to $\mathcal{R}$ of the restriction of $g_\kappa$ to $(-\infty, 0]$ and $g^{(2)}_\kappa$ to be a bounded continuous extension to $\mathcal{R}$ of the restriction of $g_\kappa$ to $(0, \infty)$; thus $g^{(1)}_\kappa(0) = -1$ and $g^{(2)}_\kappa(0) = 1$. It follows that

$$\int_{\mathcal{R}^d \times [0,1]} y_1 g_\kappa((\bar{X}^n(t))_1) \nu^n(dy \times dt) = \int_{\mathcal{R}^d \times [0,1]} y_1 g^{(1)}_\kappa((\bar{X}^n(t))_1) \nu^{(1),n}(dy \times dt) + \int_{\mathcal{R}^d \times [0,1]} y_1 g^{(2)}_\kappa((\bar{X}^n(t))_1) \nu^{(2),n}(dy \times dt).$$

Let $f$ be any bounded continuous function mapping $\mathcal{R}^2$ into $\mathcal{R}$. With probability 1, since $\nu^{(i),n} \Rightarrow \nu^{(i)}$ and $\{\bar{X}^n\}$ converges uniformly on $[0,1]$ to the continuous process $\bar{X}$, we can apply part (b) of Theorem A.3.10 to assert that

$$\lim_{n \to \infty} \int_{\mathcal{R}^d \times [0,1]} f(y) g^{(1)}_\kappa((\bar{X}^n(t))_1) \nu^{(1),n}(dy \times dt) = \int_{\mathcal{R}^d \times [0,1]} f(y) g^{(1)}_\kappa((\bar{X}(t))_1) \nu^{(1)}(dy \times dt).$$

By the uniform integrability given in Proposition 7.4.2, $f(y)$ can be replaced by $y_1$. Hence w.p.1 we obtain from (7.34)

$$4a \geq \int_{\mathcal{R}^d \times [0,1]} y_1 g^{(1)}_\kappa((\bar{X}(t))_1) \nu^{(1)}(dy \times dt) + \int_{\mathcal{R}^d \times [0,1]} y_1 g^{(2)}_\kappa((\bar{X}(t))_1) \nu^{(2)}(dy \times dt).$$

We now use the probability–1 decompositions $\nu^{(i)}(dy \times dt) = \nu^{(i)}(dy|t) \otimes dt$ [Lemma 7.4.3 (b)] and the fact that if $((\bar{X}(t))_1 < 0$, then $\nu^{(2)}(dy|t) = 0$ and if $((\bar{X}(t))_1 > 0$, then $\nu^{(1)}(dy|t) = 0$ [see (7.26)]. This yields

$$4a \geq \int_0^1 \int_{\mathcal{R}^d} y_1 g^{(1)}_\kappa((\bar{X}(t))_1) 1_{(-\infty,0]}((\bar{X}(t))_1) \nu^{(1)}(dy|t) dt + \int_0^1 \int_{\mathcal{R}^d} y_1 g^{(2)}_\kappa((\bar{X}(t))_1) 1_{[0,\infty)}((\bar{X}(t))_1) \nu^{(2)}(dy|t) dt.$$  \hfill (7.35)

The functions $\{g^{(1)}_\kappa, \kappa > 0\}$ and $\{g^{(2)}_\kappa, \kappa > 0\}$ are uniformly bounded and for $z \in \mathcal{R}$ [see (7.31)]

$$\lim_{\kappa \to 0^+} g^{(1)}_\kappa(z) 1_{(-\infty,0]}(z) = -1_{(-\infty,0]}(z) \quad \text{and} \quad \lim_{\kappa \to 0^+} g^{(2)}_\kappa(z) 1_{[0,\infty)}(z) = 1_{[0,\infty)}(z).$$
By the Lebesgue Dominated Convergence Theorem, sending \( \kappa \to 0 \) in formula (7.35) yields w.p.1
\[
\int_0^1 \int_{\mathbb{R}^d} 1_{[0,a]}((\tilde{X}(t)))_1 y_1 \nu^{(2)}(dy|t) \, dt - \int_0^1 \int_{\mathbb{R}^d} 1_{[-a,0]}((\tilde{X}(t)))_1 y_1 \nu^{(1)}(dy|t) \, dt \leq 4a.
\]
Similar proofs that are based on approximating
\[
G(z) = \begin{cases} 
z & \text{if } |z| \leq a \\
0 & \text{if } z > a \\
-a & \text{if } z < -a
\end{cases}
\]
and \(-G(z)\) show that w.p.1
\[
\left| \int_0^1 \int_{\mathbb{R}^d} 1_{[0,a]}((\tilde{X}(t)))_1 y_1 \nu^{(2)}(dy|t) \, dt + \int_0^1 \int_{\mathbb{R}^d} 1_{[-a,0]}((\tilde{X}(t)))_1 y_1 \nu^{(1)}(dy|t) \, dt \right| \leq 4a.
\]
Combining these equations gives w.p.1
\[
\int_0^1 \int_{\mathbb{R}^d} 1_{[-a,0]}((\tilde{X}(t)))_1 y_1 \nu^{(1)}(dy|t) \, dt \geq -8a
\]
and
\[
\int_0^1 \int_{\mathbb{R}^d} 1_{[0,a]}((\tilde{X}(t)))_1 y_1 \nu^{(2)}(dy|t) \, dt \leq 8a.
\]
By taking \( a \to 0 \), we conclude that w.p.1
\[
\int_0^1 \int_{\mathbb{R}^d} 1_{[0]}((\tilde{X}(t)))_1 y_1 \nu^{(1)}(dy|t) \, dt \geq 0
\]
and
\[
\int_0^1 \int_{\mathbb{R}^d} 1_{[0]}((\tilde{X}(t)))_1 y_1 \nu^{(2)}(dy|t) \, dt \leq 0.
\]
Let \([b,c]\) be any closed interval in \([0,1]\). Repeating the argument leading to the last two displays, we obtain
\[
\theta^{(1)}(b,c) \doteq \int_b^c \int_{\mathbb{R}^d} 1_{[0]}((\tilde{X}(t)))_1 y_1 \nu^{(1)}(dy|t) \, dt \geq 0
\]
and
\[
\theta^{(2)}(b,c) \doteq \int_b^c \int_{\mathbb{R}^d} 1_{[0]}((\tilde{X}(t)))_1 y_1 \nu^{(2)}(dy|t) \, dt \leq 0.
\]
With probability 1 these inequalities hold simultaneously for all intervals \([b,c] \subset [0,1]\) with rational endpoints, and thus by continuity they hold simultaneously for all intervals \([b,c] \subset [0,1]\). This implies that w.p.1 \( \theta^{(1)}(0,c) \) is nondecreasing and \( \theta^{(2)}(0,c) \) is nonincreasing for \( c \in [0,1] \). Since a nondecreasing (resp. nonincreasing) function has a nonnegative (resp., nonpositive) derivative a.s., it follows that w.p.1 we have a.s. for \( t \in [0,1] \) whenever \((\tilde{X}(t))_1 = 0\)
\[
\int_{\mathbb{R}^d} y_1 \nu^{(1)}(dy|t) = \left( \int_{\mathbb{R}^d} y \nu^{(1)}(dy|t) \right)_1 \geq 0
\]
7.4. PROOF OF UPPER BOUND

\[
\int_{\mathbb{R}^d} y_1 \nu^{(2)}(dy|t) = \left( \int_{\mathbb{R}^d} y \nu^{(2)}(dy|t) \right)_1 \leq 0.
\]

This proves part (d) of the theorem. The proof of the theorem is complete. \[\blacksquare\]

We now proceed with the proof of Proposition 7.4.1, which is the Laplace principle upper bound with rate function \( I_x \) defined in Theorem 7.2.3. The proposition assumes Condition 7.2.1 on \( H^{(i)}(x, \alpha) \) and \( \mu^{(i)}(dy|x) \), which is identical to Condition 6.2.1 on the analogous quantities in Chapter 6. Accordingly, under Condition 7.2.1, for \( i = 1, 2 \) \( H^{(i)}(x, \alpha) \) and \( L^{(i)}(x, \beta) \) satisfy the same properties as \( H(x, \alpha) \) and \( L(x, \alpha) \) in Lemma 6.2.3. Instead of repeating the facts here, we will refer to Lemma 6.2.3 as needed.

Let us recall the discussion immediately after the statement of Proposition 7.4.1. For each \( x \in \mathbb{R}^d \) it suffices to prove

\[
\liminf_{n \to \infty} W^n(x) \geq \inf_{\varphi \in C([0,1], \mathbb{R}^d)} \left\{ I_x(\varphi) + h(\varphi) \right\}.
\]

We will carry this out using the representation formula for \( W^n(x) \) given in (7.14), recalling that \( \{ W^n(x) \} \) is a convergent subsequence. For \( \varepsilon > 0 \) we have chosen a sequence \( \{ \nu^n \} \) of admissible control measures satisfying for each \( n \)

\[
V^n(x) + \varepsilon \geq \hat{E}_x \left\{ \int_{[0,1]} R(\nu^n(\cdot;|t)\|\mu(\cdot|X^n(t))) \, dt + h(X^n) \right\}. \tag{7.36}
\]

Since from its definition \( |W^n(x)| \leq \|h\|_\infty \), the boundedness condition in Theorem 7.4.4 is satisfied; namely,

\[
\sup_{n \in \mathbb{N}} \hat{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot)\|\mu(\cdot|X^n_j)) \right\} < \infty.
\]

Hence along some subsequence of \( n \in \mathbb{N} \)

\[
(\nu^n, \nu^{(1)n}, \nu^{(2)n}, \gamma^{(1)n}, \gamma^{(2)n}, X^n, \tilde{X}^n) \xrightarrow{\mathbb{P}} (\nu, \nu^{(1)}, \nu^{(2)}, \gamma^{(1)}, \gamma^{(2)}, X, \tilde{X}).
\]

By the Skorohod Representation Theorem we can assume that the convergence in this display occurs w.p.1. In the evaluation of the limit inferior of \( W^n(x) \), we will make use of part (f) of Lemma 1.4.3. This implies that if \( \theta \) is a probability measure on \([0,1] \) and \( \sigma(dy|t) \) and \( \tau(dy|t) \) are stochastic kernels on \( \mathbb{R}^d \) given \([0,1] \), then

\[
\int_0^1 R(\sigma(\cdot|t)\|\tau(\cdot|t)) \theta(dt) = R(\sigma(dy|t) \otimes \theta(dt)\|\tau(dy|t) \otimes \theta(dt)).
\]

In order to apply this formula in the present context, we will have to normalize various substochastic kernels to be stochastic kernels.

The limit inferior of \( W^n(x) \) will now be evaluated along the subsequence of \( n \in \mathbb{N} \) for which

\[
(\nu^n, \nu^{(1)n}, \nu^{(2)n}, \gamma^{(1)n}, \gamma^{(2)n}, X^n, \tilde{X}^n) \longrightarrow (\nu, \nu^{(1)}, \nu^{(2)}, \gamma^{(1)}, \gamma^{(2)}, X, \tilde{X}).
\]
The integrands that will appear in this evaluation are either relative entropies or $L^{(i)}(x, \cdot)$. Although the arguments of these integrands may include fractions whose denominators are 0, in every case this occurs only on a set of measure 0. However, rather than deal with sets of measure 0, it is more convenient to define such integrands as $\infty$. This causes no difficulty because the integrands are nonnegative and because we follow the standard convention that $0 \cdot \infty = 0$. The evaluation of the limit inferior of $W^n(x)$ is carried out in the following string of inequalities, each line of which will be explained afterwards:

\[
\lim_{n \to \infty} \inf_{\varepsilon} W^n(x) + \varepsilon \\
= \lim_{n \to \infty} \inf_{\varepsilon} V^n(x) + \varepsilon \\
\geq \lim_{n \to \infty} \inf_{\varepsilon} E_x \left\{ \int_0^1 R \left( \nu^n(\cdot \mid t) \| \mu_{\cdot \mid X^n(t)} \right) dt + h(X^n) \right\} \\
= \lim_{n \to \infty} \inf_{\varepsilon} E_x \left\{ \int_0^1 R \left( \nu^n(\cdot \mid t) \| \mu^{(1)}_{\cdot \mid X^n(t)} \right) \gamma^{(1), n}(dt) \\
+ \int_0^1 R \left( \nu^{(2)}(\cdot \mid t) \| \mu^{(2)}_{\cdot \mid X^n(t)} \right) \gamma^{(2), n}(dt) + h(\tilde{X}^n) \right\} \\
= \lim_{n \to \infty} \inf_{\varepsilon} E_x \left\{ R \left( \frac{\gamma^{(1), n}(0, 1)}{\gamma^{(1), n}(0, 1)} \right) \left( \mu^{(1)}(dy \mid \tilde{X}^n(t)) \otimes \frac{\gamma^{(1), n}(dt)}{\gamma^{(1), n}(0, 1)} \right) \gamma^{(1), n}([0, 1]) \\
+ R \left( \frac{\gamma^{(2), n}(0, 1)}{\gamma^{(2), n}(0, 1)} \right) \left( \mu^{(2)}(dy \mid \tilde{X}^n(t)) \otimes \frac{\gamma^{(2), n}(dt)}{\gamma^{(2), n}(0, 1)} \right) \gamma^{(2), n}([0, 1]) + h(\tilde{X}^n) \right\} \\
= \inf_{\varepsilon} E_x \left\{ \int_0^1 R \left( \frac{\mu^{(1)}(dy \mid \tilde{X}^n(t))}{\gamma^{(1)}(t)} \right) \tilde{\gamma}^{(1)}(t) dt \\
+ \int_0^1 R \left( \frac{\mu^{(2)}(dy \mid \tilde{X}^n(t))}{\gamma^{(2)}(t)} \right) \tilde{\gamma}^{(2)}(t) dt + h(\tilde{X}) \right\} \\
\geq \inf_{\varepsilon} E_x \left\{ \int_0^1 L^{(1)} \left( \tilde{X}(t), \frac{1}{\gamma^{(1)}(t)} \int_{\mathbb{R}^d} y \nu^{(1)}(dy \mid t) \right) \tilde{\gamma}^{(1)}(t) \\
+ L^{(2)} \left( \tilde{X}(t), \frac{1}{\gamma^{(2)}(t)} \int_{\mathbb{R}^d} y \nu^{(2)}(dy \mid t) \right) \tilde{\gamma}^{(2)}(t) dt + h(\tilde{X}) \right\}.
\]

The first three lines of this display are consequences of the representation formula $W^n(x) = V^n(x)$ and (7.36). By definition $\gamma^{(1), n} + \gamma^{(2), n}$ equals Lebesgue measure $\lambda$, and on the support of $\gamma^{(i), n}$ the one-component $(\tilde{X}^n(t))_1$ is nonpositive for $i = 1$ and is positive for $i = 2$. Since $\mu(dy \mid \tilde{X}^n(t))$ equals $\mu^{(1)}(dy \mid \tilde{X}^n(t))$ or $\mu^{(2)}(dy \mid \tilde{X}^n(t))$ depending upon whether $(\tilde{X}^n(t))_1 \leq 0$ or $(\tilde{X}^n(t))_1 > 0$, we obtain lines four and five of the display. Lines six and seven follow from part (f) of Lemma 1.4.3 and the formula $\nu^{(i), n}(dy \times dt) = \nu^{(i), n}(dy \mid t) \otimes \gamma^{(i), n}(dt)$. The justification of the next two lines is, with one important exception, almost the same as the analogous justification in the previous chapter, where
we treated continuous statistics. The exception arises because of the more complicated
dependence of the various measures upon the parameter \( n \). With probability 1 \( \nu^{(i), n} \Longrightarrow \nu^{(i)} \), and since \( \lim_{n \rightarrow \infty} \gamma^{(i), n}([0, 1]) = \gamma^{(i)}([0, 1]) \)
$$
\frac{1}{\gamma^{(i), n}([0, 1])} \nu^{(i), n} \Longrightarrow \frac{1}{\gamma^{(i)}([0, 1])} \nu^{(i)}
$$
whenever \( \gamma^{(i)}([0, 1]) > 0 \). Furthermore, for \( i = 1, 2 \) the function mapping \( x \in \mathbb{R}^d \rightarrow \mu^{(i)}(\cdot|x) \in \mathcal{P}(\mathbb{R}^d) \) is continuous [Condition 7.2.1 (b)] and w.p.1 the sequence \( \{X^n, n \in \mathbb{N}\} \) converges uniformly on \([0, 1]\) to the continuous process \( \bar{X} \). Hence w.p.1, for each \( t \in [0, 1] \)
$$
\mu^{(i)}(dy|X^n(t)) \Longrightarrow \mu^{(i)}(dy|\bar{X}(t))
$$
and whenever \( \gamma^{(i)}([0, 1]) > 0 \)
$$
\frac{1}{\gamma^{(i), n}([0, 1])} \gamma^{(i), n} \Longrightarrow \frac{1}{\gamma^{(i)}([0, 1])} \gamma^{(i)}.
$$
Hence whenever \( \gamma^{(i)}([0, 1]) > 0 \), Corollary A.5.10 yields w.p.1
$$
\mu^{(i)}(dy|X^n(t)) \otimes \frac{\gamma^{(i), n}(dt)}{\gamma^{(i), n}([0, 1])} \Longrightarrow \mu^{(i)}(dy|\bar{X}(t)) \otimes \frac{\gamma^{(i)}(dt)}{\gamma^{(i)}([0, 1])}.
$$
The lower semicontinuity of \( R(\cdot||\cdot) \) yields w.p.1
$$
\liminf_{n \rightarrow \infty} R \left( \frac{\nu^{(i), n} \otimes \mu^{(i), n}}{\gamma^{(i), n}(0, 1)} \right) \rightarrow R \left( \frac{\nu^{(i)}(dy \otimes dt)}{\gamma^{(i)}([0, 1])} \right)
$$
whenever \( \gamma^{(i)}([0, 1]) > 0 \). Finally, since \( h \) is continuous on \( C([0, 1] : \mathbb{R}^d) \) and w.p.1 the sequence \( \{X^n\} \) converges uniformly on \([0, 1]\) to \( \bar{X} \), we also have w.p.1
$$
\lim_{n \rightarrow \infty} h(X^n) = h(\bar{X}).
$$
Lines eight and nine in (7.37) now follow from Fatou’s Lemma. According to parts (b) and (c) of Lemma 7.4.3, whenever \( \gamma^{(i)}([0, 1]) > 0 \) we have the probability–1 decomposition
$$
\frac{\nu^{(i)}(dy \otimes dt)}{\gamma^{(i)}([0, 1])} = \left( 1_{\{s \in [0, 1] : \gamma^{(i)}(s) > 0\}}(t) \frac{\nu^{(i)}(dy|t)}{\gamma^{(i)}(t)} \right) \otimes \frac{\gamma^{(i)}(t dt)}{\gamma^{(i)}([0, 1])}.
$$
Another application of part (f) of Lemma 1.4.3 yields lines ten and eleven of the display. Finally, lines twelve and thirteen are implied by the lower bound involving the relative entropy and \( L^{(i)}(x, \beta) \) given in part (h) of Lemma 6.2.3.

Returning to the last two lines of (7.37), we now claim that w.p.1 we have a.s. for \( t \in [0, 1] \)
$$
L^{(1)} \left( \bar{X}(t), \frac{1}{\gamma^{(1)}(t)} \int_{\mathbb{R}^d} y \nu^{(1)}(dy|t) \right) \gamma^{(1)}(t) + L^{(2)} \left( \bar{X}(t), \frac{1}{\gamma^{(2)}(t)} \int_{\mathbb{R}^d} y \nu^{(2)}(dy|t) \right) \gamma^{(2)}(t)
$$
\[ \geq \tilde{L}(\tilde{X}(t), \dot{X}(t)). \] (7.38)

Indeed, if \((\tilde{X}(t))_1 < 0\), then \(\dot{\gamma}^{(1)}(t) = 1\), \(\dot{\gamma}^{(2)}(t) = 0\), and \(\nu^{(2)}(dy|t) = 0\) [Theorem 7.4.4 (c)]. The inequality (7.38) follows from (7.30) and from the fact that if \((\tilde{X}(t))_1 < 0\), then

\[
\tilde{L}(\tilde{X}(t), \dot{X}(t)) = L^{(1)}(\tilde{X}(t), \dot{X}(t)).
\]

If \((\tilde{X}(t))_1 > 0\), then one proves (7.38) similarly.

We now prove (7.38) when \((\tilde{X}(t))_1 = 0\). Details are given only in the case where \(\dot{\gamma}^{(1)}(t) > 0\) and \(\dot{\gamma}^{(2)}(t) > 0\); the case where either \(\dot{\gamma}^{(1)}(t) = 0\) or \(\dot{\gamma}^{(2)}(t) = 0\) is handled in an analogous way. With probability 1, a.s. for \(t \in [0, 1]\) the constraint \(\dot{\gamma}^{(1)}(t) + \dot{\gamma}^{(2)}(t) = 1\) in equation (7.27) is satisfied, and formulas (7.28) and (7.29) imply

\[
\left( \frac{1}{\dot{\gamma}^{(1)}(t)} \int_{\mathbb{R}^d} y \nu^{(1)}(dy|t) \right)_1 \geq 0 \quad \text{and} \quad \left( \frac{1}{\dot{\gamma}^{(1)}(t)} \int_{\mathbb{R}^d} y \nu^{(2)}(dy|t) \right)_1 \leq 0.
\]

Since by (7.30)

\[
\dot{\gamma}^{(1)}(t) \left( \frac{1}{\dot{\gamma}^{(1)}(t)} \int_{\mathbb{R}^d} y \nu^{(1)}(dy|t) \right) + \dot{\gamma}^{(2)}(t) \left( \frac{1}{\dot{\gamma}^{(2)}(t)} \int_{\mathbb{R}^d} y \nu^{(2)}(dy|t) \right) = \int_{\mathbb{R}^d} y \nu^{(1)}(dy|t) + \int_{\mathbb{R}^d} y \nu^{(2)}(dy|t) = \dot{X}(t),
\]

the lower bound (7.38) follows from the definition of \(L^{(0)}(x, \tilde{X}(t), \dot{X}(t))\).

This completes the proof of (7.38).

Combining (7.37) and (7.38) and sending \(\varepsilon \to 0\), we have proved that every convergent subsequence of the original sequence \(\{W^n(x), n \in \mathbb{N}\}\) has a subsequence satisfying

\[
\liminf_{n \to \infty} W^n(x) = \liminf_{n \to \infty} V^n(x)
\]

\[
\geq \mathbb{E}_x \left\{ \int_0^1 \tilde{L}(\tilde{X}(t), \dot{X}(t)) \, dt + h(\tilde{X}) \right\}
\]

\[
\geq \inf_{\varphi \in C([0,1];\mathbb{R}^d)} \{ I_x(\varphi) + h(\varphi) \}.
\]

An argument by contradiction establishes this lower limit for the entire sequence \(\{W^n(x), n \in \mathbb{N}\}\). We have completed the proof of Proposition 7.4.1, which is the Laplace principle upper bound for the random walk model with discontinuous statistics.

### 7.5 Proof of the Laplace Principle Lower Bound

In this section the following proposition will be proved.
Proposition 7.5.1. We assume Conditions 7.2.1 and 7.2.2. For \( x \in \mathbb{R}^d \), we define \( I_x \) as in Theorem 7.2.3. Then for all bounded continuous functions \( h \) mapping \( C([0, 1]: \mathbb{R}^d) \) into \( \mathbb{R} \) we have the Laplace principle lower bound

\[
\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{E}_x \{ \exp \{ -n h(X^n) \} \} \geq - \inf_{\varphi \in C([0, 1]: \mathbb{R}^d)} \{ I_x(\varphi) + h(\varphi) \}.
\]

Condition 7.2.2 states that the sets \( \text{ri} (\text{conv} S_{\mu^{(i)}(\cdot|x)}) \) are independent of \( x \in \mathbb{R}^d \) and \( i = 1, 2 \), that \( 0 \in \Sigma \equiv \text{ri} (\text{conv} S_{\mu^{(i)}(\cdot|x)}) \), and that \( \Sigma \) is not a subset of \( \partial \equiv \{ \beta \in \mathbb{R}^d : \beta_1 = 0 \} \). As in the proof of the Laplace principle lower bound in Section 6.5, we will modify Condition 7.2.2, without loss of generality, to include the condition that the interior of \( \Sigma \) relative to \( \mathbb{R}^d \) is nonempty; i.e., that \( \Sigma \) is a nonempty open convex subset of \( \mathbb{R}^d \). Hence we will assume that for each \( x \in \mathbb{R}^d \)

\[ 0 \in \Sigma = \text{int}(\text{conv} S_{\mu^{(i)}(\cdot|x)}). \]

As in part (a) of Lemma 6.2.3, for \( i = 1, 2 \) the functions \( H^{(i)}(x, \alpha) \) are continuous on \( \mathbb{R}^d \times \mathbb{R}^d \). According to part (a) of the next lemma, under our conditions this continuity property of \( H^{(i)}(x, \alpha) \) carries over to the continuity of the respective Legendre-Fenchel transform \( L^{(i)}(x, \beta) \) on \( \mathbb{R}^d \times \Sigma \).

Lemma 7.5.2. Under Condition 7.2.1 and part (a) of Condition 7.2.2 the functions \( L^{(i)}(x, \beta), i = 1, 2 \), have the following properties.

(a) For each \( x \in \mathbb{R}^d \) the set \( \text{ri}(\text{dom} L^{(i)}(x, \cdot)) = \text{ri}(\text{conv} S_{\mu^{(i)}(\cdot|x)}) \) equals \( \Sigma \).

(b) \( L^{(i)}(x, \beta) \) is a continuous function of \( (x, \beta) \in \mathbb{R}^d \times \Sigma \).

Comments on the Proof. (a) Part (d) of Lemma 6.2.3 states that for each \( x \in \mathbb{R}^d \) \( \text{ri}(\text{dom} L^{(i)}(x, \cdot)) = \text{ri}(\text{conv} S_{\mu^{(i)}(\cdot|x)}) \). According to part (a) of Condition 7.2.2 this set equals \( \Sigma \).

(b) This is proved in Section C.8. ■

Another property that we need is the continuity of \( L^{(0)}(x, \beta) \) on \( \mathbb{R}^d \times (\Sigma \cap \partial) \). This is stated in part (c) of the next lemma, which also gives several other useful facts.

Lemma 7.5.3. Under Conditions 7.2.1 and 7.2.2, \( L^{(0)}(x, \beta) \) and \( \tilde{L}(x, \beta) \) have the following properties.

(a) For \( x \) and \( \beta \) in \( \mathbb{R}^d \)

\[
L^{(0)}(x, \beta) \leq L^{(1)}(x, \beta) \text{ if } \beta_1 \geq 0 \text{ and } L^{(0)}(x, \beta) \leq L^{(2)}(x, \beta) \text{ if } \beta_1 \leq 0.
\]

(b) For each \( x \in \mathbb{R}^d \) \( \text{ri}(\text{dom} L^{(0)}(x, \cdot)) \) equals \( \Sigma \).

(c) \( L^{(0)}(x, \beta) \) is a continuous function of \( (x, \beta) \in \mathbb{R}^d \times (\Sigma \cap \partial) \).

(d) For each \( x \in \mathbb{R}^d \) \( \text{ri}(\text{dom} \tilde{L}(x, \cdot)) \) equals \( \Sigma \).
Proof. (a) Suppose that $\beta \in \mathbb{R}^d$ satisfies $\beta_1 \geq 0$. If in the definition of $L^{(0)}(x, \beta)$ we take $\rho^{(1)} = 1$, $\rho^{(2)} = 0$, $\beta^{(1)} = \beta$, and $\beta^{(2)} = 0$, then it follows that $L^{(0)}(x, \beta) \leq L^{(1)}(x, \beta)$. Similarly, if $\beta \in \mathbb{R}^d$ satisfies $\beta_1 \leq 0$, then $L^{(0)}(x, \beta) \leq L^{(2)}(x, \beta)$.

(b) Part (a) implies that if $L^{(1)}(x, \beta) < \infty$ and $L^{(2)}(x, \beta) < \infty$, then $L^{(0)}(x, \beta) < \infty$. Hence by part (a) of Lemma 7.5.2, $L^{(0)}(x, \beta) < \infty$ for $(x, \beta) \in \mathbb{R}^d \times \Sigma$. On the other hand, if $(x, \beta) \in \mathbb{R}^d \times (\text{cl} \Sigma)^c$, then for any $\rho^{(1)}$, $\rho^{(2)}$, $\beta^{(1)}$, and $\beta^{(2)}$ satisfying the constraints (7.6)–(7.8), either $\rho^{(1)} > 0$ and $\beta^{(1)} \in (\text{cl} \Sigma)^c$ or $\rho^{(2)} > 0$ and $\beta^{(2)} \in (\text{cl} \Sigma)^c$. Thus $L^{(0)}(x, \beta) = \infty$.

(c) In part (a) of Lemma 7.5.5 we will prove that under Conditions 7.2.1 and 7.2.2 $L^{(0)}(x, \beta)$ is a lower semicontinuous function of $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$. In order to prove part (c) of the present lemma, we show that $L^{(0)}(x, \beta)$ is an upper semicontinuous function of $(x, \beta) \in \mathbb{R}^d \times (\Sigma \cap \partial)$.

Fix $(x, \beta) \in \mathbb{R}^d \times (\Sigma \cap \partial)$ and let $\varepsilon > 0$ be given. Condition 7.2.2 implies that $0 \in \Sigma \cap (\text{int} L^{(i)})$ for $i = 1, 2$. We claim that there exist $\rho^{(1)}$, $\rho^{(2)}$, $\beta^{(1)}$, and $\beta^{(2)}$ satisfying

$$\rho^{(1)} > 0, \rho^{(2)} > 0, (\beta^{(1)})_1 > 0, (\beta^{(2)})_1 < 0, \beta^{(1)} \in \Sigma, \beta^{(2)} \in \Sigma,$$

and

$$\rho^{(1)} + \rho^{(2)} = 1, \rho^{(1)} \beta^{(1)} + \rho^{(2)} \beta^{(2)} = \beta, \quad (7.39)$$

and

$$\rho^{(1)} L^{(1)}(x, \beta^{(1)}) + \rho^{(2)} L^{(2)}(x, \beta^{(2)}) \leq L^{(0)}(x, \beta) + \varepsilon. \quad (7.40)$$

To see this, one starts with $\rho^{(1)}$, $\rho^{(2)}$, $\beta^{(1)}$, and $\beta^{(2)}$ satisfying

$$\rho^{(1)} \geq 0, \rho^{(2)} \geq 0, (\beta^{(1)})_1 \geq 0, (\beta^{(2)})_1 \leq 0, \beta^{(1)} \in \text{cl} \Sigma, \beta^{(2)} \in \text{cl} \Sigma,$$

and

$$\rho^{(1)} + \rho^{(2)} = 1, \rho^{(1)} \beta^{(1)} + \rho^{(2)} \beta^{(2)} = \beta,$$

and

$$\rho^{(1)} L^{(1)}(x, \beta^{(1)}) + \rho^{(2)} L^{(2)}(x, \beta^{(2)}) \leq L^{(0)}(x, \beta) + \varepsilon/4.$$

First, if one of the $\rho^{(i)}$'s equals 0, then set the corresponding $\beta^{(i)} = 0$ and perturb $\rho^{(1)}$ and $\rho^{(2)}$ so that both are positive and (7.39) and (7.40) hold with $\varepsilon$ replaced by $\varepsilon/3$. Second, if $\beta^{(1)}$ or $\beta^{(2)}$ lies in $\partial \Sigma$, then use the fact that $\beta \in \Sigma, \beta_1 = 0$, the convexity of $\Sigma$, and a continuity property of $L^{(1)}(x, \cdot)$ and $L^{(2)}(x, \cdot)$ [Theorem D.2.2 (c)] to perturb $\beta^{(1)}$ and $\beta^{(2)}$ into $\Sigma$ so that $(\beta^{(1)})_1 \geq 0, (\beta^{(2)})_1 \leq 0$, and (7.39) and (7.40) continue to hold with $\varepsilon/3$ replaced by $\varepsilon/2$. Finally, if $(\beta^{(1)})_1 = 0$ or $(\beta^{(2)})_1 = 0$, then use the continuity of $L^{(i)}(x, \cdot)$ on $\Sigma$ [Lemma 7.5.2 (b)] to perturb $\beta^{(1)}$ and $\beta^{(2)}$ inside $\Sigma$ so that $(\beta^{(1)})_1 > 0, (\beta^{(2)})_1 < 0$, and (7.39) and (7.40) hold.

Now let $(\xi, v)$ be any point in $\mathbb{R}^d \times \Sigma$ and let $v^{(1)}$ and $v^{(2)}$ be any points in $\mathbb{R}^d$ satisfying

$$(v^{(1)})_1 \geq 0, (v^{(2)})_1 \leq 0, \rho^{(1)} v^{(1)} + \rho^{(2)} v^{(2)} = v.$$ 

Then

$$L^{(0)}(\xi, v) - L^{(0)}(x, \beta) \leq \rho^{(1)} L^{(1)}(\xi, v^{(1)}) + \rho^{(2)} L^{(2)}(\xi, v^{(2)}) - \rho^{(1)} L^{(1)}(x, \beta^{(1)}) - \rho^{(2)} L^{(2)}(x, \beta^{(2)}) + \varepsilon. \quad (7.41)$$
We now define

\[ v^{(1)} = \beta^{(1)} \quad \text{and} \quad v^{(2)} = \frac{v - \rho^{(1)} v^{(1)}}{\rho^{(2)}} = \frac{v - \rho^{(1)} \beta^{(1)}}{\rho^{(2)}}. \]

Then \((v^{(1)})_1 = (\beta^{(1)})_1 > 0\), \(v^{(1)} \in \Sigma\),

\[ (v^{(2)})_1 = \frac{(v - \beta)_1}{\rho^{(2)}} + \frac{(\beta - \rho^{(1)} \beta^{(1)})_1}{\rho^{(2)}} = \frac{(v - \beta)_1}{\rho^{(2)}} + (\beta^{(2)})_1 \leq \frac{v - \beta_1}{\rho^{(2)}} + (\beta^{(2)})_1, \]

and

\[ \|v^{(2)} - \beta^{(2)}\| \leq \frac{1}{\rho^{(2)}} \|v - \beta\|. \]

Since \((\beta^{(2)})_1 < 0\) and \(\beta^{(2)}\) lies in the open set \(\Sigma\), we can guarantee that \((v^{(2)})_1 < 0\) and that \(v^{(2)}\) lies in \(\Sigma\) by making \(\|v - \beta\|\) sufficiently small. We now insert \(v^{(1)}\) and \(v^{(2)}\) into (7.41). The continuity of \(L^{(1)}(\cdot, \cdot)\) and \(L^{(2)}(\cdot, \cdot)\) on \(\mathbb{R}^d \times \Sigma\) implies that if \(\|\langle \xi, v \rangle - (x, \beta)\|\) is sufficiently small, then

\[ L^{(0)}(\xi, v) - L^{(0)}(x, \beta) \leq 2\varepsilon. \]

This proves that \(L^{(0)}(\cdot, \cdot)\) is upper semicontinuous on \(\mathbb{R}^d \times \Sigma\) and completes the proof of part (c).

(d) This follows from part (a) of Lemma 7.5.2 and part (b) of the present lemma.

In order to prove the Laplace principle lower bound, we must show that the functions

\[ W^n(x) = -\frac{1}{n} \log E\{\exp[-n h(X^n)]\} \]

satisfy

\[ \limsup_{n \to \infty} W^n(x) \leq \inf_{\varphi \in C([0,1]:\mathbb{R}^d)} \left\{ I_x(\varphi) + h(\varphi) \right\}. \]

As in the corresponding proof in the previous chapter, the strategy for proving this is to follow the Approximation Procedure 6.4.3. For a class \(\mathcal{N}_0\) of sufficiently well-behaved functions \(\psi^*\), we will first show for each \(x \in \mathbb{R}^d\)

\[ \limsup_{n \to \infty} W^n(x) \leq I_x(\psi^*) + h(\psi^*), \]

which is the first requirement of the approximation procedure. The class \(\mathcal{N}_0\) is a slight restriction of the class \(\mathcal{N}\) introduced in Section 6.5.

**Identification of the class \(\mathcal{N}_0\).** \(\mathcal{N}_0\) consists of all functions \(\psi^* \in C([0,1]:\mathbb{R}^d)\) satisfying the following conditions:

(a) \(\dot{\psi}^*(t)\) is piecewise constant with only finitely many jumps in the interval \((0,1)\).

(b) \(\dot{\psi}^*(t) \in \Sigma\) for all \(t \in [0,1)\).

(c) Either \((\psi^*(t))_1 \neq 0\) or \((\psi^*(t))_1 = 0\) on each interval of constancy of \(\dot{\psi}^*\).
In order to have $\hat{\psi}^*(t)$ defined for all $t \in [0, 1)$, we replace the almost everywhere defined function $\hat{\psi}$ by its right continuous regularization. For the proof of the Laplace principle lower bound, we must show that $\mathcal{N}_0$ also satisfies the second requirement of the Approximation Procedure 6.4.3. This is carried out in the following lemma. The proof is postponed until the end of the section.

**Lemma 7.5.4.** Assume Conditions 7.2.1 and 7.2.2. For $x \in \mathbb{R}^d$, let $\psi \in C([0, 1] : \mathbb{R}^d)$ satisfy $I_x(\psi) < \infty$. Then for each $\delta > 0$, there exists $\psi^* \in \mathcal{N}_0$ such that

$$
\|\psi^* - \psi\|_\infty \leq \delta \quad \text{and} \quad I_x(\psi^*) \leq I_x(\psi) + \delta,
$$

and for each $k \in \{1, 2, \ldots, r\}$ either $(\psi^*(t))_1 \neq 0$ for all $t \in (t_k, t_{k+1})$ or $(\psi^*(t))_1 = 0$ for all $t \in (t_k, t_{k+1})$. The intervals $(t_k, t_{k+1}), k = 1, 2, \ldots, r$, denote the interiors of the successive intervals on which $\psi^*$ is constant.

Let us assume that we have proved

$$
\limsup_{n \to \infty} W^n(x) \leq I_x(\psi^*) + h(\psi^*) \tag{7.42}
$$

for $\psi^* \in \mathcal{N}_0$ and that we have proved this lemma. We then complete the proof of the Laplace principle upper bound exactly as in Section 6.4. Indeed, given $\varepsilon > 0$, we choose $\psi \in C([0, 1] : \mathbb{R}^d)$ to satisfy

$$
I_x(\psi) + h(\psi) \leq \inf_{\varphi \in C([0, 1] : \mathbb{R}^d)} \{I_x(\varphi) + h(\varphi)\} + \varepsilon < \infty.
$$

Since $h$ is bounded, this implies that $I_x(\psi) < \infty$. Since $h$ is continuous, we can pick $\psi^* \in \mathcal{N}_0$ so that

$$
h(\psi^*) \leq h(\psi) + \varepsilon \quad \text{and} \quad I_x(\psi^*) \leq I_x(\psi) + \varepsilon.
$$

It follows that

$$
\liminf_{n \to \infty} \frac{1}{n} \log E_x \left\{ \exp \left\{ -n h(X^n) \right\} \right\} 
\geq -I_x(\psi^*) - h(\psi^*) 
\geq -I_x(\psi) - h(\psi) - 2\varepsilon 
\geq - \inf_{\varphi \in C([0, 1] : \mathbb{R}^d)} \{I_x(\varphi) + h(\varphi)\} - 3\varepsilon.
$$

Sending $\varepsilon \to 0$ yields the Laplace principle lower bound.

Let $\psi^* \in \mathcal{N}_0$ satisfy $I_x(\psi^*) < \infty$. We prove (7.42) by using the same representation formula used in the proof of the Laplace principle upper bound; namely,

$$
W^n(x) = V^n(x) = \inf_{\{\nu^n_T\}} \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_T(\cdot) \mu(\cdot | \tilde{X}^n_T)) \right\}.
$$
7.5. PROOF OF THE LAPLACE PRINCIPLE LOWER BOUND

The infimum is taken over all admissible control sequences \( \{ \nu_j^n \} \), and as the first argument of the relative entropy \( \nu_j^n (\cdot) = \nu_j^n (\cdot | \bar{X}_0^n, \bar{X}_1^n, \ldots, \bar{X}_n^n) \). The process \( \bar{X}_n \), defined in formula (5.6), is the piecewise linear interpolation of the controlled random vectors \( \{ \bar{X}_j^n \} \) defined in equations (5.1).

We now use \( \psi^* \) to design admissible control sequences \( \{ \nu_j^n \} \) and controlled random vectors \( \{ \bar{X}_j^n \} \) that can be used to prove (7.42). For each \( k \in \{ 0, 1, \ldots, r \} \), define \( \beta_k = \psi^*(t) \), where \( t \) is any point in the interior of the interval of constancy \( (t_k, t_{k+1}) \) of this derivative. Since \( \psi^* \in \mathcal{N}_0 \), \( \beta_k \) lies in the open set \( \Sigma \). If \( (\psi^*(t))_{1} \neq 0 \) for all \( t \in (t_k, t_{k+1}) \), then we define \( \beta_k^{(1)} = \beta_k^{(2)} = \beta_k \). If \( (\psi^*(t))_{1} = 0 \) for all \( t \in (t_k, t_{k+1}) \) and hence \( (\beta_k) = 0 \), then we must prescribe a control that will yield a running cost close to \( L^{(0)}(\psi^*(t), \beta_k) \).

To do this, we make use of the facts that \( \text{ri}(\text{dom} \ L^{(1)}(x, \cdot)) \) and \( \text{ri}(\text{dom} \ L^{(2)}(x, \cdot)) \) equal \( \Sigma \) for each \( x \in \mathbb{R}^d \); that \( L^{(1)}(\cdot, \cdot) \) and \( L^{(2)}(\cdot, \cdot) \) are continuous on \( \mathbb{R}^d \times \Sigma \); and that \( \Sigma \cap (\text{int} \Lambda^{(i)}) \neq \emptyset \) for \( i = 1, 2 \), where \( \Lambda^{(1)} \) and \( \Lambda^{(2)} \) denote the respective halfspaces \( \{ \beta \in \mathbb{R}^d : \beta_1 \leq 0 \} \) and \( \{ \beta \in \mathbb{R}^d : \beta_1 > 0 \} \). The first two facts are given in Lemma 7.5.2, and the third is a consequence of part (b) of Condition 7.2.2. As we showed in the proof of part (c) of Lemma 7.5.3, there exist \( \rho_k^{(1)}, \rho_k^{(2)}, \beta_k^{(1)}, \) and \( \beta_k^{(2)} \) that satisfy

\[
\rho_k^{(1)} > 0, \rho_k^{(2)} > 0, (\beta_k^{(1)}) = 0, (\beta_k^{(2)}) = 0, \beta_k^{(1)} \in \Sigma, \beta_k^{(2)} \in \Sigma, \rho_k^{(1)} + \rho_k^{(2)} = 1, \rho_k^{(1)} \beta_k^{(1)} + \rho_k^{(2)} \beta_k^{(2)} = \beta_k,
\]

(7.44)

and

\[
\rho_k^{(1)} L^{(1)}(\psi^*(t), \beta_k^{(1)}) + \rho_k^{(2)} L^{(2)}(\psi^*(t), \beta_k^{(2)}) \leq L^{(0)}(\psi^*(t), \beta_k) + \varepsilon \tag{7.45}
\]

for all \( t \in (t_k, t_{k+\lambda}) \), where \( \lambda \) is some positive number. Since \( \psi^* \) is in general not a constant function, it may not actually be the case that (7.45) can be guaranteed for all \( t \in (t_k, t_{k+1}) \) simultaneously. However, the continuity properties of \( L^{(0)} \), \( L^{(1)} \), and \( L^{(2)} \) imply that if necessary we can add points to the original subdivision \( 0 = t_1 < t_2 < \ldots < t_{r+1} = 1 \) of \( (0, 1) \) and obtain a new subdivision \( 0 = t_0 < t_1 < \ldots < t_{\ell+1} = 1 \) of \( (0, 1) \) with the following property: if \( (\psi^*(t))_{1} = 0 \) for all \( t \in (t_k, t_{k+1}) \), then (7.45) holds for all \( t \in (t_k, t_{k+1}) \). For simplicity we will retain the same notation for the original subdivision and the refined subdivision.

For each \( n \in \mathbb{N} \) we now define the admissible control sequence \( \{ \nu_j^n \} \) by adapting the construction given in Procedure 6.4.1. For \( j \in \{ 0, 1, \ldots, n - 1 \} \), points in \( (\mathbb{R}^d)^{j+1} \) are denoted by \( \xi_j = (\xi_0, \xi_1, \ldots, \xi_{j}) \), where each coordinate \( \xi_j \) is an element of \( \mathbb{R}^d \). The admissible control \( \nu_j^n \) applied at time \( j \) is a stochastic kernel on \( \mathbb{R}^d \) given \( (\mathbb{R}^d)^{j+1} \) which we denote by \( \nu_j^n(dy|\xi_j) \). We first assume that \( j/n \in [t_k, t_{k+1}] \) and that the one-component of \( \xi_j \) satisfies \( (\xi_j)_{1} \leq 0 \). Then part (g) of Lemma 6.2.3 guarantees that there exists a stochastic kernel \( \gamma_j^n(dy|\xi_j) \) on \( \mathbb{R}^d \) given \( \mathbb{R}^d \) satisfying

\[
R(\gamma_j^n (\cdot|\xi_j) \| \mu^{(1)} (\cdot|\xi_j) ) = L^{(1)}(\xi_j, \beta_k^{(1)}) \quad \text{and} \quad \int_{\mathbb{R}^d} y \gamma_j^n(dy|\xi_j) = \beta_k^{(1)}.
\]

If \( j/n \in [t_k, t_{k+1}] \) and, on the other hand, \( (\xi_j)_{1} > 0 \), then there exists a stochastic kernel \( \gamma_j^n(dy|\xi_j) \) on \( \mathbb{R}^d \) given \( \mathbb{R}^d \) satisfying

\[
R(\gamma_j^n (\cdot|\xi_j) \| \mu^{(2)} (\cdot|\xi_j) ) = L^{(2)}(\xi_j, \beta_k^{(2)}) \quad \text{and} \quad \int_{\mathbb{R}^d} y \gamma_j^n(dy|\xi_j) = \beta_k^{(2)}.
\]
Now fix \( \delta > 0 \); the definition of \( \nu_j^n(dy|\xi) \) depends on whether \( \max_{i=0,1,\ldots,j} \|\xi_i-\psi(i/n)\| \leq \delta \) or \( \max_{i=0,1,\ldots,j} \|\xi_i-\psi(i/n)\| > \delta \). Specifically, for \( \xi \in (\mathbb{R}^{d})^{j+1} \) we define

\[
\nu_j^n(dy|\xi) = \begin{cases} 
\gamma_j^n(dy|\xi_j) & \text{if } \max_{i=0,1,\ldots,j} \|\xi_i-\psi(i/n)\| \leq \delta \\
\mu(dy|\xi_j) & \text{if } \max_{i=0,1,\ldots,j} \|\xi_i-\psi(i/n)\| > \delta.
\end{cases}
\]

Having defined the admissible control sequence \( \{\nu_j^n\} \), we set \( \bar{X}_0^n = x \), where \( x \) is a fixed point in \( \mathbb{R}^d \), and take \( \bar{Y}_j^n \) to be a random vector with conditional distribution

\[
P_x\{\bar{Y}_j^n \in dy|\bar{X}_0^n, \bar{X}_1^n, \ldots, \bar{X}_j^n\} = \nu_j^n(dy|\bar{X}_0^n, \bar{X}_1^n, \ldots, \bar{X}_j^n).
\]

We then define

\[
\bar{X}_{j+1}^n = \bar{X}_j^n + \frac{1}{n} \bar{Y}_j^n.
\]

In terms of the stopping time

\[
\tau^n \overset{\Delta}{=} \frac{1}{n} \left( \min \left\{ i \in \{0, 1, \ldots, n\} : \|\bar{X}_i^n - \psi(i/n)\| > \delta \right\} \right) \wedge n,
\]
the admissible controls \( \nu_j^n(\cdot) = \nu_j^n(\cdot|\bar{X}_0^n, \bar{X}_1^n, \ldots, \bar{X}_j^n) \) have the following properties. For \( j \in \{0, 1, \ldots, n\tau^n - 1\} \) satisfying \( j/n \in [t_k, t_{k+1}) \), if \( (\bar{X}_j^n)_1 \leq 0 \), then

\[
R\left(\nu_j^n(\cdot)\|\mu^{(1)}(\cdot|\bar{X}_j^n)\right) = L^{(1)}(\bar{X}_j^n, \beta_k^{(1)}) \quad \text{and} \quad \int_{\mathbb{R}^d} y \nu_j^n(dy) = \beta_k^{(1)},
\]
and if \( (\bar{X}_j^n)_1 > 0 \), then

\[
R\left(\nu_j^n(\cdot)\|\mu^{(2)}(\cdot|\bar{X}_j^n)\right) = L^{(2)}(\bar{X}_j^n, \beta_k^{(2)}) \quad \text{and} \quad \int_{\mathbb{R}^d} y \nu_j^n(dy) = \beta_k^{(2)}.
\]

For \( j \in \{n\tau^n, n\tau^n+1, \ldots, n-1\} \) \( \nu_j^n(\cdot) \) equals \( \mu(\cdot|\bar{X}_j^n) \) and thus \( R(\nu_j^n(\cdot)\|\mu(\cdot|\bar{X}_j^n)) = 0 \). This completes the definition of the admissible control sequences and the controlled random vectors.

Using the control sequence \( \{\nu_j^n\} \) just constructed, we consider the random subprobability measures \( \nu^n, \nu^{(1),n}, \nu^{(2),n}, \gamma^{(1),n}, \) and \( \gamma^{(2),n} \), the piecewise linear process \( \bar{X}^n \), and the piecewise constant process \( \bar{X}^n \), all defined as in the proof of the Laplace principle upper bound. For each \( n \in \mathbb{N} \), associated with the admissible control sequence \( \{\nu_j^n\} \) is the running cost

\[
E_x \left\{ \sum_{j=0}^{n-1} \frac{1}{n} R\left(\nu_j^n(\cdot)\|\mu(\cdot|\bar{X}_j^n)\right) \right\} = E_x \left\{ \sum_{k=1}^{r} \sum_{\{j \in \{0,1,\ldots,n\tau^n-1\} : j/n \in [t_k, t_{k+1})\}} \frac{1}{n} \left[ L^{(1)}(\bar{X}_j^n, \beta_k^{(1)}) \cdot 1\{(\bar{X}_j^n)_1 \leq 0\} + L^{(2)}(\bar{X}_j^n, \beta_k^{(2)}) \cdot 1\{(\bar{X}_j^n)_1 > 0\} \right] \right\}.
\]
The vectors \( \{ \beta_k^{(i)}, i = 1, 2; k = 1, 2, \ldots, r \} \) lie in \( \Sigma \). The boundedness of the set \( \{ \| \psi(t) \|, t \in [0, 1] \} \), the definition of the stopping times \( \tau^n \), and the continuity of \( L^{(1)}(\cdot, \cdot) \) and \( L^{(2)}(\cdot, \cdot) \) on \( \mathbb{R}^d \times \Sigma \) then imply that the sequence of running costs in the last display is bounded for \( n \in \mathbb{N} \). Since part (a) of Condition 7.2.1 gives Condition 5.3.1, the hypotheses of Theorem 7.4.4 and Proposition 5.3.8 are satisfied, and so we may apply the compactness and convergence results stated there.

For \( n \in \mathbb{N} \) we define the process \( S^n(t) = \{ S^n(t), t \in [0, 1] \} \) by

\[
S^n(t) = x + \int_{\mathbb{R}^d \times [0, 1]} y \nu^n(dy \times ds).
\]

Since \( \tau^n \) takes values in the compact set \([0, 1]\), Theorem 7.4.4, Proposition 5.3.8, and Prohorov’s Theorem imply that any subsequence of the sequence

\[
\left\{ \left( \nu^n, \nu^{(1),n}, \nu^{(2),n}, \gamma^{(1),n}, \gamma^{(2),n}, \tilde{X}^n, \tilde{X}^n, S^n, \tau^n \right), n \in \mathbb{N} \right\}
\]

is tight and that given any subsequence, there exists a probability space \((\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}_2)\) and a subsequence that satisfies

\[
\left( \nu^n, \nu^{(1),n}, \nu^{(2),n}, \gamma^{(1),n}, \gamma^{(2),n}, \tilde{X}^n, \tilde{X}^n, S^n, \tau^n \right) \overset{D}{\longrightarrow} \left( \nu, \nu^{(1)}, \nu^{(2)}, \gamma^{(1)}, \gamma^{(2)}, \tilde{X}, \tilde{X}, \tau \right).
\]

The quantity \( \nu \) is a stochastic kernel on \( \mathbb{R}^d \times [0, 1] \) given \( \tilde{\Omega} \), \( \nu^{(1)} \) and \( \nu^{(2)} \) are substochastic kernels on \( \mathbb{R}^d \times [0, 1] \) given \( \tilde{\Omega} \), \( \gamma^{(1)} \) and \( \gamma^{(2)} \) are substochastic kernels on \([0, 1] \) given \( \tilde{\Omega} \), \( \tilde{X} = \{ \tilde{X}(t), t \in [0, 1] \} \) is a random variable mapping \( \tilde{\Omega} \) into \( \mathcal{C}([0, 1] : \mathbb{R}^d) \), and \( \tau \) is a random variable mapping \( \tilde{\Omega} \) into \([0, 1] \). With probability 1, for every \( t \in [0, 1] \) \( \tilde{X} \) is related to \( \nu \) by

\[
\tilde{X}(t) = x + \int_{\mathbb{R}^d \times [0, t]} y \nu(dy \times ds).
\]

By the Skorohod Representation Theorem, we can assume that the subsequence

\[
\left\{ \left( \nu^n, \nu^{(1),n}, \nu^{(2),n}, \gamma^{(1),n}, \gamma^{(2),n}, \tilde{X}^n, \tilde{X}^n, S^n, \tau^n \right) \right\}
\]

converges w.p.1. Furthermore, Lemma 7.4.3 gives the probability–1 decompositions

\[
\nu(dy \times dt) = \nu(dy|t) \otimes dt, \nu^{(0)}(dy \times dt) = \nu^{(0)}(dy|t) \otimes dt, \text{ and } \gamma^{(i)}(dt) = \gamma^{(i)}(t) \, dt.
\]

We next prove that w.p.1 \( \tilde{X}(s) = \psi^*(s) \) for each \( k \in \{1, 2, \ldots, r + 1\} \) and all \( s \in [0, t_k \wedge \tau] \). The proof is by induction on \( k \). The equality holds for \( k = 1 \) since \( t_1 = 0 \) and w.p.1

\[
\tilde{X}(0) = \lim_{n \to \infty} \tilde{X}^n(0) = x = \psi^*(0);
\]

\( \psi^*(0) = x \) since \( I_x(\psi^*) < \infty \). Assuming that w.p.1 \( \tilde{X}(s) = \psi^*(s) \) for some \( k \in \{1, 2, \ldots, r\} \) and all \( s \in [0, t_k \wedge \tau] \), we will prove that w.p.1 \( \tilde{X}(s) = \psi^*(s) \) for all \( s \in [0, t_{k+1} \wedge \tau] \). Since \( t_{r+1} = 1 \), this will yield w.p.1

\[
\tilde{X}(s) = \psi^*(s) \text{ for all } s \in [0, \tau].
\]
Assuming that this has been proved, we claim that w.p.1 $\tau = 1$. Indeed, the definition of the stopping times $\tau^n$, the probability–1 convergence of $\tilde{X}^n$ to $\tilde{X}$ uniformly on $[0, 1]$, and the continuity of $\tilde{X}$ and $\psi^*$ imply that on the event \{$\tau < 1$\}

$$\lim_{t \to \tau} \|X(t) - \psi^*(t)\| \geq \delta.$$ 

However, if $\tilde{P}_x \{\tau < 1\}$ were positive, then the previous two displays would contradict the continuity of $\psi^*(t)$ and the probability–1 continuity of $\tilde{X}(t)$ at $t = \tau$. Hence from the inductive argument we will conclude that w.p.1 $\tau = 1$ and

$$\tilde{X}(s) = \psi^*(s) \text{ for all } s \in [0, 1].$$

In carrying out the inductive proof, we assume that $t_k < \tau$ because otherwise we are done. First consider the case where $(\psi^*(s))_1 \neq 0 \text{ for all } s \in (t_k, t_{k+1})$. By the definition of the admissible control sequence, for any $j \in \{0, 1, \ldots, n-1\}$ such that $j/n \in [t_k, t_{k+1} \land \tau)$

$$\int_{R^d} y \nu^n_j (dy) = \beta_k;$$

by definition $\beta_k$ equals the constant value $\dot{\psi}(s)$ for any $s \in (t_k, t_{k+1})$. As in the proof of Lemma 6.4.2, for any $t \in [t_k, t_{k+1} \land \tau)$ this equality gives rise to the probability–1 limit

$$\tilde{X}(t) - \tilde{X}(t_k) = \lim_{n \to \infty} \int_{R^d \times (t_k, t]} y \nu^n (dy \times ds) = \int_{t_k}^t \dot{\psi}^*(s) ds = \psi^*(t) - \psi^*(t_k),$$

which extends by continuity to $t = t_{k+1} \land \tau$. The inductive hypothesis is that w.p.1 $\tilde{X}(t) = \psi^*(t)$ for all $t \in [0, t_k]$. The last display now yields that w.p.1 $\tilde{X}(t) = \psi^*(t)$ for all $t \in [t_k, t_{k+1} \land \tau]$ and thus for all $t \in [0, t_{k+1} \land \tau]$. This is what we want to prove.

We next consider the case where $(\psi^*(s))_1 = 0 \text{ for all } s \in (t_k, t_{k+1})$. The aim is to prove that the probability–1 equality $\tilde{X}(t) = \psi^*(t)$ for all $t \in [0, t_k]$ implies the same probability–1 equality for all $t \in [0, t_{k+1} \land \tau]$. Given numbers $\sigma$ and $t$ satisfying $t_k < \sigma < t < t_{k+1} \land \tau$, the definition of the admissible control sequence gives

$$\int_{R^d \times [\sigma, t]} y \nu^n (dy \times ds) = \beta_k^{(1)} \gamma^{(1)n}((\sigma, t]) + \beta_k^{(2)} \gamma^{(2)n}((\sigma, t]).$$

According to part (c) of Lemma 7.4.3, w.p.1 $\gamma^{(i)}$ is absolutely continuous with respect to Lebesgue measure on $[0, 1]$. Hence by normalizing we can apply part (e) of the Portmanteau Theorem [Theorem A.3.4] to the weakly convergent sequences $\{\gamma^{(i)n}\}$. We conclude that w.p.1, for all $\sigma$ and $t$ satisfying $t_k < \sigma < t < t_{k+1} \land \tau$

$$\tilde{X}(t) - \tilde{X}(\sigma) = \int_{R^d \times [\sigma, t]} y \nu (dy \times ds) = \beta_k^{(1)} \gamma^{(1)}((\sigma, t]) + \beta_k^{(2)} \gamma^{(2)}((\sigma, t]).$$

Since w.p.1 we have a.s. for $t \in [0, 1]$

$$\tilde{X}(t) = x + \int_0^t \dot{X}(s) ds \text{ and } \gamma^{(i)}(dt) = \dot{\gamma}^{(i)}(t) dt,$$
the previous display yields

\[ \dot{X}(s) = \dot{\gamma}^{(1)}(s) \beta_k^{(1)} + \dot{\gamma}^{(2)}(s) \beta_k^{(2)}, \]

which w.p.1 holds a.s. for \( s \in (t_k, t_{k+1} \wedge \tau) \). Combining this with the properties of \( \dot{\gamma}^{(i)} \) expressed in (7.26) and the inequalities \( (\beta_k^{(1)})_1 > 0, (\beta_k^{(2)})_1 < 0 \), w.p.1 we have a.s. for \( s \in (t_k, t_{k+1} \wedge \tau) \) the implications

if \( (\dot{X}(s))_1 < 0 \), then \( (\dot{X}(s))_1 = (\beta_k^{(1)})_1 > 0 \),

and

if \( (\dot{X}(s))_1 > 0 \), then \( (\dot{X}(s))_1 = (\beta_k^{(2)})_1 < 0 \).

An elementary stability argument that uses the Lyapunov function \( f(x) \equiv |x_1| \) shows that w.p.1 \( (\dot{X}(s))_1 = 0 \) for all \( s \in (t_k, t_{k+1} \wedge \tau) \). Therefore w.p.1 we have a.s. for \( s \in (t_k, t_{k+1} \wedge \tau) \)

\[ (\dot{X}(s))_1 = \dot{\gamma}^{(1)}(s) (\beta_k^{(1)})_1 + \dot{\gamma}^{(2)}(s) (\beta_k^{(2)})_1 = 0 \]

and, according to equation (7.27),

\[ \dot{\gamma}^{(1)}(s) + \dot{\gamma}^{(2)}(s) = 1. \]

Since \( (\beta_k^{(1)})_1 > 0 \) and \( (\beta_k^{(2)})_1 < 0 \), the quantities \( \rho_k^{(1)} \) and \( \rho_k^{(2)} \) are uniquely determined by

\[ \rho_k^{(1)} + \rho_k^{(2)} = 1, \quad \rho_k^{(1)} (\beta_k^{(1)})_1 + \rho_k^{(2)} (\beta_k^{(2)})_1 = 0. \]

Since these are the same two equations as in the previous two displays, the uniqueness implies that w.p.1

\[ \dot{\gamma}^{(1)}(s) = \rho_k^{(1)} \quad \text{and} \quad \dot{\gamma}^{(2)}(s) = \rho_k^{(2)} \] (7.48)

a.s. for \( s \in (t_k, t_{k+1} \wedge \tau) \). Hence w.p.1 it follows from the second equality in (7.44) and the definition of \( \beta_k \) that a.s. for \( s \in (t_k, t_{k+1} \wedge \tau) \)

\[ \dot{X}(s) = \dot{\gamma}^{(1)}(s) \beta_k^{(1)} + \dot{\gamma}^{(2)}(s) \beta_k^{(2)} = \rho_k^{(1)} \beta_k^{(1)} + \rho_k^{(2)} \beta_k^{(2)} = \beta_k = \psi^*(s). \]

With probability 1, the inductive hypothesis states that \( \dot{X}(t_k) = \psi^*(t_k) \). The last display now yields that w.p.1 \( \dot{X}(s) = \psi^*(s) \) for all \( s \in [t_k, t_{k+1} \wedge \tau] \) and thus for all \( s \in [0, t_{k+1} \wedge \tau] \). This is what we want to prove. The proof by induction is complete, and the probability–1 equality \( \dot{X}(s) = \psi^*(s) \) for all \( s \in [0, 1] \) is shown.

For \( k \in \{1, 2, \ldots, r\} \) and \( t \in [0, 1] \), we define

\[ g_k(t) = \begin{cases} L_1^{(1)}(\psi^*(t), \beta_k), & \text{if } (\psi^*(t))_1 < 0 \\ L_2^{(1)}(\psi^*(t), \beta_k), & \text{if } (\psi^*(t))_1 > 0 \\ \rho_k^{(1)} L_1^{(1)}(\psi^*(t), \beta_k^{(1)}) + \rho_k^{(2)} L_2^{(2)}(\psi^*(t), \beta_k^{(2)}) & \text{if } (\psi^*(t))_1 = 0. \end{cases} \]

Because of formula (7.45)

\[ g_k(t) \leq \begin{cases} L_1^{(1)}(\psi^*(t), \beta_k), & \text{if } (\psi^*(t))_1 < 0 \\ L_2^{(2)}(\psi^*(t), \beta_k), & \text{if } (\psi^*(t))_1 > 0 \\ L_0^{(0)}(\psi^*(t), \beta_k) + \varepsilon & \text{if } (\psi^*(t))_1 = 0, \end{cases} \]
and thus
\[ g_k(t) \leq \tilde{L}(\psi^*(t), \dot{\psi}^*(t)) + \varepsilon. \]  
(7.49)

The limit superior of \( W^n(x) \) will now be evaluated along the subsequence of \( n \in \mathbb{N} \) for which
\[
\left( \nu^n, \nu^{(1),n}, \nu^{(2),n}, \gamma^{(1),n}, \gamma^{(2),n}, \tilde{X}^n, \tilde{X}^\gamma \right) \rightarrow \left( \nu, \nu^{(1)}, \nu^{(2)}, \gamma^{(1)}, \gamma^{(2)}, \tilde{X}, \tilde{X}^\gamma \right).
\]

Each line of the following display will be explained afterwards:

\[
\begin{align*}
\limsup_{n \to \infty} W^n(x) &= \limsup_{n \to \infty} V^n(x) \\
&\leq \limsup_{n \to \infty} \tilde{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(x)||\mu(.)|\tilde{X}^n_j) + h(\tilde{X}^n) \right\} \\
&\leq \limsup_{n \to \infty} \tilde{E}_x \left\{ \sum_{k=1}^{r} \sum_{\{l_j: i_{j-1}(t_{k}) < t_j / n \}} \frac{1}{n} \left[ L^{(1)}(\tilde{X}^n_j, \beta^{(1)}_k) \cdot 1_{\{l_j: i_{j-1}(t_{k}) < t_j / n \}} + L^{(2)}(\tilde{X}^n_j, \beta^{(2)}_k) \cdot 1_{\{l_j: i_{j-1}(t_{k}) < t_j / n \}} \right] \\
&\quad \quad + L^{(2)}(\tilde{X}^n_j, \beta^{(2)}_k) \cdot 1_{\{l_j: i_{j-1}(t_{k}) < t_j / n \}} + h(\tilde{X}^n) \right\} \\
&= \limsup_{n \to \infty} \tilde{E}_x \left\{ \sum_{k=1}^{r} \int_{t_{k}}^{t_{k+1}} L^{(1)}(\tilde{X}^n(t), \beta^{(1)}_k) \gamma^{(1),n}(dt) \\
&\quad \quad + \sum_{k=1}^{r} \int_{t_{k}}^{t_{k+1}} L^{(2)}(\tilde{X}^n(t), \beta^{(2)}_k) \gamma^{(2),n}(dt) + h(\tilde{X}^n) \right\} \\
&= \tilde{E}_x \left\{ \sum_{k=1}^{r} \int_{t_{k}}^{t_{k+1}} \left[ L^{(1)}(\tilde{X}(t), \beta^{(1)}_k) \gamma^{(1)}(t) \\
&\quad \quad + L^{(2)}(\tilde{X}(t), \beta^{(2)}_k) \gamma^{(2)}(t) \right] dt + h(\tilde{X}) \right\} \\
&= \sum_{k=1}^{r} \int_{t_{k}}^{t_{k+1}} g_k(t) dt + \tilde{E}_x \{ h(\tilde{X}) \} \\
&\leq \int_0^1 \tilde{L}(\psi^*(t), \dot{\psi}^*(t)) ds + h(\psi^*) + \varepsilon \\
&= \tilde{I}_x(\psi^*) + h(\psi^*) + \varepsilon
\end{align*}
\]

Lines two and three of the display are consequences of the representation formula (7.43). Lines four and five follow from the definition of the admissible control sequence \( \nu^n_j \), from the properties of \( \nu^n_j \) expressed in (7.46) and (7.47), and the nonnegativity of \( L^{(i)} \). Lines six and seven follow from the definitions of the piecewise linear process \( \tilde{X} \) and the measures \( \gamma^{(i),n} \). With probability 1, for \( i = 1, 2 \) \( \gamma^{(i),n} \Rightarrow \gamma^{(i)} \), which is absolutely continuous with respect to Lebesgue measure and has the decomposition \( \gamma^{(i)}(dt) = \gamma^{(i)}(t) dt \).
7.5. PROOF OF THE LAPLACE PRINCIPLE LOWER BOUND

Lines eight and nine of the display are thus consequences of the probability-1 uniform convergence of $X^n$ to $\bar{X}$ on $[0,1]$, the continuity on $\mathbb{R}^d$ of $L^{(i)}(\cdot, \beta)$ for each $\beta \in \Sigma$ [Lemma 7.5.2 (b)], part (b) of Theorem A.3.10, the continuity of $h$ on $C([0,1]: \mathbb{R}^d)$, and the Lebesgue Dominated Convergence Theorem. With probability 1 $\bar{X}(t) = \psi^*(t)$ for all $t \in [0,1]$. The last three lines of the display now follow from the probability-1 equality $\tau = 1$, the properties of $\hat{\gamma}(i)$ given in (7.26) and (7.48), the inequality (7.49) relating $g_k$ and $\hat{L}$, and the definition of $I_x(\psi^*)$.

Sending $\varepsilon \to 0$ yields for the given subsequence the upper limit

$$\limsup_{n \to \infty} W^n(x) \leq I_x(\psi^*) + h(\psi^*).$$

An argument by contradiction applied to an arbitrary subsequence of the original sequence $\{W^n(x), n \in \mathbb{N}\}$ yields the same upper limit for the entire sequence. Except for the proof of Lemma 7.5.4, we have completed the proof of the Laplace principle lower bound stated in Proposition 7.5.1. In order to carry this out, we need a number of properties of $L^{(i)}(x, \beta)$ and $\hat{L}(x, \beta)$. The complication in the statement of part (c) arises since $\hat{L}(x, \beta)$ need not be lower semicontinuous on all of $\mathbb{R}^d \times \mathbb{R}^d$. Interestingly, this does not prevent the rate function $I_x$ in the Laplace principle from being lower semicontinuous.

**Lemma 7.5.5.** We assume Conditions 7.2.1 and 7.2.2. The function $L^{(0)}(x, \beta)$ defined in equation (7.5) and the function $\hat{L}(x, \beta)$ defined in equation (7.9) have the following properties.

(a) For each $x \in \mathbb{R}^d$, $L^{(0)}(x, \beta)$ is a convex function of $\beta \in \mathbb{R}^d$. In addition, $L^{(0)}(x, \beta)$ is a nonnegative, lower semicontinuous function of $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$.

(b) $L^{(0)}(x, \beta)$ is uniformly superlinear in the sense that

$$\lim_{N \to \infty} \inf_{x \in \mathbb{R}^d} \inf_{\{\beta \in \mathbb{R}^d : \|\beta\| = N\}} \frac{1}{\|\beta\|} L^{(0)}(x, \beta) = \infty.$$

(c) For each $x \in \mathbb{R}^d$, $\hat{L}(x, \beta)$ is a convex function of $\beta \in \mathbb{R}^d$. In addition, $\hat{L}(x, \beta)$ is a nonnegative function satisfying the following restricted lower semicontinuity property:

$$\liminf_{n \to \infty} \hat{L}(x_n, \beta_n) \geq \hat{L}(x, \beta)$$

whenever $(x_n, \beta_n) \to (x, \beta)$ satisfying $x_1 \neq 0, \beta \in \mathbb{R}^d$ or $x_1 = 0, \beta_1 = 0$.

(d) $\hat{L}(x, \beta)$ is uniformly superlinear in the sense that

$$\lim_{N \to \infty} \inf_{x \in \mathbb{R}^d} \inf_{\{\beta \in \mathbb{R}^d : \|\beta\| = N\}} \frac{1}{\|\beta\|} \hat{L}(x, \beta) = \infty.$$

**Proof.** Parts (a) and (b) can both be proved by rewriting $L^{(0)}(x, \beta)$ as a Legendre-Fenchel transform. However, we prefer the proofs given below since they are more elementary.

(a) The nonnegativity of $L^{(0)}(x, \beta)$ follows from its definition and the nonnegativity of $L^{(1)}(x, \beta)$ and $L^{(2)}(x, \beta)$. In addition, part (b) of Lemma 7.5.3 implies that $L^{(0)}(x, \beta) < \infty$ for $(x, \beta) \in \mathbb{R}^d \times \Sigma$. We now show that for each $x, \beta$, and $\gamma$ in $\mathbb{R}^d$ and each $s \in (0,1)$

$$L^{(0)}(x, s\beta + (1-s)\gamma) \leq sL^{(0)}(x, \beta) + (1-s)L^{(0)}(x, \gamma).$$
Without loss of generality, we can assume that $L^{(0)}(x, \beta) < \infty$ and $L^{(0)}(x, \gamma) < \infty$ since otherwise the convexity inequality is automatic. Let $\varepsilon > 0$ be given. There exist $\rho^{(1)}, \rho^{(2)}, \beta^{(1)}, \beta^{(2)}$ that satisfy

$$
\rho^{(1)} \geq 0, \rho^{(2)} \geq 0, \rho^{(1)} + \rho^{(2)} = 1, (\beta^{(1)})_1 \geq 0, (\beta^{(2)})_1 \leq 0, \rho^{(1)} \beta^{(1)} + \rho^{(2)} \beta^{(2)} = \beta,
$$

$$
L^{(0)}(x, \beta) + \varepsilon \geq \rho^{(1)} L^{(1)}(x, \beta^{(1)}) + \rho^{(2)} L^{(2)}(x, \beta^{(2)}),
$$

and there exist $\sigma^{(1)}, \sigma^{(2)}, \gamma^{(1)}, \gamma^{(2)}$ that satisfy

$$
\sigma^{(1)} \geq 0, \sigma^{(2)} \geq 0, \sigma^{(1)} + \sigma^{(2)} = 1, (\gamma^{(1)})_1 \geq 0, (\gamma^{(2)})_1 \leq 0, \sigma^{(1)} \gamma^{(1)} + \sigma^{(2)} \gamma^{(2)} = \gamma,
$$

$$
L^{(0)}(x, \gamma) + \varepsilon \geq \sigma^{(1)} L^{(1)}(x, \gamma^{(1)}) + \sigma^{(2)} L^{(2)}(x, \gamma^{(2)}).
$$

We complete the proof under the assumption that $\rho^{(1)} > 0, \rho^{(2)} > 0, \sigma^{(1)} > 0, \sigma^{(2)} > 0$, omitting the minor modifications that are necessary when this assumption is not valid. The convexity of $L^{(1)}(x, \cdot)$ and $L^{(2)}(x, \cdot)$ [Lemma 6.2.3 (b)] implies that

$$
s L^{(0)}(x, \beta) + (1 - s) L^{(0)}(x, \gamma) + \varepsilon \\
\geq \rho^{(1)} L^{(1)}(x, \beta^{(1)}) + (1 - s) \sigma^{(1)} L^{(1)}(x, \gamma^{(1)}) + \rho^{(2)} L^{(2)}(x, \beta^{(2)}) + (1 - s) \sigma^{(2)} L^{(2)}(x, \gamma^{(2)}) \\
\geq \left( s \rho^{(1)} + (1 - s) \sigma^{(1)} \right) L^{(1)} \left( x, \frac{s \rho^{(1)}}{s \rho^{(1)} + (1 - s) \sigma^{(1)}} \beta^{(1)} + \frac{(1 - s) \sigma^{(1)}}{s \rho^{(1)} + (1 - s) \sigma^{(1)}} \gamma^{(1)} \right) \\
+ \left( s \rho^{(2)} + (1 - s) \sigma^{(2)} \right) L^{(2)} \left( x, \frac{s \rho^{(2)}}{s \rho^{(2)} + (1 - s) \sigma^{(2)}} \beta^{(2)} + \frac{(1 - s) \sigma^{(2)}}{s \rho^{(2)} + (1 - s) \sigma^{(2)}} \gamma^{(2)} \right) \\
\geq L^{(0)}(x, s \beta + (1 - s) \gamma).
$$

Sending $\varepsilon \to 0$ gives what we want to prove.

We now prove that $L^{(0)}(\cdot, \cdot)$ is lower semicontinuous on $\mathbb{R}^d \times \mathbb{R}^d$. We do this by showing that for any sequence $\{(x_n, \beta_n), n \in \mathbb{N} \}$ converging to $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$

$$
\liminf_{n \to \infty} L^{(0)}(x_n, \beta_n) \geq L^{(0)}(x, \beta).
$$

Without loss of generality we can assume that $A = \liminf_{n \to \infty} L^{(0)}(x_n, \beta_n)$ is finite since otherwise there is nothing to prove. Let $\{(x_n, \beta_n)\}$ be a subsequence for which $L^{(0)}(x_n, \beta_n) < \infty$ and $L^{(0)}(x_n, \beta_n) \to A$. For each $n$ there exist $\rho^{(1)}_n, \rho^{(2)}_n, \beta^{(1)}_n, \beta^{(2)}_n$ that satisfy

$$
\rho^{(1)}_n \geq 0, \rho^{(2)}_n \geq 0, \rho^{(1)}_n + \rho^{(2)}_n = 1, (\beta^{(1)}_n)_1 \geq 0, (\beta^{(2)}_n)_1 \leq 0, \rho^{(1)}_n \beta^{(1)}_n + \rho^{(2)}_n \beta^{(2)}_n = \beta_n
$$

and

$$
L^{(0)}(x_n, \beta_n) + \frac{1}{n} \geq \rho^{(1)}_n L^{(1)}(x_n, \beta^{(1)}_n) + \rho^{(2)}_n L^{(2)}(x_n, \beta^{(2)}_n).
$$

Since $[0, 1]$ is compact, there exists a subsubsequence of $n \in \mathbb{N}$ such that $\rho^{(1)}_n \to \rho^{(1)}$ and $\rho^{(2)}_n \to \rho^{(2)}$, where $\rho^{(1)} \geq 0, \rho^{(2)} \geq 0$, and $\rho^{(1)} + \rho^{(2)} = 1$. 

7.5. PROOF OF THE LAPLACE PRINCIPLE LOWER BOUND

Either $\rho^{(1)}$ or $\rho^{(2)}$ or both are positive. We consider $i \in \{1, 2\}$ for which $\rho^{(i)} > 0$. Since $L^{(i)}(\cdot, \cdot)$ is uniformly superlinear [Lemma 6.2.3 (c)] and $L^{(1)}(\cdot, \cdot)$ and $L^{(2)}(\cdot, \cdot)$ are nonnegative, the bound

$$\sup_{n \in \mathbb{N}} \rho_n^{(i)}(x_n, \beta_n) \leq \sup_{n \in \mathbb{N}} \left( L^{(0)}(x_n, \beta_n) + \frac{1}{n} \right) < \infty$$

implies that the subsequence $\{\beta_n^{(i)}\}$ lies in a compact set. Hence there exists $\beta^{(i)} \in \mathbb{R}^d$ such that $(\beta^{(i)})_1 \geq 0$ if $i = 1$, $(\beta^{(i)})_1 \leq 0$ if $i = 2$, and along a further subsequence of $n \in \mathbb{N}$ $\beta_n^{(i)} \to \beta^{(i)}$. Along this same subsequence $x_n \to x$ and $L^{(0)}(x_n, \beta_n) \to A$.

We complete the proof of lower semicontinuity by considering two cases. If both $\rho^{(1)} > 0$ and $\rho^{(2)} > 0$, then $\rho^{(1)}(x, \beta^{(1)}) + \rho^{(2)}(x, \beta^{(2)}) = \beta$, and since $L^{(1)}(\cdot, \cdot)$ and $L^{(2)}(\cdot, \cdot)$ are lower semicontinuous on $\mathbb{R}^d \times \mathbb{R}^d$ [Lemma 6.2.3 (b)],

$$A = \lim_{n \to \infty} L^{(0)}(x_n, \beta_n)$$

$$\geq \liminf_{n \to \infty} \left( \rho_n^{(1)} L^{(1)}(x_n, \beta_n^{(1)}) + \rho_n^{(2)} L^{(2)}(x_n, \beta_n^{(2)}) - \frac{1}{n} \right)$$

$$\geq \rho^{(1)} L^{(1)}(x, \beta^{(1)}) + \rho^{(2)} L^{(2)}(x, \beta^{(2)}) \geq L^{(0)}(x, \beta).$$

Now suppose that $\rho^{(1)} = 1$ and $\rho^{(2)} = 0$; the case $\rho^{(1)} = 0$ and $\rho^{(2)} = 1$ is handled similarly. Then $\beta^{(1)} = \beta$ and since $L^{(1)}(\cdot, \cdot)$ is lower semicontinuous on $\mathbb{R}^d \times \mathbb{R}^d$ and $L^{(2)}(\cdot, \cdot)$ is nonnegative,

$$A = \lim_{n \to \infty} L^{(0)}(x_n, \beta_n) \geq \liminf_{n \to \infty} \left( \rho_n^{(1)} L^{(1)}(x_n, \beta_n^{(1)}) - \frac{1}{n} \right)$$

$$\geq L^{(1)}(x, \beta^{(1)}) \geq L^{(0)}(x, \beta).$$

This completes the proof that $L^{(0)}(\cdot, \cdot)$ is lower semicontinuous.

(b) For $x, \alpha$, and $\beta$ in $\mathbb{R}^d$ we define

$$\bar{H}(x, \alpha) = H^{(1)}(x, \alpha) \lor H^{(2)}(x, \alpha) \text{ and } \bar{L}(x, \beta) = \sup_{\alpha \in \mathbb{R}^d} \{ \langle \alpha, \beta \rangle - \bar{H}(x, \alpha) \}.$$

For each $x$ $\bar{H}(x, \beta)$ is a finite convex function of $\beta \in \mathbb{R}^d$. According to Corollary D.4.3, if in the definition of $L^{(0)}(x, \beta)$ given in (7.5) we remove the constraints $(\beta^{(1)})_1 \geq 0$, $(\beta^{(2)})_1 \leq 0$, then we obtain $\bar{L}(x, \beta)$. In other words,

$$L^{(0)}(x, \beta) \geq \inf \left\{ \rho^{(1)} L^{(1)}(x, \beta^{(1)}) + \rho^{(2)} L^{(2)}(x, \beta^{(2)}) : \rho^{(1)} \geq 0, \rho^{(2)} \geq 0, \rho^{(1)} + \rho^{(2)} = 1, \beta^{(1)} \in \mathbb{R}^d, \beta^{(2)} \in \mathbb{R}^d, \rho^{(1)} \beta^{(1)} + \rho^{(2)} \beta^{(2)} = \beta \right\}$$

$$= \bar{L}(x, \beta).$$

Exactly the same proof used to establish the uniform superlinearity of $L(x, \beta)$ in part (c) of Lemma 6.2.3 shows that $\bar{L}(x, \beta)$ is uniformly superlinear. The inequality $L^{(0)}(x, \beta) \geq \bar{L}(x, \beta)$ implies the uniform superlinearity of $L^{(0)}(x, \beta)$.

(c) The convexity and nonnegativity properties of $\bar{L}(x, \beta)$ follow from the convexity and nonnegativity properties of $L^{(i)}(x, \beta), i = 0, 1, 2$. The lower semicontinuity property
of \( \tilde{L}(x, \beta) \) when \( x_1 < 0 \) or when \( x_1 > 0 \) is a consequence of the lower semicontinuity of \( L^{(1)}(x, \beta) \) or of \( L^{(2)}(x, \beta) \), respectively. In order to prove the lower semicontinuity property of \( \tilde{L}(x, \beta) \) when \( x_1 = 0, \beta_1 = 0 \), it suffices to consider three types of sequences \((x_n, \beta_n) \to (x, \beta)\): all \( (x_n)_1 < 0 \), all \( (x_n)_1 > 0 \), and all \( (x_n)_1 = 0 \). In the first two cases, the lower semicontinuity property is a consequence of the lower semicontinuity of \( L^{(1)}(x, \beta) \) and of \( L^{(2)}(x, \beta) \), respectively, and part (a) of Lemma 7.5.3. In the third case, we use the lower semicontinuity of \( L^{(3)}(x, \beta) \).

(d) The uniform superlinearity of \( \tilde{L}(x, \beta) \) follows from the uniform superlinearity of \( L^{(i)}(x, \beta), i = 0, 1, 2 \). For \( i = 0 \) this property is a consequence of part (b) of the present lemma, while for \( i = 1, 2 \) this property is a consequence of part (c) of Lemma 6.2.3.

We now prove Lemma 7.5.4, which is related to the Approximation Procedure 6.4.3.

**Proof of Lemma 7.5.4.** The first step is to adapt Lemma 6.5.3 to the present situation, in which the rate function

\[
I_x(\varphi) \triangleq \int_0^1 \tilde{L}(\varphi(t), \varphi(t)) \, dt
\]

replaces the rate function \( \int_0^1 L(\varphi(t), \varphi(t)) \, dt \) for absolutely continuous \( \varphi \in C([0, 1] : \mathbb{R}^d) \) satisfying \( \varphi(0) = x \). The four properties of \( \tilde{L}(\cdot, \cdot) \) needed for this adaptation are listed in Remark 6.5.4. Specifically, parts (c) and (d) of Lemma 7.5.5 shows that \( \tilde{L}(x, \beta) \) is uniformly superlinear, nonnegative, and convex with respect to \( \beta \in \mathbb{R}^d \) for each \( x \in \mathbb{R}^d \).

We now verify that if \( \psi \) is any function in \( C([0, 1] : \mathbb{R}^d) \) satisfying \( I_x(\psi) < \infty \), then for some \( c \in (0, \infty) \)

\[
\sup_{t \in [0, 1]} \sup_{\beta \in K(0, c)} \tilde{L}(\psi(t), \beta) < \infty, \tag{7.51}
\]

where \( K(0, c) \) is a closed ball of radius \( c \) centered at 0 and contained in \( \Sigma \). We have agreed that without loss of generality the set \( \Sigma \) in Condition 7.2.2 can be taken to be a nonempty open subset of \( \mathbb{R}^d \). Since by part (b) of this condition 0 lies in \( \Sigma \), \( c \in (0, 1) \) can be chosen so that

\[
K(0, c) = \{ \beta \in \mathbb{R}^d : \|\beta\| \leq c \} \subset \Sigma.
\]

Since for \( i = 1, 2 \) \( L^{(i)}(\cdot, \cdot) \) is continuous on \( \mathbb{R}^d \times \Sigma \) [Lemma 7.5.2], it follows that

\[
\sup_{t \in [0, 1]} \sup_{\beta \in K(0, c)} L^{(i)}(\psi(t), \beta) < \infty.
\]

This implies (7.51) since by part (a) of Lemma 7.5.3

\[
\tilde{L}(x, \beta) \leq \max_{i=1, 2} L^{(i)}(x, \beta)
\]

for each \( x \) and \( \beta \) in \( \mathbb{R}^d \). Having verified these four properties of \( \tilde{L}(\cdot, \cdot) \), we can now conclude that for each \( \delta > 0 \), there exist \( \theta \in (0, \infty) \) and \( \zeta \in C([0, 1] : \mathbb{R}^d) \) with the following properties:

\[
\|\zeta - \psi\|_\infty \leq \delta/2 \text{ and } I_x(\zeta) \leq I_x(\psi) + \delta/2
\]
and a.s. for \( t \in [0, 1] \), \( \hat{\zeta}(t) \in \Sigma^{-\theta} \) and \( \|\hat{\zeta}(t)\| \leq 1/\theta \).

We consider the closed ball \( K(0, 1/\theta) = \{ \beta \in \mathbb{R}^d : \|\beta\| \leq 1/\theta \} \) and define
\[
\Phi = \{ t \in [0, 1] : \hat{\zeta}(t) \in \Sigma^{-\theta} \cap K(0, 1/\theta) \},
\]
which has Lebesgue measure 1. Let \( Q \) be any compact subset of \( \mathbb{R}^d \). \( L^{(1)}(x, \beta) \) and \( L^{(2)}(x, \beta) \) are uniformly continuous on \( Q \times (\Sigma^{-\theta} \cap K(0, 1/\theta)) \), and \( L^{(0)}(x, \beta) \) is uniformly continuous on \( Q \times (\Sigma^{-\theta} \cap K(0, 1/\theta) \cap \partial) \). Hence for \( i = 1, 2 \) (resp., \( i = 0 \)) there exists \( \lambda > 0 \) such that
\[
|L^{(i)}(x, \beta) - L^{(i)}(\xi, \beta)| \leq \delta/2
\]
whenever \( x \in Q, \xi \in Q, \|x - \xi\| \leq \lambda \), and \( \beta \in \Sigma^{-\theta} \cap K(0, 1/\theta) \) (resp., \( \beta \in \Sigma^{-\theta} \cap K(0, 1/\theta) \cap \partial) \). In addition, there exists \( \sigma_1 \in (0, 1] \) such that whenever \( s \) and \( t \) in \( [0, 1] \) satisfy \( |s - t| \leq \sigma_1 \), then
\[
\|\hat{\zeta}(s) - \hat{\zeta}(t)\| \leq \lambda/2.
\]
Now take \( Q \) to be any compact, convex subset of \( \mathbb{R}^d \) containing \( \{ \zeta(t), t \in [0, 1] \} \). Then for \( i = 1, 2 \) (resp., \( i = 0 \)), whenever \( s \in \Phi \) and \( t \in [0, 1] \) satisfy \( |s - t| \leq \sigma_1 \) (resp., satisfy \( |s - t| \leq \sigma_1 \) and \( \hat{\zeta}(s) = 0 \), we have
\[
|L^{(i)}(\zeta(s), \zeta(t)) - L^{(i)}(\zeta(t), \zeta(s))| \leq \delta/2.
\]

For \( \sigma \in (0, \sigma_1] \) we define
\[
G^{(0)} = \{ t \in [0, 1] : \zeta(t)_1 = 0 \},
G^{(1)} = \{ t \in [0, 1] : \zeta(t)_1 < 0 \}, \quad G^{(2)} = \{ t \in [0, 1] : \zeta(t)_1 > 0 \}.
\]
Let \( B^{(0)} \subset [0, 1] \) be a finite union of intervals \( [c_j, d_j], j \in \{1, 2, \ldots, J\} \), such that \( d_j \leq c_{j+1} \) for each \( j \in \{1, 2, \ldots, J - 1\} \) and such that \( G^{(0)} \subset B^{(0)} \). Without loss of generality, we can assume that for each \( j \in \{1, 2, \ldots, J\} \)
\[
(\zeta(c_j))_1 = (\zeta(d_j))_1 = 0
\]
and that \( \max_{j\in\{1,2,\ldots,J\}} (d_j - c_j) < \sigma \). The proof will be completed assuming that \( c_1 = 0 \) and \( d_1 = 1 \) and that the sets \( G^{(i)}, i = 0, 1, 2 \), each have positive Lebesgue measure. We will omit the minor changes that are required when these assumptions do not hold.

For each \( j \in \{1, 2, \ldots, J\} \) such that \( c_j < d_j \), we define
\[
u_j^0 = \frac{1}{d_j - c_j} \int_{c_j}^{d_j} \zeta(s) \, ds = \frac{\zeta(d_j) - \zeta(c_j)}{d_j - c_j}.
\]
For all \( j \in \{1, 2, \ldots, J - 1\} \) such that \( c_{j+1} > d_j \), we choose finitely many numbers \( \{e^j_k, k = 1, 2, \ldots, K_j\} \) such that
\[
d_j = e^j_1 < e^j_2 < \cdots < e^j_{K_j} = c_{j+1} \quad \text{and} \quad \max_{k \in \{1, 2, \ldots, K_j-1\}} (e^j_{k+1} - e^j_k) < \sigma.\]
Note that \((\zeta(t))_1 \neq 0\) for \(t \in (d_j, c_{j+1})\). For each \(j \in \{1, 2, \ldots, J\}\) and \(k \in \{1, 2, \ldots, K_j - 1\}\), we define

\[
\begin{align*}
u_j^k & = \frac{1}{e_j^{k+1} - e_j^k} \int_{e_j^k}^{e_j^{k+1}} \zeta(s) \, ds = \frac{\zeta(e_j^{k+1}) - \zeta(e_j^k)}{e_j^{k+1} - e_j^k}.
\end{align*}
\]

Finally, for \(t \in [0, 1]\) we set

\[
\begin{align*}
u^\sigma(t) & = \sum_{j=1}^J \left( 1_{[c_j, d_j]}(t) \nu_j^0 + \sum_{k=1}^{K_j-1} 1_{[e_j^k, e_j^{k+1}]}(t) \nu_j^k \right) \\
\psi^\sigma(t) & = x + \int_0^t \nu^\sigma(s) \, ds.
\end{align*}
\]

The function \(\psi^\sigma(t)\) is the piecewise linear interpolation of \(\zeta(t)\) with interpolation points \(c_j, d_j,\) and \(e_j^k\). We have \(\psi^\sigma(t) = \psi^\sigma(t)\) for almost all \(t \in [0, 1]\) and \(\psi^\sigma(t) = \zeta(t)\) whenever \(t = c_j, t = d_j,\) or \(t = e_j^k\). In addition, for all \(t \in [0, 1]\) \(\nu^\sigma(t)\) lies in the convex set \(\Sigma^\sigma \cap K(0, 1/\theta)\). Since \(\|\psi^\sigma - \zeta\|_{\infty} \to 0\) as \(\sigma \to 0\), there exists \(\sigma_2 \in (0, \sigma_1]\) such that

\[
\|\psi^\sigma - \zeta\|_{\infty} \leq (\lambda/4) \land (\delta/2) \text{ whenever } \sigma \in (0, \sigma_2]. \tag{7.55}
\]

We claim that for each \(j \in \{1, 2, \ldots, J\}\) and \(k \in \{1, 2, \ldots, K_j - 1\}\), \((\psi^\sigma(t))_1 = 0\) for all \(t \in (c_j, d_j)\) and \((\psi^\sigma(t))_1 \neq 0\) for all \(t \in (e_j^k, e_j^{k+1})\). This follows from the facts that \((\zeta(c_j))_1 = 0 = (\zeta(d_j))_1\), that \((\zeta(t))_1 \neq 0\) for all \(t \in (d_j, c_{j+1})\), and that \(\psi^\sigma\) is the piecewise linear interpolation of \(\zeta\) with interpolation points \(c_j, d_j,\) and \(e_j^k\). We next prove that for any \(\sigma \in (0, \sigma_2]\)

\[
I_x(\psi^\sigma) \leq I_x(\zeta) + \delta/2.
\]

The function \(\psi^\sigma\) in the statement of Lemma 7.5.4 can then be taken to be \(\psi^\sigma\) for any such value of \(\sigma\).

Let us consider a nonempty interval \((c_j, d_j), j \in \{1, 2, \ldots, J\}\), and for \(i = 0, 1, 2\) define

\[
\begin{align*}
\alpha^{(i)} & = \int_{c_j}^{d_j} 1_{G^{(i)}}(s) \, ds \quad \text{and} \quad \beta^{(i)} = \frac{1}{\alpha^{(i)}} \int_{c_j}^{d_j} \zeta(s) \, 1_{G^{(i)}}(s) \, ds.
\end{align*}
\]

Clearly \(\sum_{i=0}^2 \alpha^{(i)} = d_j - c_j\), and by assumption each \(\alpha^{(i)} > 0, i = 0, 1, 2\). Since

\[
\left\{ s \in [0, 1] : (\zeta(s))_1 = 0, (\zeta(s))_1 \neq 0 \right\}
\]

has Lebesgue measure 0 [Theorem A.6.3], \((\beta^{(0)})_1\) equals 0. Since the open sets \(G^{(1)}\) and \(G^{(2)}\) can each be written as a countable union of open intervals at each endpoint of which \(\zeta(s) = 0, (\beta^{(1)})_1\) and \((\beta^{(2)})_1\) also equal 0. It follows from part (a) of Lemma 7.5.3 that

\[
L^{(1)}(\zeta(c_j), \beta^{(1)}) \geq L^{(0)}(\zeta(c_j), \beta^{(1)}) \quad \text{and} \quad L^{(2)}(\zeta(c_j), \beta^{(2)}) \geq L^{(0)}(\zeta(c_j), \beta^{(2)}). \tag{7.56}
\]

In addition,

\[
\frac{1}{d_j - c_j} \sum_{i=0}^2 \alpha^{(i)} \beta^{(i)} = \frac{1}{d_j - c_j} \int_{c_j}^{d_j} \zeta(s) \, ds = u_j^0. \tag{7.57}
\]
7.5. PROOF OF THE LAPLACE PRINCIPLE LOWER BOUND

For any \( \sigma \in (0, \sigma_2] \) and \( s \in (c_j, d_j) \), we have \( 0 < s - c_j < d_j - c_j < \sigma \) and so by (7.53) and (7.55)
\[
\|\psi^{\sigma}(s) - \psi^{\sigma}(c_j)\| \leq \|\zeta(s) - \zeta(c_j)\| + 2 \|\psi^{\sigma} - \zeta\|_{\infty} \leq \lambda.
\]
In addition, for such \( s \) \((\psi^{\sigma}(s))_1 = 0 \). Since the compact convex set \( Q \) contains \( \{\zeta(t), t \in [0, 1]\} \), \( Q \) also contains \( \{\psi^{\sigma}(t), t \in [0, 1]\} \). In addition, a.s. for \( s \in (0, 1) \) (7.57) implies that \( \psi^{\sigma}(s) \in \Sigma^{-\theta} \cap K(0, 1/\theta) \). Hence (7.52) implies that for any \( \sigma \in (0, \sigma_2] \) and a.s. for \( s \in (c_j, d_j) \)
\[
\left| L^{(0)}(\psi^{\sigma}(s), \dot{\psi}^{\sigma}(s)) - L^{(0)}(\psi^{\sigma}(c_j), \dot{\psi}^{\sigma}(s)) \right| \leq \delta/2. \tag{7.58}
\]
For any \( \sigma \in (0, \sigma_2] \), we now prove
\[
\int_{c_j}^{d_j} \tilde{L}\left(\psi^{\sigma}(s), \dot{\psi}^{\sigma}(s)\right) ds \leq \int_{c_j}^{d_j} \tilde{L}\left(\zeta(s), \dot{\zeta}(s)\right) ds + \delta(d_j - c_j). \tag{7.59}
\]
A similar estimate applies for each interval of the form \((c_j^k, e_j^{k+1})\). Combining the estimates over the different intervals gives the desired inequality \( I_{\sigma}(\psi^{\sigma}) \leq I_{\sigma}(\zeta) + \delta/2 \) for any \( \sigma \in (0, \sigma_2] \). For \( \sigma \in (0, \sigma_2] \), the following yields (7.59):
\[
\int_{c_j}^{d_j} \tilde{L}\left(\zeta(s), \dot{\zeta}(s)\right) ds
\]
\[
= \int_{c_j}^{d_j} \sum_{i=0}^{2} 1_{G(i)}(s) L^{(i)}(\zeta(s), \dot{\zeta}(s)) ds
\]
\[
\geq \int_{c_j}^{d_j} \sum_{i=0}^{2} 1_{G(i)}(s) L^{(i)}(\zeta(c_j), \dot{\zeta}(s)) ds - \delta(d_j - c_j)/2
\]
\[
\geq \sum_{i=0}^{2} a(\beta(\zeta(c_j), \dot{\beta}(\zeta(c_j)), \gamma(\zeta(c_j), \dot{\gamma}(\zeta(c_j))) - \delta(d_j - c_j)/2
\]
\[
\geq \sum_{i=0}^{2} a(\beta(\zeta(c_j), \dot{\beta}(\zeta(c_j)), \gamma(\zeta(c_j), \dot{\gamma}(\zeta(c_j))) - \delta(d_j - c_j)/2
\]
\[
= \int_{c_j}^{d_j} L^{(0)}(\psi^{\sigma}(s), \dot{\psi}^{\sigma}(s)) ds - \delta(d_j - c_j)/2
\]
\[
\geq \int_{c_j}^{d_j} L^{(0)}(\psi^{\sigma}(s), \dot{\psi}^{\sigma}(s)) ds - \delta(d_j - c_j)
\]
\[
= \int_{c_j}^{d_j} \tilde{L}\left(\psi^{\sigma}(s), \dot{\psi}^{\sigma}(s)\right) ds - \delta(d_j - c_j).
\]
Line two follows from the definition of \( \tilde{L} \); line three from (7.54); line four from the convexity of \( L^{(i)}(x, \cdot) \) and Jensen’s Inequality; line five from (7.56); line six from the convexity of \( L^{(0)}(x, \cdot) \), Jensen’s Inequality, and (7.57); line seven from the facts that \( \psi^{\sigma}(c_j) = \zeta(c_j) \) and \( \psi^{\sigma}(s) = v_j^0 \) for all \( s \in (c_j, d_j) \); line eight from (7.58); line nine from the definition of \( L \) and the fact that \((\psi^{\sigma}(s))_1 = 0 \) for all \( s \in (c_j, d_j) \). The proof of Lemma 7.5.4 is complete. 

\[\blacksquare\]
We have finished the proof of the Laplace principle lower bound stated in Proposition 7.5.1. In the next section we complete the proof of the Laplace principle by showing that $I_x$ has compact level sets.

### 7.6 Compactness of the Level Sets of $I_x$

In the rest of the book the conditions required to prove that a given rate function has compact level sets are the same as those that are used to prove the Laplace principle upper bound. We deviate from this procedure in the present section in order to simplify the proof. In particular, in order to avoid certain measurability issues, it will greatly simplify the proof to appeal to Lemma 7.5.4. This lemma, which was also used in the proof of the Laplace principle lower bound, requires Condition 7.2.2.

**Proposition 7.6.1.** Under Conditions 7.2.1 and 7.2.2, the function $I_x$ defined in Theorem 7.2.3 has compact level sets in $\mathcal{C}([0, 1]: \mathbb{R}^d)$.

**Proof.** In Proposition 6.2.4, we proved that the rate function $I_x$ for the random walk model with continuous statistics has compact level sets. In proving the present proposition, we follow the pattern of that earlier proof, which of course must be modified because of the discontinuous statistics. Our task is to prove that for any $M < \infty$ the level set

$$Z(M) \doteq \{ \varphi \in \mathcal{C}([0, 1]: \mathbb{R}^d) : I_x(\varphi) \leq M \}$$

is a compact subset of $\mathcal{C}([0, 1]: \mathbb{R}^d)$. Suppose that $\{ \varphi^n, n \in \mathbb{N} \}$ is any sequence in $Z(M)$. Thus for each $n$ $I_x(\varphi^n) \leq M$. We will first prove that there exists $\varphi \in \mathcal{C}([0, 1]: \mathbb{R}^d)$ such that for some subsequence of $n \in \mathbb{N}$

$$\lim_{n \to \infty} \| \varphi^n - \varphi \|_{\infty} = 0.$$

This will show that $Z(M)$ is relatively compact. We will then prove that $I_x(\varphi) \leq M$. This will show that $Z(M)$ is also closed and thus is compact.

We begin by appealing to Lemma 7.5.4, which guarantees that for each $n \in \mathbb{N}$ there exists $f^n \in \mathcal{N}_0$ satisfying

$$\| f^n - \varphi^n \|_{\infty} \leq \frac{1}{n} \text{ and } I_x(f^n) \leq I_x(\varphi^n) + \frac{1}{n} \leq M + \frac{1}{n}.$$

In particular, since $f^n \in \mathcal{N}_0$, $f^n(t)$ is piecewise constant with only finitely many jumps in $(0, 1)$ and it lies in $\Sigma$ for all $t \in (0, 1)$.

The next step is to determine sequences $\{ \nu^{(1), n}, n \in \mathbb{N} \}$ and $\{ \nu^{(2), n}, n \in \mathbb{N} \}$ of sub-probability measures on $\mathbb{R}^d \times [0, 1]$ such that for each $n \in \mathbb{N}$ and $t \in [0, 1]$

$$f^n(t) = x + \int_{\mathbb{R}^d \times [0, t]} y \nu^{(1), n}(dy \times ds) + \int_{\mathbb{R}^d \times [0, t]} y \nu^{(2), n}(dy \times ds) \quad (7.60)$$
and

\[ I_x(f^n) + \frac{1}{n} \geq \int_{\mathbb{R}^d \times [0,1]} L^{(1)}(f^n(t), y) \nu^{(1), n}(dy \times dt) + \int_{\mathbb{R}^d \times [0,1]} L^{(2)}(f^n(t), y) \nu^{(2), n}(dy \times dt). \]  \hspace{1cm} (7.61)

We introduce the set

\[ \mathcal{D} = \left\{ t \in [0,1] : \hat{f}^n(t) \text{ exists for all } n \in \mathcal{N} \right\}, \]

which has Lebesgue measure 1, and for each \( n \in \mathcal{N} \) define

\[ \Gamma^{(0), n} = \left\{ t \in [0,1] : (f^n(t))_1 = 0 \right\}, \]

\[ \Gamma^{(1), n} = \left\{ t \in [0,1] : (f^n(t))_1 < 0 \right\}, \Gamma^{(2), n} = \left\{ t \in [0,1] : (f^n(t))_1 > 0 \right\}. \]

By the continuity of \( L^{(0)}(\cdot, \cdot) \) on \( \mathbb{R}^d \times (\Sigma \cap \partial) \) [Lemma 7.5.3 (c)], one can find for \( i = 1, 2 \) piecewise constant functions \( \rho^{(i), n} \) mapping \([0,1]\) into \( \mathcal{R} \) and \( \beta^{(i), n} \) mapping \([0,1]\) into \( \mathcal{R}^d \) and having the following properties:

(a) For all \( t \in \Gamma^{(0), n} \)

\[ \rho^{(1), n}(t) \geq 0, \rho^{(2), n}(t) \geq 0, \rho^{(1), n}(t) + \rho^{(2), n}(t) = 1, \]

and

\[ (\beta^{(1), n}(t))_1 \geq 0, (\beta^{(2), n}(t))_1 \leq 0. \]

(b) For all \( t \in \Gamma^{(0), n} \cap \mathcal{D} \)

\[ \rho^{(1), n}(t) \beta^{(1), n}(t) + \rho^{(2), n}(t) \beta^{(2), n}(t) = \hat{f}^n(t) \]

and

\[ \rho^{(1), n}(t) L^{(1)}(f^n(t), \beta^{(1), n}(t)) + \rho^{(2), n}(t) L^{(2)}(f^n(t), \beta^{(2), n}(t)) \leq L^{(0)}(f^n(t), \hat{f}^n(t)) + \frac{1}{n}. \]

(c) For all \( t \in \Gamma^{(1), n} \cup \Gamma^{(2), n} \)

\[ \rho^{(1), n}(t) = \rho^{(2), n}(t) = 0 \text{ and } \beta^{(1), n}(t) = \beta^{(2), n}(t) = 0. \]

For each \( i = 1, 2 \) and \( n \in \mathcal{N} \) and for Borel subsets \( A \) of \( \mathbb{R}^d \) and \( B \) of \([0,1]\) we define subprobability measures \( \nu^{(i), n} \) on \( \mathbb{R}^d \times [0,1] \) by

\[ \nu^{(i), n}(A \times B) = \int_B \nu^{(i), n}(A|t) dt, \]

where for \( t \in \mathcal{D} \), and thus a.s.,

\[ \nu^{(i), n}(A|t) = \delta_{\hat{f}^n(t)}(A) 1_{\Gamma^{(i), n}(t)} + \delta_{\beta^{(i), n}(t)}(A) \rho^{(i), n}(t). \]
With this definition, one easily verifies (7.60) and (7.61). We also define probability measures \( \nu^n \) on \( \mathbb{R}^d \times [0,1] \) by \( \nu^n = \nu^{(1),n} + \nu^{(2),n} \).

Since
\[
\infty > M + 2
\]
\[
\geq \sup_{n \in \mathbb{N}} \left( I_x(f^n) + \frac{1}{n} \right)
\]
\[
\geq \sup_{n \in \mathbb{N}} \left( \int_{\mathbb{R}^d \times [0,1]} L^{(1)}(f^n(t), y) \nu^{(1),n}(dy \times dt) + \int_{\mathbb{R}^d \times [0,1]} L^{(2)}(f^n(t), y) \nu^{(2),n}(dy \times dt) \right),
\]
the uniform superlinearity and nonnegativity of \( L^{(1)}(x, \beta) \) and \( L^{(2)}(x, \beta) \) [Lemma 6.2.3 (b)-(c)] imply for \( i = 1,2 \) the uniform integrability
\[
\lim_{C \to \infty} \sup_{n \in \mathbb{N}} \int_{\{y \in \mathbb{R}^d : ||y|| > C\} \times [0,1]} \|y\| \nu^{(i),n}(dy \times dt) = 0.
\]
Thus there exist a subsequence of \( n \in \mathbb{N} \), a probability measure \( \nu \), and subprobability measures \( \nu^{(1)} \) and \( \nu^{(2)} \) such that along this subsequence \( \nu^{(1),n} \Rightarrow \nu^{(1)} \), \( \nu^{(2),n} \Rightarrow \nu^{(2)} \), and \( \nu^n \Rightarrow \nu \). Just as in the proof of Proposition 6.2.4, one shows that
\[
\lim_{n \to \infty} \|f^n - \varphi\|_\infty = 0,
\]
where \( \varphi \) is the function in \( C([0,1]: \mathbb{R}^d) \) defined by
\[
\varphi(t) = x + \int_{\mathbb{R}^d \times [0,t]} y \nu^{(1)}(dy \times ds) + \int_{\mathbb{R}^d \times [0,t]} y \nu^{(2)}(dy \times ds).
\]
Since \( \|f^n - \varphi^n\|_\infty \leq 1/n \), it follows that
\[
\lim_{n \to \infty} \|\varphi^n - \varphi\|_\infty = 0,
\]
This shows that \( Z(M) \) is relatively compact.

In order to complete the proof that \( Z(M) \) is compact, we must show that \( \varphi \) lies in \( Z(M) \); i.e., that \( I_x(\varphi) \leq M \). Since the second marginal of \( \nu^n \) is Lebesgue measure \( \lambda \) on \( [0,1] \), the weak convergence \( \nu^n \Rightarrow \nu \) implies that the second marginal of \( \nu \) is \( \lambda \). Hence the equality \( \nu = \nu^{(1)} + \nu^{(2)} \) yields that for \( i = 1,2 \) the second marginal of \( \nu^{(i)} \) is absolutely continuous with respect to \( \lambda \) with Radon–Nikodym derivative in \([0,1]\) \( \lambda \)-a.s. As in the proof of part (b) of Lemma 7.4.3, it follows that there exist substochastic kernels \( \nu^{(i)}(dy|t) \) and \( \nu^{(2)}(dy|t) \) on \( \mathbb{R}^d \) given \([0,1]\) such that for \( i = 1,2 \)
\[
\nu^{(i)}(dy \times dt) = \nu^{(i)}(dy|t) \otimes dt.
\]
A simplified version of Theorem 7.4.4 yields the following properties of \( \nu^{(1)}(dy|t) \) and \( \nu^{(2)}(dy|t) \):

(i) \( \lambda \)-a.s. for \( t \in [0,1] \)

whenever \( (\varphi(t))_1 < 0 \), \( \nu^{(1)}(\mathbb{R}^d|t) = 1 \) and \( \nu^{(2)}(\mathbb{R}^d|t) = 0 \),
whenever \( (\varphi(t))_1 > 0 \), \( \nu^{(2)}(\mathbb{R}^d|t) = 1 \) and \( \nu^{(1)}(\mathbb{R}^d|t) = 0 \),
and for any value of \((\varphi(t))_1\) \(\nu^{(1)}(\mathbb{R}^d|t) + \nu^{(2)}(\mathbb{R}^d|t) = 1\);
(ii) \(\lambda\)-a.s. for \(t \in [0, 1]\), if \((\varphi(t))_1 = 0\), then
\[
\left(\int_{\mathbb{R}^d} y \nu^{(1)}(dy|t)\right)_1 \geq 0 \quad \text{and} \quad \left(\int_{\mathbb{R}^d} y \nu^{(2)}(dy|t)\right)_1 \leq 0;
\]
(iii) \(\lambda\)-a.s. for \(t \in [0, 1]\)
\[
\hat{\varphi}(t) = \int_{\mathbb{R}^d} y \nu^{(1)}(dy|t) + \int_{\mathbb{R}^d} y \nu^{(2)}(dy|t).
\]

The subsequence \(\{f^n\}\) converges uniformly on \([0, 1]\) to \(\varphi\). In addition \(L^{(1)}(x, \beta)\) and \(L^{(2)}(x, \beta)\) are nonnegative, lower semicontinuous functions of \((x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d\) [Lemma 6.2.3 (b)]. Since for \(i = 1, 2\) \(\nu^{(i),n} \Rightarrow \nu^{(i)}\), by normalizing these subprobability measures we can apply part (b) of Theorem A.3.13. It follows that

\[
M \geq \liminf_{n \to \infty} I_x(f^n) \\
\geq \liminf_{n \to \infty} \left(\int_{\mathbb{R}^d \times [0, 1]} L^{(1)}(f^n(t), y) \nu^{(1),n}(dy \times dt) + \int_{\mathbb{R}^d \times [0, 1]} L^{(2)}(f^n(t), y) \nu^{(2),n}(dy \times dt)\right) \\
\geq \int_{\mathbb{R}^d \times [0, 1]} L^{(1)}(\varphi(t), y) \nu^{(1)}(dy \times dt) + \int_{\mathbb{R}^d \times [0, 1]} L^{(2)}(\varphi(t), y) \nu^{(2)}(dy \times dt).
\]

Properties (i), (ii), and (iii) in the previous paragraph, the decompositions \(\nu^{(i)}(dy \times dt) = \nu^{(i)}(dy|t) \otimes dt\), the convexity of \(L^{(i)}(x, \cdot)\), and the definition of \(\bar{L}(x, \beta)\) yield
\[
\int_{\mathbb{R}^d \times [0, 1]} L^{(1)}(\varphi(t), y) \nu^{(1)}(dy \times dt) + \int_{\mathbb{R}^d \times [0, 1]} L^{(2)}(\varphi(t), y) \nu^{(2)}(dy \times dt) \\
\geq \int_{0}^{1} \bar{L}(\varphi(t), \hat{\varphi}(t)) dt.
\]

Combining the last two displays shows that
\[
M \geq \int_{0}^{1} \bar{L}(\varphi(t), \hat{\varphi}(t)) dt = I_x(\varphi).
\]

This completes the proof that \(I_x\) has compact level sets. \(\blacksquare\)

In the next chapter we prove the Laplace principle for the empirical measures of a Markov chain.
CHAPTER 7. RANDOM WALK WITH DISCONTINUOUS STATISTICS
Chapter 8

Laplace Principle for the Empirical Measures of a Markov Chain

8.1 Introduction

In the previous two chapters we applied the weak convergence approach to prove Laplace principles for the random walk model with continuous statistics and with discontinuous statistics. In the present chapter, we apply the approach to the empirical measures of a Markov chain, which is a large deviation problem of a totally different sort.

The problem to be studied in this chapter was introduced in Section 4.2. Let $\mathcal{S}$ be a Polish space, $\mathcal{M}(\mathcal{S})$ the set of Borel subprobability measures on $\mathcal{S}$, and $\mathcal{P}(\mathcal{S})$ the subset of $\mathcal{M}(\mathcal{S})$ consisting of probability measures. Let $\{X_j, j \in \mathbb{N}_0\}$ be a Markov chain that takes values in $\mathcal{S}$ and has stationary transition probabilities. We consider the sequence of empirical measures

$$L^n = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{X_j},$$

which take values in $\mathcal{P}(\mathcal{S})$. The Laplace principle for the sequence $\{L^n, n \in \mathbb{N}\}$ is stated in Theorem 8.4.3 for the weak topology on $\mathcal{P}(\mathcal{S})$ and includes bounds that are uniform over initial conditions in compact subsets of $\mathcal{S}$. The proofs of the Laplace principle bounds follow the pattern of the proofs given for the random walk model in Chapter 6. Our methods also yield results for the $\tau$-topology on $\mathcal{P}(\mathcal{S})$, which will be given in the next chapter.

The pioneering work on large deviations for the empirical measures of a Markov chain is that of Donsker and Varadhan [32, 34]. Our approach to this problem adapts many ideas from the second of these papers. Related works include [18, 22, 23, 28, 29, 31, 60, 74, 75]. All of these authors who prove the large deviation upper bound assume that the transition probability function $p(x, dy)$ of the Markov chain satisfies the Feller property, as we do in our Theorem 8.4.3. The Feller property stipulates that the function mapping $x \in \mathcal{S} \mapsto p(x, \cdot)$ is continuous in the weak topology on $\mathcal{P}(\mathcal{S})$. Hence Theorem 8.4.3 may be regarded as a Laplace principle for Markov chains with continuous statistics. In Section 9.2 of the next chapter we prove the Laplace principle upper bound when the transition probability
function satisfies a weakened version of the Feller property.

Our methods also allow us to prove the Laplace principle for the multivariate empirical measures of the Markov chain. For each \( \alpha \in \mathbb{N} \) satisfying \( \alpha \geq 2 \), the multivariate empirical measures are the empirical measures of the Markov chain \( \{(X_j, X_{j+1}, \ldots, X_{j+\alpha-1}), j \in \mathbb{N}_0\} \). The Laplace principle for the multivariate empirical measures is a consequence of the Laplace principle for the empirical measures proved in Theorem 8.4.3. We omit the proof, which as in [46] is based on the fact that the conditions on the transition probability function assumed in Theorem 8.4.3 are inherited by the transition probability function of the Markov chain \( \{(X_j, X_{j+1}, \ldots, X_{j+\alpha-1}), j \in \mathbb{N}_0\} \).

The point of the present chapter is not just to prove the Laplace principle for the empirical measures, but also to continue to demonstrate the power and naturalness of the weak convergence approach. For example, as in Chapter 6 the calculations leading to a Laplace principle upper bound will lead us, in Section 8.3, the form of the rate function. It does not have to be known beforehand. Nevertheless, the full power of the weak convergence approach is not revealed here. It is readily applicable to much more complicated problems involving measure-valued processes with state dependencies. An example is the random walk model with state-dependent noise, which was discussed in Section 4.4.

We mention a number of related works in the literature. Using a restrictive comparison condition on the transition probability function, Stroock [84] proves a large deviation principle for the empirical measures which is uniform over all initial conditions \( X_0 = x \in \mathcal{S} \). The works [23, 28, 29, 46] give extensions. The references [11, 17, 23, 25, 28, 29] prove the large deviation principle for the empirical measures in the \( \tau \)-topology, a topic that we will consider in the next chapter. Results for other topologies have also been obtained in [33, 55]. Large deviation theorems for the multivariate empirical measures and for infinite dimensional extensions known as the empirical processes have been proved by a number of authors including [11, 23, 28, 29, 35, 46, 49].

In the next section, we introduce a class of admissible control measures that are suited to the asymptotic analysis of the empirical measures. In Section 8.3, we prove the Laplace principle upper bound and identify the form of the rate function. Section 8.4 gives the statement of the Laplace principle, which includes bounds that are uniform over initial conditions in compact subsets of \( \mathcal{S} \). In Section 8.5, we prove several properties of the rate function, including its convexity and the fact that its level sets are compact. Finally, the Laplace principle lower bounds are proved in Section 8.6.

In this chapter we will deviate from the pattern of previous chapters. There, the Laplace principle lower bound was proved in each case under a set of conditions that included the assumptions used for the upper bound. In the present chapter we will follow the prevailing custom in the literature of proving upper and lower bounds under disjoint conditions. One consequence is that the compactness and convergence results used to prove the upper bound are not applied to obtain the lower bound. In their place we use the ergodic theorem, which could have been avoided had we continued our previous practice. In the present setting, the use of two sets of conditions allows one to obtain the lower bound in greater generality. However, the reader should keep in mind that a lower bound without the corresponding upper bound could be far from optimal and
its significance unclear. For this reason we generally prefer the economy achieved by including upper bound assumptions with those for the lower bound and making use of a single set of convergence results.

8.2 Compactness and Limit Properties of Controls and Controlled Processes

Throughout this chapter we use the weak topology on \( \mathcal{P}(S) \) and on \( \mathcal{M}(S) \), recalling that in each space a sequence \( \{ \theta_n, n \in \mathbb{N} \} \) converges in this topology to \( \theta \) if for each bounded continuous function \( g \) mapping \( S \) into \( \mathbb{R} \)

\[
\lim_{n \to \infty} \int_S g \, d\theta_n = \int_S g \, d\theta.
\]

We write \( \theta_n \to \theta \) to indicate this convergence. \( \mathcal{P}(S) \) is metrized by the Lévy–Prohorov metric \( L(\cdot, \cdot) \), which is compatible with the weak topology. With respect to this metric \( \mathcal{P}(S) \) is a Polish space [Theorem A.3.2]. In Section A.3 we introduce a metric on \( \mathcal{M}(S) \) which is also compatible with the weak topology on this space and with respect to which \( \mathcal{M}(S) \) is a Polish space. Given \( \tau \) a probability measure on \( S \times S \), \( \tau_1 \) and \( \tau_2 \) denote, respectively, the first marginal and the second marginal of \( \tau \). For Borel sets \( A \) and \( B \), these measures are defined by

\[
\tau_1(A) \overset{\text{def}}{=} \tau(A \times S) \quad \text{and} \quad \tau_2(B) \overset{\text{def}}{=} \tau(S \times B).
\]

The purpose of this chapter is to prove the Laplace principle for the empirical measures of a Markov chain \( \{ X_j, j \in \mathbb{N}_0 \} \) taking values in \( S \). We denote by \( p(x, dy) \) the transition probability function of the Markov chain and let \( X_0 = x \) be the initial condition. \( E_x \) denotes expectation conditioned on \( X_0 = x \). Theorem 4.2.2 states a representation formula for

\[
W^n(x) \overset{\text{def}}{=} -\frac{1}{n} \log E_x \{ \exp \left[ -n \, h(L^n) \right] \},
\]

where \( h \) is any bounded measurable function mapping \( \mathcal{P}(S) \) into \( \mathbb{R} \). The representation formula for \( W^n(x) \) is stated in terms of the minimal cost function \( V^n(x) \) of an associated optimal control problem, which we now recall. In order to prove the Laplace principle for the empirical measures, we will rewrite the representation for \( W^n(x) \) slightly.

Recapitulation of the optimal control problem. For \( n \in \mathbb{N} \) the control that is applied at time \( j \in \{ 0, 1, \ldots, n - 1 \} \) is a stochastic kernel on \( S \) given \( S \times \mathcal{M}_{j/n}(S) \) which is denoted by \( \nu^n_{j} = \nu^n_{j}(dy|x, \mu) \). Recall that \( \mathcal{M}_{j/n}(S) \overset{\text{def}}{=} \{ \mu \in \mathcal{M}(S) : \mu(S) = j/n \} \). A sequence of controls \( \{ \nu^n_{j}, j = 0, 1, \ldots, n - 1 \} \) is called an admissible control sequence.

At time \( j \in \{ 0, 1, \ldots, n \} \) the controlled process takes values in \( S \times \mathcal{M}_{j/n}(S) \) and is denoted by \( (\bar{X}^n_j, \bar{L}^n_j) \). For \( x \in S \) and all \( n \in \mathbb{N} \) the controlled processes are defined on a common probability space \( (\Omega, \mathcal{F}, \mathbb{P}_x) \). We set \( \bar{X}^n_0 = X_0 = x \) and specify the conditional distribution of \( \bar{X}^n_{j+1} \) to be

\[
\mathbb{P}_x \{ \bar{X}^n_{j+1} \in dy | (\bar{X}^n_i, \bar{L}^n_i), i = 0, 1, \ldots, j \} \overset{\text{def}}{=} \nu^n_{j}(dy|\bar{X}^n_j, \bar{L}^n_j).
\]
The measures $\bar{L}^n_j$ evolves according to $\bar{L}^n_0 = L^n_0 = 0$ and
\[
\bar{L}^n_{j+1} = \bar{L}^n_j + \frac{1}{n} \delta_{X^n_j}.
\] (8.3)

We write $\bar{L}^n_i$ instead of $\bar{L}^n_i$.

Given these dynamics, we define for $x \in \mathcal{S}$ the minimal cost function
\[
V^n(x) \doteq \inf_{\{\nu^n_j\}} \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R\left(\nu^n_j(\cdot|X^n_j, \bar{L}^n_j)|p(X^n_j, \cdot)\right) + h(\bar{L}^n) \right\}.
\] (8.4)

The infimum is taken over all admissible control sequences $\{\nu^n_j\}$, $\mathbb{E}_x$ denotes expectation with respect to $\bar{P}_x$, $\{(X^n_j, \bar{L}^n_j)\}$ is the controlled process that is associated with a particular admissible control sequence $\{\nu^n_j\}$, and $h$ is the same bounded measurable function mapping $\mathcal{P}(\mathcal{S})$ into $\mathbb{R}$ which appears in the definition (8.1) of $W^n(x)$. Theorem 4.2.2 gives the representation formula $W^n(x) = V^n(x)$, which for easy reference is restated here.

**Theorem 8.2.1.** Let $h$ be a bounded measurable function mapping $\mathcal{P}(\mathcal{S})$ into $\mathbb{R}$. Then for all $n \in \mathbb{N}$ and $x \in \mathcal{S}$
\[
W^n(x) \doteq -\frac{1}{n} \log E_x \{\exp [-n h(\bar{L}^n)]\}
\]
equals the minimal cost function $V^n(x)$ defined in formula (8.4).

In order to prove the Laplace principle upper bound, it will be convenient to apply Jensen’s Inequality to the sum of the relative entropy terms appearing in $V^n(x)$. This will replace the sum of the relative entropy terms by a relative entropy involving a certain admissible control measure. The form of this measure will allow us, in Theorem 8.2.8, to easily relate its limits with the limits of the controlled empirical measures $\bar{L}^n_i$. In order to motivate the definition of the control measure, let us recall how we proceeded in the proof of Sanov’s Theorem, where the control measure was defined simply as the time average of the control sequence $\{\nu^n_j\}$ [see (2.16)]. Since the only difference between the representation formulas in the present chapter and in Sanov’s Theorem is that the running cost now depends explicitly on $\bar{X}^n_i$ through $p(\bar{X}^n_i, \cdot)$, one might suspect that now we may also have to keep track of the dependence of the control upon $\bar{X}^n_i$. This turns out to be true. The reader is referred to Section 3.6 for a detailed discussion of the form of the control measure and the use of Jensen’s Inequality in the proof of an arbitrary Laplace principle.

We now define the admissible control measure that will be used in the proof of the Laplace principle upper bound. For $n \in \mathbb{N}$, let $\{\nu^n_j, j = 0, 1, \ldots, n-1\}$ be an admissible control sequence. We define the **admissible control measure** $\nu^n$ to be the random probability measure on $\mathcal{S} \times \mathcal{S}$ given by
\[
\nu^n(dx \times dy) \doteq \frac{1}{n} \sum_{j=0}^{n-1} \delta_{\bar{X}^n_j}(dx) \times \nu^n_j(dy|\bar{X}^n_j, \bar{L}^n_j).
\] (8.5)
The quantity \( \nu^n \) is a stochastic kernel on \( S \times S \) given \( \tilde{\Omega} \), and its marginals are

\[
(v^n)_1 = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{X_j} = \tilde{L}^n \quad \text{and} \quad (v^n)_2 = \frac{1}{n} \sum_{j=0}^{n-1} \nu^n_j.
\] (8.6)

In parallel with our analysis of the random walk model, the proof of the Laplace principle upper bound will require that we prove the tightness of any sequence of admissible control measures \( \{\nu^n\} \) for which the sequence of associated running costs is bounded. It will also require us to relate the limit of any convergent subsequence of \( \{\nu^n\} \) with the limit of the same subsequence of the controlled empirical measures \( \{\tilde{L}^n\} \). These steps are carried out in Proposition 8.2.5 and Theorem 8.2.8, respectively.

In order to perform these steps, we need a strong positive recurrence probability condition on the underlying transition probability function. One such condition that could have been assumed is Hypothesis \( \text{H}^* \) in Section 4 of [34], where it is used to prove the large deviation upper bound. Instead, we will adopt the following closely related condition. It is slightly easier to check because the boundedness in part (c) is on \( U(x) \) rather than on \( \int_S \exp[U(y)] p(x, dy) \) as in [34]. As we will see, if the state space \( S \) is compact, then the following condition is not needed.

**Condition 8.2.2.** There exists a measurable function \( U \) mapping \( S \) into \([0, \infty)\) and having the following properties:

(a) \( \inf_{x \in S} \left\{ U(x) - \log \int_S e^{U(y)} p(x, dy) \right\} > -\infty \).

(b) For each \( M < \infty \), the level set

\[
Z(M) \triangleq \left\{ x \in S : U(x) - \log \int_S e^{U(y)} p(x, dy) \leq M \right\}
\]

is a relatively compact subset of \( S \).

(c) \( U \) is bounded above on every compact subset of \( S \).

The nonnegativity of \( U \) and part (a) of Condition 8.2.2 imply that for each \( x \in S \)

\[
1 \leq \int_S e^{U(y)} p(x, dy) < \infty.
\]

This bound will be used later on.

We next give an example of a Markov chain on \( \mathbb{R}^d \) for which Condition 8.2.2 is satisfied. This example seems to be typical of when the condition can be expected to hold in the case \( S = \mathbb{R}^d \).

**Example 8.2.3.** Let \( p(x, dy) \) be a transition probability function on \( \mathbb{R}^d \). Given a measurable function \( b \) mapping \( \mathbb{R}^d \) into \( \mathbb{R}^d \), we define for \( x \in \mathbb{R}^d \) and \( \alpha \in \mathbb{R}^d \) the shifted cumulant generating function

\[
H_b(x, \alpha) \triangleq \log \int_{\mathbb{R}^d} \exp(\alpha, y - x - b(x)) p(x, dy).
\]
A natural choice for $b$ is 

$$b(x) \triangleq \int_{\mathbb{R}^d} (y - x) p(x, dy),$$

but we are not restricted to this. We make the following assumptions. Assumption (iii) may be considered to be a stability condition on the discrete-time dynamical system $x_{n+1} = x_n + b(x_n)$. At a minimum, $U$ should be a Lyapunov function for the mean dynamics given in terms of $b$. This turns out not to be strong enough, however, and the strengthening in assumption (iii) is needed.

(i) $b$ is bounded on compact subsets of $\mathbb{R}^d$.

(ii) There exists $r > 0$ such that

$$\sup_{x \in \mathbb{R}^d} H_b(x, \alpha) < \infty \text{ for each } \alpha \in \mathbb{R}^d \text{ satisfying } \|\alpha\| \leq r.$$ 

(iii) There exists a Lipschitz continuous function $U$ that maps $\mathbb{R}^d$ into $[0, \infty)$ and satisfies

$$\lim_{\|x\| \to \infty} [U(x + b(x)) - U(x)] = -\infty.$$

We claim that under these three assumptions, $p(x, dy)$ and $U$ satisfy Condition 8.2.2. Since $U$ is Lipschitz continuous, there exists a Lipschitz constant $\kappa < \infty$ such that for all $x$ and $y$ in $\mathbb{R}^d$

$$|U(x) - U(y)| \leq \kappa \|x - y\|.$$ 

Without loss of generality, we can assume that $\kappa \leq r/\sqrt{d}$, where $r > 0$ is the number in assumption (ii). Indeed, if this inequality is not satisfied, then for sufficiently small $v > 0$ the function $vU(x)$ has Lipschitz constant less than or equal to $r/\sqrt{d}$ and satisfies assumption (iii); we continue to write $U$ for this new function. The Lipschitz continuity of $U$ implies that $U$ is bounded above on compact subsets of $\mathbb{R}^d$. This is part (c) of Condition 8.2.2.

We also use the Lipschitz continuity of $U$ to verify part (b) of Condition 8.2.2. For any $x \in \mathbb{R}^d$

$$U(x) - \log \int_{\mathbb{R}^d} e^{U(y)} p(x, dy)$$

$$= U(x) - U(x + b(x)) - \log \int_{\mathbb{R}^d} \exp[U(y) - U(x + b(x))] p(x, dy)$$

$$\geq U(x) - U(x + b(x)) - \log \int_{\mathbb{R}^d} \exp[\kappa \|y - x - b(x)\|] p(x, dy)$$

$$\geq U(x) - U(x + b(x)) - \log \prod_{i=1}^{2d} \exp[\kappa|y_i - x_i - b_i(x)|] p(x, dy)$$

$$\geq U(x) - U(x + b(x)) - \log \sum_{i=1}^{2d} \exp[H_b(x, \kappa \alpha^i)],$$
where \( \{ \alpha^i, i = 1, 2, \ldots, 2^d \} \) are the \( 2^d \) vectors in \( \mathbb{R}^d \) with coordinates 1 or \(-1\). Since \( \kappa \| \alpha^i \| = \kappa \sqrt{d} \leq r \), assumption (ii) yields the existence of \( C < \infty \) such that

\[
U(x) - \log \int_{\mathbb{R}^d} e^{U(y)} p(x, dy) \geq U(x) - U(x + b(x)) - C. \tag{8.7}
\]

Assumption (iii) now guarantees that the level set

\[
Z(M) = \left\{ x \in \mathbb{R}^d : U(x) - \log \int_{\mathbb{R}^d} e^{U(y)} p(x, dy) \leq M \right\}
\]

is a relatively compact subset of \( \mathbb{R}^d \). This verifies part (b) of Condition 8.2.2.

Finally, equation (8.7) and assumption (iii) imply that for some constant \( D < \infty \)

\[
c(x) \doteq U(x) - \log \int_{\mathbb{R}^d} e^{U(y)} p(x, dy)
\]

is bounded below on \( \{ x \in \mathbb{R}^d : \| x \| > D \} \). Since

\[
U(x) - U(x + b(x)) \geq -\kappa \| b(x) \|,
\]

equation (8.7) and assumption (i) guarantee that \( c(x) \) is also bounded below on \( \{ x \in \mathbb{R}^d : \| x \| \leq D \} \). This verifies part (a) of Condition 8.2.2 and completes the discussion of the example. \( \blacksquare \)

Before proving a compactness result for sequences of admissible control measures, we need a result from weak convergence theory which exploits a bootstrap procedure implicit in Prohorov’s Theorem. This result has already been applied in the proof of Proposition 5.3.2. Given \( \mathcal{X} \) a Polish space and \( g \) measurable function mapping \( \mathcal{X} \) into \( \mathbb{R} \cup \{ \infty \} \), we call \( g \) a **tightness function** on \( \mathcal{X} \) if \( \inf_{x \in \mathcal{X}} g(x) > -\infty \) and if for each \( M < \infty \) the level set \( \{ x \in \mathcal{X} : g(x) \leq M \} \) is relatively compact. Given \( g \) a tightness function on \( \mathcal{X} \), we define a function \( G \) mapping \( \mathcal{P}(\mathcal{X}) \) into \( \mathbb{R} \cup \{ \infty \} \) by

\[
G(\theta) \doteq \int_{\mathcal{X}} g \, d\theta.
\]

According to Theorem A.3.17, for each \( M < \infty \) the set \( \{ \theta \in \mathcal{P}(\mathcal{X}) : G(\theta) \leq M \} \) is tight and \( G \) is a tightness function on \( \mathcal{P}(\mathcal{X}) \). This yields the next result, which identifies a tightness function appearing in the statement of Condition 8.2.2.

**Lemma 8.2.4.** We assume that Condition 8.2.2 is satisfied by a measurable function \( U \) mapping \( \mathcal{S} \) into \([0, \infty)\). For \( x \in \mathcal{S} \) we define

\[
c(x) \doteq U(x) - \log \int_{\mathbb{R}^d} e^{U(y)} p(x, dy).
\]

The following conclusions hold.

(a) \( c \) is a tightness function on \( \mathcal{S} \).
(b) For each $M < \infty$ the set
\[
\left\{ \theta \in \mathcal{P}(\mathcal{S}) : \int_{\mathcal{S}} c \, d\theta \leq M \right\}
\]
is tight.

(c) For each $M < \infty$ the set
\[
\left\{ Q \in \mathcal{P}(\mathcal{P}(\mathcal{S})) : \int_{\mathcal{P}(\mathcal{S})} \left( \int_{\mathcal{S}} c \, d\theta \right) Q(d\theta) \leq M \right\}
\]
is tight.

The following compactness result for sequences of admissible control measures will play the same role in the analysis of the empirical measures of a Markov chain which Proposition 5.3.2 played in the analysis of the random walk model. Inequality (8.8) is a bound on the sequence of running costs associated with this sequence of controls. The compactness result will be used in the proof of the Laplace principle to be given later in this chapter. The inequality (8.8) will be automatically satisfied by the admissible control sequences that arise in the course of the proof.

**Proposition 8.2.5.** We assume Condition 8.2.2. Let $K$ be a compact subset of $\mathcal{S}$ and for $n \in \mathbb{N}$ let $\bar{X}_0^n = x^n$ be any sequence of initial conditions in $K$. For each $n$ consider any admissible control sequence $\{\nu^n_j\}$ such that
\[
\sup_{n \in \mathbb{N}} E_{\bar{x}^n} \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot | \bar{X}_j^n, \bar{L}_j^n)\|p(\bar{X}_j^n, \cdot)) \right\} \equiv \Delta < \infty.
\]  
(8.8)

Let the sequence of admissible control measures $\{\nu^n, n \in \mathbb{N}\}$ on $\mathcal{S} \times \mathcal{S}$ be defined by formula (8.5). Then $\{\nu^n\}$ is tight.

**Remark 8.2.6.** If $\mathcal{S}$ is compact, then $\mathcal{P}(\mathcal{S})$ is compact as is $\mathcal{P}(\mathcal{S} \times \mathcal{S})$, in which the admissible control measures $\nu^n$ take values. Hence in the case of compact $\mathcal{S}$, the tightness of any sequence of admissible control measures $\{\nu^n\}$ is valid without Condition 8.2.2 and without (8.8). ■

**Proof of Proposition 8.2.5.** The heart of the proof is to show that for each $\varepsilon > 0$ there exist compact sets $A$ and $B$ in $\mathcal{P}(\mathcal{S})$ satisfying
\[
\inf_{n \in \mathbb{N}} \bar{P}_{\bar{x}^n} \left\{ \bar{L}_n \in A \right\} \geq 1 - \varepsilon
\]  
(8.9)
and
\[
\inf_{n \in \mathbb{N}} \bar{P}_{\bar{x}^n} \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \nu^n_j \in B \right\} \geq 1 - \varepsilon.
\]  
(8.10)
8.2. COMPACTNESS AND LIMIT PROPERTIES

We will carry this out in a moment. Assuming that it is true, we can easily complete the proof. Prohorov’s Theorem guarantees that for each $\delta > 0$ there exist compact sets $K$ and $\Lambda$ in $\mathcal{S}$ such that $\tau(K) \geq 1 - \delta$ for each $\tau \in A$ and $\tau(\Lambda) \geq 1 - \delta$ for each $\tau \in B$. Then

$$C = \{\tau \in \mathcal{P}(\mathcal{S} \times \mathcal{S}) : \tau_1 \in A, \tau_2 \in B\}$$

is closed, and $\tau(K \times \Lambda) \geq 1 - 2\delta$ for all $\tau \in C$. It follows again by Prohorov’s Theorem that $C$ is compact. Furthermore, by formulas (8.6), (8.9), and (8.10)

$$\inf_{n \in \mathbb{N}} \hat{P}_x^n \{\nu^n \in C\} \geq 1 - 2\varepsilon.$$

Thus the inequalities (8.9) and (8.10) will imply that $\{\nu^n\}$ is tight.

We first consider (8.9). For any bounded measurable function $\psi$ mapping $\mathcal{S}$ into $\mathbb{R}$, any $\gamma \in \mathcal{P}(\mathcal{S})$, and any $\xi \in \mathcal{S}$, the Donsker–Varadhan variational formula for the relative entropy [Lemma 1.4.3 (a)] implies that

$$\int_{\mathcal{S}} \psi(y) \gamma(dy) - \log \int_{\mathcal{S}} e^{\psi(y)} p(\xi, dy) \leq R(\gamma(\cdot)\|p(\xi, \cdot)).$$

In particular, this inequality holds for $\psi = U \wedge n$, $n \in \mathbb{N}$, where $U$ is the nonnegative function appearing in Condition 8.2.2. By the Monotone Convergence Theorem and the bound

$$\int_{\mathcal{S}} e^{U(y)} p(\xi, dy) < \infty,$$

which was pointed out after the statement of Condition 8.2.2, the inequality also holds for $U$ itself. Thus

$$\int_{\mathcal{S}} U(y) \gamma(dy) - \log \int_{\mathcal{S}} e^{U(y)} p(\xi, dy) \leq R(\gamma(\cdot)\|p(\xi, \cdot)). \tag{8.11}$$

For $n \in \mathbb{N}$ and $j \in \{0, 1, \ldots, n - 1\}$ we define $\mathcal{F}_n^n$ to be the $\sigma$–field generated by $\{(X_i^n, L_i^n), i = 0, 1, \ldots, j\}$. Since $\nu_j^n(dy|X_j^n, L_j^n)$ is a regular conditional distribution for $X_{j+1}^n$ given $\mathcal{F}_j^n$,

$$\hat{E}_x^n\{g(X_{j+1}^n)|\mathcal{F}_j^n\} = \int_{\mathcal{S}} g(y) \nu_j^n(dy|X_j^n, L_j^n) \tag{8.12}$$

holds $\hat{P}_x^n$–a.s. when $g \equiv U$. Define

$$c(x) \equiv U(x) - \log \int_{\mathcal{S}} e^{U(y)} p(x, dy).$$

Then for each $j \in \{0, 1, \ldots, n - 1\}$

$$\hat{E}_x^n\{U(X_{j+1}^n) - U(X_j^n)\}$$

$$= \hat{E}_x^n\{\hat{E}_x^n\{U(X_{j+1}^n) - U(X_j^n)|\mathcal{F}_j^n\}\}\$$

$$= \hat{E}_x^n\left\{\int_{\mathcal{S}} U(y) \nu_j^n(dy|X_j^n, L_j^n) - U(X_j^n)\right\}$$

$$= \hat{E}_x^n\left\{\int_{\mathcal{S}} U(y) \nu_j^n(dy|X_j^n, L_j^n) - \log \int_{\mathcal{S}} e^{U(y)} p(X_j^n, dy)\right\}$$

$$+ \hat{E}_x^n\left\{\log \int_{\mathcal{S}} e^{U(y)} p(X_j^n, dy) - U(X_j^n)\right\}$$

$$\leq \hat{E}_x^n\left\{R\left(\nu_j^n(\cdot|X_j^n, L_j^n)\|p(X_j^n, \cdot)\right)\right\} - \hat{E}_x^n\{c(X_j^n)\}.$$
Summing these inequalities over \( j \in \{0, 1, \ldots, n - 1\} \) yields
\[
\mathbb{E}_x^n \left\{ U(\bar{X}_n^x) - U(x^n) \right\} \\
\leq \mathbb{E}_x^n \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R \left( \varphi^n_j(\cdot|\bar{X}_j^n, \bar{L}_j^n) || p(\bar{X}_j^n, \cdot) \right) \right\} - \mathbb{E}_x^n \left\{ \frac{1}{n} \sum_{j=0}^{n-1} c(\bar{X}_j^n) \right\}.
\]
Since by assumption \( U \) is nonnegative, it follows that for each \( n \in \mathbb{N} \)
\[
\mathbb{E}_x^n \left\{ \int_S c \, d\bar{L}_n \right\} = \mathbb{E}_x^n \left\{ \frac{1}{n} \sum_{j=0}^{n-1} c(\bar{X}_j^n) \right\} \\
\leq \mathbb{E}_x^n \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R \left( \varphi^n_j(\cdot|\bar{X}_j^n, \bar{L}_j^n) || p(\bar{X}_j^n, \cdot) \right) \right\} + \frac{1}{n} U(x^n) \\
\leq \Delta + \frac{1}{n} U(x^n).
\]
Thus, since \( \sup_{n \in \mathbb{N}} U(x^n) \leq \sup_{x \in K} U(x) < \infty \),
\[
\mathbb{E}_x^n \left\{ \int_S c \, d\bar{L}_n \right\} = \sup_{n \in \mathbb{N}} \mathbb{E}_x^n \left\{ \int_S c \, d\mu \right\} P_x^n \left( L^n \in d\mu \right) \leq \Delta + \sup_{n \in \mathbb{N}} U(x^n) < \infty.
\]
(8.13)
Part (c) of Lemma 8.2.4 now implies that the sequence \( \{\bar{L}_n \in d\mu\}, n \in \mathbb{N} \) is tight. Thus for each \( \varepsilon > 0 \), there exists a compact set \( A \) in \( \mathcal{P}(S) \) satisfying (8.9).

We now consider (8.10), which states that the sequence of random measures \( \{\frac{1}{n} \sum_{j=0}^{n-1} \varphi^n_j\} \) is tight. It is essentially a consequence of Lemma 8.2.4 and the tightness of the sequence \( \{\bar{L}_n\} \), which has just been proved. Given the initial condition \( \bar{X}_0^n = x^n \), we define the probability measure \( \nu^n_{-1} \) on \( S \) by
\[
\nu^n_{-1}(dy) = \delta_{x^n}(dy).
\]
We also define the random probability measure \( \lambda^n \) on \( S \) by
\[
\lambda^n(dy) = \frac{1}{n} \left( \nu^n_{-1}(dy) + \sum_{j=0}^{n-2} \nu^n_j(dy|\bar{X}_j^n, \bar{L}_j^n) \right).
\]
In the next paragraph the sequence \( \{\lambda^n\} \) will be shown to be tight. Assuming this, let us denote by \( \| \cdot \|_v \) the total variation norm on \( \mathcal{P}(S) \). The tightness of \( \{\lambda^n\} \) will imply its relative compactness, and since for each \( n \in \mathbb{N} \)
\[
\left\| \frac{1}{n} \sum_{j=0}^{n-1} \varphi^n_j(\cdot|\bar{X}_j^n, \bar{L}_j^n) - \lambda^n(\cdot) \right\|_v = \frac{1}{n} \left\| \nu^n_{-1}(\cdot|\bar{X}_{n-1}^n, \bar{L}_{n-1}^n) - \nu^n_{-1}(\cdot) \right\|_v \leq \frac{2}{n},
\]
the relative compactness of \( \{\lambda^n\} \) will imply the relative compactness of \( \{\frac{1}{n} \sum_{j=0}^{n-1} \varphi^n_j\} \). Prohorov’s Theorem will then imply the existence, for each \( \varepsilon > 0 \), of a compact set \( B \) in \( \mathcal{P}(S) \) satisfying (8.10).
We now prove that \( \{ \lambda^n \} \) is tight. By (8.12) with \( g = c \) and (8.13)

\[
\sup_{n \in \mathbb{N}} E_n \left\{ \int_S c \, d\lambda^n \right\} = \sup_{n \in \mathbb{N}} E_n \left\{ \frac{1}{n} \sum_{j=-1}^{n-2} \int_S c \, d\nu^j \right\} \\
= \sup_{n \in \mathbb{N}} E_n \left\{ \frac{1}{n} \sum_{j=0}^{n-1} c(\bar{X}^n_j) \right\} = \sup_{n \in \mathbb{N}} E_n \left\{ \int_S c \, d\tilde{L}^n \right\} \leq \Delta + \sup_{n \in \mathbb{N}} U(x^n) < \infty.
\]

Thus

\[
\sup_{n \in \mathbb{N}} E_n \left\{ \int_S c \, d\lambda^n \right\} = \sup_{n \in \mathbb{N}} \int_{\mathcal{P}(S)} \left( \int_S c \, d\mu \right) \tilde{P}_n \{ \lambda^n \in d\mu \} < \infty. \tag{8.14}
\]

Again, part (c) of Lemma 8.2.4 implies that \( \{ \lambda^n \} \) is tight. The proof of the lemma is complete. \( \blacksquare \)

The next lemma will be used to prove Theorem 8.2.8 and will also be needed in Section 9.3.

**Lemma 8.2.7.** Let \( g \) be any bounded measurable function mapping \( S \) into \( \mathbb{R} \). Then for any \( \varepsilon > 0 \), any sequence \( \{ x^n, n \in \mathbb{N} \} \) in \( S \), and any \( n \geq 4\|g\|_{\infty}/\varepsilon \)

\[
\tilde{P}_n \left\{ \left| \int_S g(y) \tilde{L}^n(dy) - \int_{S \times S} g(y) \nu^n(dx \times dy) \right| \geq \varepsilon \right\} = \tilde{P}_n \left\{ \left| \int_S g(y) \tilde{L}^n(dy) - \int_S g(y) (\nu^n)^2(dy) \right| \geq \varepsilon \right\} \leq \frac{16\|g\|_{\infty}^2}{n^2},
\]

where for any Borel set \( A \) \( (\nu^n)^2(A) = \nu^n(S \times A) \).

**Proof.** For \( n \in \mathbb{N} \) and \( j \in \{0, 1, \ldots, n - 1\} \), we define \( \mathcal{F}_j^n \) to be the \( \sigma \)-field generated by \( \{(\bar{X}^n_i, \bar{L}^n_i), i = 0, 1, \ldots, j\} \). As noted in equation (8.12), for each bounded measurable function \( g \) mapping \( S \) into \( \mathbb{R} \)

\[
\tilde{E}_n \left\{ g(\bar{X}^n_{j+1}) - \int_S g(y) \nu^n(dy | \bar{X}^n_j, \bar{L}^n_j) \right| \mathcal{F}_j^n \} = 0
\]

\( P_n \)-a.s. This implies that the sequence

\[
\left\{ g(\bar{X}^n_{j+1}) - \int_S g(y) \nu^n(dy | \bar{X}^n_j, \bar{L}^n_j), j = 0, 1, \ldots, n - 1 \right\}
\]

forms a martingale difference sequence with respect to \( \{\mathcal{F}_j^n, j = 0, 1, \ldots, n - 1\} \). Since

\[
\int_S g(y) \tilde{L}^n(dy) - \int_{S \times S} g(y) \nu^n(dx \times dy)
\]

\[
= \frac{1}{n} \sum_{j=0}^{n-1} \left( g(\bar{X}^n_{j+1}) - \int_S g(y) \nu^n(dy | \bar{X}^n_j, \bar{L}^n_j) \right) + \frac{1}{n} \left( g(\bar{X}^n_0) - g(\bar{X}^n_n) \right)
\]
and \( |g(\bar{X}_n^n) - g(\overline{X}_n^n)| \leq 2\|g\|_{\infty} \), Chebyshev’s Inequality and a standard conditioning argument yield for any \( \varepsilon > 0 \) and all \( n \geq 4\|g\|_{\infty}/\varepsilon \)

\[
P_{x^n}\left\{ \left| \int_{\mathcal{S}} g(y) \, \bar{L}_n(dy) - \int_{\mathcal{S} \times \mathcal{S}} g(y) \, \nu^n(dx \times dy) \right| \geq \varepsilon \right\} \\
\leq P_{x^n}\left\{ \frac{1}{n} \sum_{j=0}^{n-1} \left| g(\bar{X}_{j+1}^n) - \int_{\mathcal{S}} g(y) \, \nu^n_j(dy|\overline{X}_j^n, \bar{L}_j^n) \right| \geq \varepsilon/2 \right\} \\
\leq \frac{4}{\varepsilon^2} \overline{E}_{x^n} \left\{ \frac{1}{n^2} \sum_{j=0}^{n-1} \left( g(\bar{X}_{j+1}^n) - \int_{\mathcal{S}} g(y) \, \nu^n_j(dy|\overline{X}_j^n, \bar{L}_j^n) \right)^2 \right\} \\
= \frac{4}{\varepsilon^2} \overline{E}_{x^n} \left\{ \frac{1}{n^2} \sum_{j=0}^{n-1} \left( g(\bar{X}_{j+1}^n) - \int_{\mathcal{S}} g(y) \, \nu^n_j(dy|\overline{X}_j^n, \bar{L}_j^n) \right)^2 \right\} \\
\leq \frac{16\|g\|_{\infty}^2}{n\varepsilon^2}.
\]

This completes the proof. \( \blacksquare \)

Proposition 8.2.5 proves the tightness of any sequence of admissible control measures for which the sequence of running costs is bounded. Given the tightness of these measures, the next step is to understand the convergence properties of the control measures and the controlled empirical measures. These properties are stated in Theorem 8.2.8. This theorem is the analogue of Theorem 5.3.5, which proves convergence properties of the admissible control measures and the controlled processes for the random walk model. The hearts of both theorems are elementary martingale bounds such as the one just proved in Lemma 8.2.7.

**Theorem 8.2.8.** We assume Condition 8.2.2. Let \( K \) be a compact subset of \( \mathcal{S} \) and for \( n \in \mathbb{N} \) let \( \overline{X}_n^n = x^n \) be any sequence of initial conditions in \( K \). We define the controlled processes \( \{(\bar{X}_j^n, \bar{L}_j^n)\} \) by formulas (8.2) and (8.3). For each \( n \in \mathbb{N} \) consider any admissible control sequence \( \{\nu^n_j\} \) such that

\[
\sup_{n \in \mathbb{N}} \overline{E}_{x^n} \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot|\bar{X}_j^n, \bar{L}_j^n))\|p(\bar{X}_j^n, \cdot)\) \right\} < \infty. \tag{8.15}
\]

Let the sequence of admissible control measures \( \{\nu^n, n \in \mathbb{N}\} \) on \( \mathcal{S} \times \mathcal{S} \) be defined by equation (8.5). Then every subsequence of \( \{(\nu^n, \bar{L}_n, x^n), n \in \mathbb{N}\} \) has a subsequence that converges in distribution to a triplet \( (\nu, \bar{L}, x) \) having the following properties.

(a) There exists a probability space \( (\Omega, \mathcal{F}, \tilde{P}_x) \) such that the limiting quantities \( \nu \) and \( \bar{L} \) can be realized respectively as a stochastic kernel on \( \mathcal{S} \times \mathcal{S} \) given \( \tilde{\Omega} \) and as a stochastic kernel on \( \mathcal{S} \) given \( \tilde{\Omega} \). \( \tilde{P}_x \)-a.s. for \( \omega \in \tilde{\Omega} \) \( \bar{L}(dx|\omega) \) is the first marginal of \( \nu(dx \times dy|\omega) \); i.e., for any Borel set \( A \)

\[
\tilde{L}(A|\omega) = \nu(A \times \mathcal{S}|\omega).
\]
8.2. COMPACTNESS AND LIMIT PROPERTIES

(b) There exists a stochastic kernel $\nu(dy|x) = \nu(dy|x,\omega)$ on $S$ given $S \times \bar{\Omega}$ such that $\bar{P}_x$-a.s. for $\omega \in \bar{\Omega}$

$$\nu(A \times B|\omega) = \int_A \nu(B|x,\omega) \bar{L}(dx|\omega)$$

for all Borel sets $A$ and $B$.

(c) For each $\omega \in \bar{\Omega}$ $q(x,dy|\omega) = \nu(dy|x,\omega)$ is a transition probability function on $S$ which $\bar{P}_x$-a.s. for $\omega \in \bar{\Omega}$ satisfies

$$\bar{L}(B|\omega) = \int_S q(x,B|\omega) \bar{L}(dx|\omega) = \int_S \nu(B|x,\omega) \bar{L}(dx|\omega) = \nu_2(B|\omega).$$

for all Borel sets $B$. In other words, $\bar{P}_x$-a.s. for $\omega \in \bar{\Omega}$ $\bar{L}(dx|\omega)$ is an invariant measure of $q(x,dy|\omega)$ and $\bar{L}(dx|\omega)$ is the second marginal of $\nu(dx \times dy|\omega)$. In the sequel $\omega$ will be suppressed in the notation for $\nu(dx \times dy)$, $\bar{L}(dx)$, $\nu(dy|x)$, and $q(x,dy)$ and the display in part (b) will be summarized as

$$\nu(dx \times dy) = \bar{L}(dx) \otimes \nu(dy|x) = \bar{L}(dx) \otimes q(x,dy).$$

Proof. Since (8.15) coincides with the bound (8.8) in Proposition 8.2.5, we may apply the proposition to deduce that any subsequence of $\{\nu^n\}$ is tight. By Prohorov’s Theorem, there exists a subsequence converging in distribution to a random variable $\nu$. Since the mapping that takes $\nu^n$ into $(\nu^n, (\nu^n)_1) = (\nu^n, \bar{L}^n)$ is continuous, $(\nu^n, \bar{L}^n) \overset{D}{\rightarrow} (\nu, \bar{L})$ [Theorem A.3.6] and w.p.1 $\bar{L}$ equals the first marginal of $\nu$. In addition, since $\{x^n\}$ lies in the compact set $K$, there exists $x \in K$ and a further subsequence such that $(\nu^n, \bar{L}^n, x^n) \overset{D}{\rightarrow} (\nu, \bar{L}, x)$. Let $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{P}_x)$ be a probability space such that $\nu$ is a stochastic kernel on $S \times S$ given $\bar{\Omega}$. Because $\bar{L}(dx) = \bar{L}(dx|\omega)$ is w.p.1 the first marginal of $\nu(dx \times dy|\omega)$, Theorem A.5.6 guarantees that there exists a stochastic kernel $\nu(dy|x) = \nu(dy|x,\omega)$ on $S$ given $S \times \bar{\Omega}$ for which the decomposition given in part (b) of the theorem is valid $\bar{P}_x$-a.s. for $\omega \in \bar{\Omega}$.

In order to complete the proof of the theorem, we must verify that $\bar{L}(dx)$ and $\nu(dy|x)$ are related as in part (c). For the rest of the proof we fix a convergent subsequence of $\{\nu^n, \bar{L}^n, n \in \mathbb{N}\}$, retaining $n$ as the index. According to Theorem A.6.1, $S$ admits an equivalent metric $m(x,y)$ with the property that the space of bounded, uniformly continuous functions $\mathcal{U}_b(S,m)$ is separable. Let $\Xi$ be a countable dense subset of $\mathcal{U}_b(S,m)$.

The previous lemma implies that for each $g \in \Xi$

$$\int_S g(y) \bar{L}^n(dy) - \int_{S \times S} g(y) \nu^n(dx \times dy) \overset{P}{\rightarrow} 0$$

and thus by Theorem A.3.7

$$\int_S g(y) \bar{L}^n(dy) - \int_{S \times S} g(y) \nu^n(dx \times dy) \overset{D}{\rightarrow} 0.$$

We have already proved that $(\nu^n, \bar{L}^n) \overset{D}{\rightarrow} (\nu, \bar{L})$. By the Skorohod Representation Theorem, we can assume that w.p.1 the following three limits are valid for all $g \in \Xi$:

$$\lim_{n \to \infty} \left( \int_S g(y) \bar{L}^n(dy) - \int_{S \times S} g(y) \nu^n(dx \times dy) \right) = 0,$$
\[
\lim_{n \to \infty} \int_{\mathcal{S}} g(y) \, \tilde{L}^n(dy) = \int_{\mathcal{S}} g(y) \, \tilde{L}(dy), \quad \lim_{n \to \infty} \int_{\mathcal{S} \times \mathcal{S}} g(y) \, \nu^n(dx \times dy) = \int_{\mathcal{S} \times \mathcal{S}} g(y) \, \nu(dx \times dy).
\]

Combining these limits and using the decomposition given in part (b) of the theorem, we conclude that w.p.1, for all \( g \in \Xi \) and hence all \( g \in \mathcal{U}_b(\mathcal{S}, m) \),

\[
\int_{\mathcal{S}} g(y) \, \tilde{L}(dy) = \int_{\mathcal{S} \times \mathcal{S}} g(y) \, \nu(dx \times dy) = \int_{\mathcal{S} \times \mathcal{S}} g(y) \, \nu(dy|x) \, \tilde{L}(dx).
\]

For probability measures \( \alpha \) and \( \beta \) on \( \mathcal{S} \) the equality \( \int_{\mathcal{S}} g \, d\alpha = \int_{\mathcal{S}} g \, d\beta \) for all \( g \in \mathcal{U}_b(\mathcal{S}, m) \) implies that \( \alpha = \beta \) [Theorem A.2.2 (b)]. It follows that w.p.1

\[
\tilde{L}(B) = \int_{\mathcal{S}} \nu(B|x) \, \tilde{L}(dx) = \int_{\mathcal{S}} \tilde{q}(x, B) \, \tilde{L}(dx) = \nu_2(B)
\]

for all Borel sets \( B \). The proof of the theorem is complete. \( \blacksquare \)

### 8.3 Proof of the Laplace Principle Upper Bound and the Identification of the Rate Function

For the purposes of proving a Laplace principle upper bound and identifying the rate function, we will impose on \( p(x, dy) \) the following condition known as the Feller property. In Section 9.2 we will introduce another condition that will allow us to weaken the Feller property.

**Condition 8.3.1 (Feller property).** The function mapping \( x \in \mathcal{S} \mapsto p(x, \cdot) \in \mathcal{P}(\mathcal{S}) \) is continuous in the topology of weak convergence on \( \mathcal{P}(\mathcal{S}) \); i.e., if \( \{x_n, n \in \mathbb{N}\} \) is any sequence in \( \mathcal{S} \) such that \( x_n \to x \in \mathcal{S} \), then \( p(x_n, \cdot) \to p(x, \cdot) \).

Our main aim in this section is to identify a function \( I \) mapping \( \mathcal{P}(\mathcal{S}) \) into \([0, \infty]\) for which the Laplace principle upper bound is valid. That is, for each point \( x \in \mathcal{S} \) and each bounded continuous function \( h \) mapping \( \mathcal{S} \) into \( \mathbb{R} \) we want the upper bound

\[
\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{E}_x \{\exp \{-n \, h(L^n)\}\} \leq - \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}.
\]

This upper bound is equivalent to the lower limit

\[
\liminf_{n \to \infty} W^n(x) \geq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}, \quad (8.16)
\]

where

\[
W^n(x) \doteq - \frac{1}{n} \log \mathbb{E}_x \{\exp \{-n \, h(L^n)\}\}.
\]

The form of \( I \) will literally be handed to us in the course of proving this lower limit. As in the proof of the corresponding lower limit in our treatment of Sanov’s Theorem, the present proof will be greatly simplified by the use of Jensen’s Inequality. Throughout this section we assume Conditions 8.2.2 and 8.3.1.
8.3. PROOF OF UPPER BOUND

We prove (8.16) using the representation formula for \( W^n(x) \) given in Theorem 8.2.1. This representation formula states that

\[
W^n(x) = V^n(x) = \inf_{\nu^n} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot|\bar{X}_j^n, \bar{L}_j^n)\|p(\bar{X}_j^n, \cdot)) + h(\bar{L}^n) \right\},
\]

the infimum being taken over all admissible control sequences \( \{\nu^n_j\} \). It suffices to prove (8.16) when \( n \) is replaced by any subsequence along which the functions \( W^n(x) \) converge. Such a sequence exists since \( |W^n(x)| \leq \|h\|_\infty \). We will work with a fixed such subsequence for the remainder of the proof, indexing it, as usual, by \( n \in \mathbb{N} \).

Let \( \varepsilon > 0 \) be given. For each \( n \in \mathbb{N} \) let \( \{\nu^n_j\} \) be an admissible control sequence satisfying

\[
V^n(x) + \varepsilon \geq E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot|\bar{X}_j^n, \bar{L}_j^n)\|p(\bar{X}_j^n, \cdot)) + h(\bar{L}^n) \right\}.
\]

(8.17)

We recall the admissible control measure \( \nu^n \) defined in terms of this admissible control sequence by

\[
\nu^n(A \times B) = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{\bar{X}_j^n}(A) \nu^n_j(B|\bar{X}_j^n, \bar{L}_j^n)
\]

for Borel sets \( A \) and \( B \). The bound \( \sup_{n \in \mathbb{N}} V^n(x) = \sup_{n \in \mathbb{N}} W^n(x) \leq \|h\|_\infty \) implies that

\[
\sup_{n \in \mathbb{N}} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot|\bar{X}_j^n, \bar{L}_j^n)\|p(\bar{X}_j^n, \cdot)) \right\} < \infty.
\]

Hence we may apply Theorem 8.2.8 with \( x^n = x \) for each \( n \in \mathbb{N} \). The theorem implies that there exists a subsequence of \( \{(\nu^n, \bar{L}^n)\} \) satisfying

\[
(\nu^n, \bar{L}^n) \overset{D}{\to} (\nu, \bar{L}).
\]

The limiting quantities \( \nu \) and \( \bar{L} \) are respectively a stochastic kernel on \( \mathcal{S} \times \mathcal{S} \) given \( \bar{\Omega} \) and a stochastic kernel on \( \mathcal{S} \) given \( \bar{\Omega} \), where \((\bar{\Omega}, \bar{\mathcal{F}}, \bar{P}_x)\) is some probability space. By the Skorohod Representation Theorem we can assume that the convergence occurs w.p.1.

We now rewrite formula (8.17) in order that limits can conveniently be taken as \( n \to \infty \). This procedure will allow us to rewrite the running cost in terms of the admissible control measures, whose compactness and limit properties were studied in the previous section. Given \( \alpha, \beta, \) and \( \gamma \) probability measures on \( \mathcal{S} \), we denote by \( \alpha \times \beta \) and \( \alpha \times \gamma \) the product measures on \( \mathcal{S} \times \mathcal{S} \) that assign to the measurable rectangle \( A \times B \) the values \( \alpha(A) \cdot \beta(B) \) and \( \alpha(A) \cdot \gamma(B) \), respectively. According to Corollary C.3.3

\[
R(\beta\|\gamma) = R(\alpha\|\alpha) + R(\beta\|\gamma) = R(\alpha \times \beta\|\alpha \times \gamma).
\]

Applying this formula in the present context yields

\[
R(\nu^n_j(\cdot|\bar{X}_j^n, \bar{L}_j^n)\|p(\bar{X}_j^n, \cdot)) = R(\delta_{\bar{X}_j^n}(\cdot) \times \nu^n_j(\cdot|\bar{X}_j^n, \bar{L}_j^n)\|\delta_{\bar{X}_j^n}(\cdot) \times p(\bar{X}_j^n, \cdot)).
\]
CHAPTER 8. EMPIRICAL MEASURES OF A MARKOV CHAIN

For any probability measure $\theta$ on $\mathcal{S}$ and any transition probability function $q(x, dy)$ on $\mathcal{S}$, we define the probability measure $\theta \otimes q = \theta(dx) \otimes q(x, dy)$ on $\mathcal{S} \times \mathcal{S}$ by

$$(\theta \otimes q)(A \times B) = \int_{A \times B} \theta(dx) q(x, dy) = \int_A q(x, B) \theta(dx)$$

for Borel sets $A$ and $B$. In terms of this notation

$$\delta_{\tilde{X}^n}(dx) \times p(\tilde{X}^n_j, dy) = \delta_{\tilde{X}^n_j}(dx) \otimes p(x, dy),$$

and thus

$$V^n(x) + \varepsilon \geq E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R\left( \nu^n_j(\cdot|\tilde{X}^n_j, \tilde{L}^n_j) \parallel p(\tilde{X}^n_j, \cdot) \right) + h(\tilde{L}^n) \right\} = E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \delta_{\tilde{X}^n_j}(dx) \times \nu^n_j(dy|\tilde{X}^n_j, \tilde{L}^n_j) \parallel \delta_{\tilde{X}^n_j}(dx) \otimes p(x, dy) \right\} + h(\tilde{L}^n).$$

We now apply Jensen’s Inequality to the convex function $R(\cdot \parallel \cdot)$, obtaining

$$V^n(x) + \varepsilon \geq E_x \left\{ R\left( \frac{1}{n} \sum_{j=0}^{n-1} \delta_{\tilde{X}^n_j}(dx) \times \nu^n_j(dy|\tilde{X}^n_j, \tilde{L}^n_j) \parallel \frac{1}{n} \sum_{j=0}^{n-1} \delta_{\tilde{X}^n_j}(dx) \otimes p(x, dy) \right) + h(\tilde{L}^n) \right\} = E_x \left\{ R(\nu^n \parallel \tilde{L}^n \otimes p) + h(\tilde{L}^n) \right\}.$$

The limit inferior of $W^n(x)$ is now evaluated along the subsequence of $n \in \mathbb{N}$ for which $(\nu^n, \tilde{L}^n) \to (\nu, \tilde{L})$ w.p.1. This evaluation is carried out in the following display, which will lead immediately to the required lower bound (8.16):

$$\lim_{n \to \infty, W^n(x) + \varepsilon = \lim_{n \to \infty} V^n(x) + \varepsilon \geq \liminf_{n \to \infty} E_x \left\{ R(\nu^n \parallel \tilde{L}^n \otimes p) + h(\tilde{L}^n) \right\} \geq E_x \left\{ R(\nu \parallel \tilde{L} \otimes p) + h(\tilde{L}) \right\} = E_x \left\{ R(\tilde{L} \otimes \tilde{q} \parallel \tilde{L} \otimes p) + h(\tilde{L}) \right\} = E_x \left\{ \int_\mathcal{X} R(\tilde{q}(\xi, \cdot)) \parallel p(\xi, \cdot) \tilde{L}(d\xi) + h(\tilde{L}) \right\}.$$

The first two lines are consequences of the representation formula and (8.18). With probability 1 $\nu^n \to \nu$ and $\tilde{L}^n \to \tilde{L}$, and we will prove in a moment that the Feller property implies that w.p.1 $\tilde{L}^n \otimes p \to \tilde{L} \otimes p$ [Lemma 8.3.2]. Since $R(\cdot \parallel \cdot)$ is lower semicontinuous and $h$ is continuous, w.p.1

$$\liminf_{n \to \infty} R(\nu^n \parallel \tilde{L}^n \otimes p) \geq R(\nu \parallel \tilde{L} \otimes p) \text{ and } \lim_{n \to \infty} h(\tilde{L}^n) = h(\tilde{L}).$$

Fatou’s Lemma and the Lebesgue Dominated Convergence Theorem now yield line three of (8.19). The equality on line four uses the decomposition $\nu(dx \times dy) = \tilde{L}(dx) \otimes \tilde{q}(x, dy)$ given in Theorem 8.2.8. Finally, the last line is implied by part (f) of Lemma 1.4.3.

We use the probability–1 convergence $\tilde{L}^n \to \tilde{L}$ and the following lemma to show that $\tilde{L}^n \otimes p \to \tilde{L} \otimes p$ w.p.1. This will complete the proof of (8.19).
Lemma 8.3.2. We assume that \( p(x, dy) \) satisfies the Feller property (Condition 8.3.1). If \( \{\mu_n, n \in \mathbb{N}\} \) is a sequence of probability measures on \( \mathcal{S} \) converging weakly to \( \mu \), then \( \mu_n \otimes p \Longrightarrow \mu \otimes p \) on \( \mathcal{S} \times \mathcal{S} \).

Proof. For any bounded continuous function \( f \) mapping \( \mathcal{S} \times \mathcal{S} \) into \( \mathbb{R} \), Theorem A.3.18 shows that the function mapping

\[
(\xi, \gamma) \in \mathcal{S} \times \mathcal{P}(\mathcal{S}) \longrightarrow \int_{\mathcal{S}} f(\xi, b) \gamma(db)
\]

is bounded and continuous. In combination with the Feller property of \( p(x, dy) \), this implies that the function mapping

\[
\xi \in \mathcal{S} \longrightarrow \int_{\mathcal{S}} f(\xi, y) p(\xi, dy)
\]

is bounded and continuous. Since \( \mu_n \Longrightarrow \mu \),

\[
\lim_{n \to \infty} \int_{\mathcal{S} \times \mathcal{S}} f(\xi, y) (\mu_n \otimes p)(d\xi \times dy) = \lim_{n \to \infty} \int_{\mathcal{S}} \left( \int_{\mathcal{S}} f(\xi, y) p(\xi, dy) \right) \mu_n(d\xi)
\]

\[
= \int_{\mathcal{S}} \left( \int_{\mathcal{S}} f(\xi, y) p(\xi, dy) \right) \mu(d\xi)
\]

\[
= \int_{\mathcal{S} \times \mathcal{S}} f(\xi, y) (\mu \otimes p)(d\xi \times dy).
\]

This completes the proof. \( \blacksquare \)

Sending \( \varepsilon \to 0 \) in (8.19), we have proved that

\[
\liminf_{n \to \infty} W^n(x) \geq \bar{E}_x \left\{ \int_{\mathcal{S}} R(\bar{q}(\xi, \cdot)\| p(\xi, \cdot)) \bar{L}(d\xi) + h(\bar{L}) \right\}. \tag{8.20}
\]

Since w.p.1 \( \bar{L} \) is an invariant measure of \( \bar{q}(x, dy) \) [Theorem 8.2.8 (c)], this lower limit leads naturally to the following definition of a function \( I(\mu) \) for which the Laplace principle upper bound is valid. We denote by \( \mathcal{T} \) be the set of all transition probability functions on \( \mathcal{S} \). For any \( \mu \in \mathcal{P}(\mathcal{S}) \) and \( q(x, dy) \in \mathcal{T} \), we define the probability measure \( \mu q \) on \( \mathcal{S} \) by

\[
\mu q(A) \doteq \int_{\mathcal{S}} q(x, A) \mu(dx)
\]

for Borel sets \( A \). Thus \( \mu \) is an invariant measure of \( q(x, dy) \) if and only if \( \mu q = q \). If for \( \mu \in \mathcal{P}(\mathcal{S}) \) we define

\[
I(\mu) \doteq \inf_{q \in \mathcal{T} : \mu q = \mu} \int_{\mathcal{S}} R(q(x, \cdot)\| p(x, \cdot)) \mu(dx), \tag{8.21}
\]

then (8.20) implies that w.p.1

\[
\int_{\mathcal{S}} R(\bar{q}(\xi, \cdot)\| p(\xi, \cdot)) \bar{L}(d\xi) \geq I(\bar{L}).
\]
We have shown that every convergent subsequence of the original sequence \( \{W^n(x), n \in \mathbb{N}\} \) has a subsubsequence satisfying the desired lower limit

\[
\liminf_{n \to \infty} W^n(x) \geq \mathbb{E}_x \{I(\bar{L}) + h(\bar{L})\} \geq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}. \tag{8.22}
\]

An argument by contradiction applied to an arbitrary convergent subsequence establishes this lower limit for the entire sequence \( \{W^n(x), n \in \mathbb{N}\} \). This completes the proof of the Laplace principle upper bound

\[
\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{E}_x \{\exp[-n h(L^n)]\} \leq - \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}. \tag{8.23}
\]

We claim that this Laplace principle upper bound actually holds uniformly over \( x \) in any compact subset \( K \) of \( \mathcal{S} \). In other words,

\[
\limsup_{n \to \infty} \sup_{x \in K} \frac{1}{n} \log \mathbb{E}_x \{\exp[-n h(L^n)]\} \leq - \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}. \tag{8.24}
\]

If this upper bound were not true, then there would exist \( \delta > 0 \) and a subsequence \( \{x^n\} \) in \( K \) such that for all sufficiently large \( n \)

\[
W^n(x^n) < \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\} - \delta.
\]

Using the compactness of \( K \) and applying Theorem 8.2.8, we conclude the existence of a subsubsequence of \( \{(\nu^n, \bar{L}^n, x^n)\} \) such that \( (\nu^n, \bar{L}^n) \xrightarrow{P} (\nu, \bar{L}) \) and \( x^n \to x \in K \). As in the proofs of the lower limits (8.19) and (8.22), for any \( \varepsilon > 0 \)

\[
\liminf_{n \to \infty} W^n(x^n) + \varepsilon = \liminf_{n \to \infty} V^n(x^n) + \varepsilon
\]

\[
\geq \mathbb{E}_x \left\{ \int_{\mathcal{S}} \mathbb{R}(\bar{q}(\xi, \cdot) \| p(\xi, \cdot)) \bar{L}(d\xi) + h(\bar{L}) \right\}
\]

\[
\geq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}.
\]

Sending \( \varepsilon \to 0 \) gives a contradiction to (8.24). This yields the uniform Laplace principle upper bound (8.23).

The following proposition has been proved.

**Proposition 8.3.3.** We assume Conditions 8.2.2 and 8.3.1 and define \( I \) in (8.21). Then for each compact subset \( K \) of \( \mathcal{S} \) and any bounded continuous function \( h \) mapping \( \mathcal{P}(\mathcal{S}) \) into \( \mathbb{R} \), we have the uniform Laplace principle upper bound

\[
\limsup_{n \to \infty} \sup_{x \in K} \frac{1}{n} \log \mathbb{E}_x \{\exp[-n h(L^n)]\} \leq - \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}.
\]

We end this section by showing that a transition probability function that satisfies Condition 8.2.2 and the Feller property (Condition 8.3.1) has an invariant measure. This result can also be derived independently of the present approach under weaker conditions.
8.4. STATEMENT OF THE LAPLACE PRINCIPLE

Proposition 8.3.4. Under Condition 8.2.2 and the Feller property (Condition 8.3.1), $p(x, dy)$ has an invariant measure. In particular, if $S$ is compact and $p(x, dy)$ satisfies the Feller property, then $p(x, dy)$ has an invariant measure.

Proof. If $S$ is compact, then Condition 8.2.2 is satisfied by the constant function $U = 0$. Hence the second statement in the proposition is a consequence of the first statement, which we now prove. In Theorem 8.2.8, for each $n \in \mathbb{N}$, $j \in \{0, 1, \ldots, n - 1\}$, $x \in S$, and $\mu \in \mathcal{M}_{j/n}(S)$, we choose as the admissible sequence of controls $\nu^n_j(dy|x, \mu) \equiv p(x, dy)$. Then

$$
\nu^n(dx \times dy) = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{X_j^n}(dx) \times p(X_j^n, dy) = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{X_j^n}(dx) \otimes p(x, dy) = \bar{L}^n(dx) \otimes p(x, dy),
$$

and since with $\nu^n_j(dy|x, \mu) \equiv p(x, dy)$ the bound (8.15) is trivially satisfied, Theorem 8.2.8 can be applied. There exists a subsequence of $\{(\nu^n, \bar{L}^n), n \in \mathbb{N}\}$ which converges in distribution to a pair of stochastic kernels $(\nu, \bar{L})$, where w.p.1 $\bar{L}$ is both the first marginal and the second marginal of $\nu$. Since by the Skorohod Representation Theorem we can assume that the convergence $(\nu^n, \bar{L}^n) \to (\nu, \bar{L})$ occurs w.p.1, Lemma 8.3.2 applied to the previous display implies that w.p.1 $\nu(dx \times dy) = \bar{L}(dx) \otimes p(x, dy)$. With probability 1, since for all Borel sets $B$

$$
\bar{L}(B) = \nu_2(B) = \int_S p(x, B) \bar{L}(dx),
$$

$\bar{L}$ is an invariant measure of $p(x, dy)$. $\blacksquare$

In the next section, we introduce two other conditions and then state the Laplace principle for the empirical measures $\{L^n\}$.

8.4 Statement of the Laplace Principle

The Laplace principle to be stated in Theorem 8.4.3 holds under a set of conditions that includes the Feller property on the underlying transition probability function $p(x, dy)$. Before presenting the theorem, we introduce conditions on $p(x, dy)$ that are needed to prove the Laplace principle lower bound. The proof of the lower bound does not require that $p(x, dy)$ satisfy the Feller property.

Given $q(x, dy)$ a transition probability function on $S$ and $k \in \mathbb{N}$, we write $q^{(1)}(x, dy) \equiv q(x, dy)$ and denote by $q^{(k)} = q^{(k)}(x, dy)$ the $k$-step transition probability function defined recursively by

$$
q^{(k+1)}(x, A) = \int_S q(y, A) q^{(k)}(x, dy)
$$

for Borel sets $A$. In the following condition, we impose a transitivity assumption on $p(x, dy)$ as well as the requirement that $p(x, dy)$ have an invariant measure. The transitivity assumption, weaker than Hypothesis H in [34], is a slight variation of Hypothesis (gH) in [18].
Condition 8.4.1.  
(a) There exist positive integers \( \ell_0 \) and \( n_0 \) such that for all \( x \) and \( \zeta \) in \( S \)
\[
\sum_{i=0}^{\infty} \frac{1}{2^i} p^{(i)}(x, dy) \leq \sum_{j=n_0}^{\infty} \frac{1}{2^j} p^{(j)}(\zeta, dy).
\]

In other words, if for \( i \in \mathbb{N} \) satisfying \( i \geq \ell_0 \), \( x \in S \), and a Borel set \( E \) \( p^{(i)}(x, E) > 0 \), then for each \( \zeta \in S \) there exists \( j \geq n_0 \) such that \( p^{(j)}(\zeta, E) > 0 \).

(b) \( p(x, dy) \) has an invariant measure.

If \( S \) is compact and \( p(x, dy) \) satisfies the Feller property, then as we pointed out in
Proposition 8.3.4, part (b) of this condition is automatically valid. If \( S \) is not compact,
then according to the same proposition the existence of an invariant measure of \( p(x, dy) \)
is guaranteed by Condition 8.2.2 and the Feller property. Another construction of an
invariant measure is given, for example, in Exercise 4.1.48 of [29].

The next assumption will allow us to prove uniformity of the Laplace principle lower
bound over initial conditions \( X_0 = x \) in any compact subset of \( S \). This assumption is a
locally uniform lower bound on the transition probability function.

Condition 8.4.2. Let \( d(\cdot, \cdot) \) denote the metric on \( S \). For any \( x \in S \) there exists a
probability measure \( \beta \) on \( S \), \( k \in \mathbb{N} \), and \( c > 0 \) such that for all \( \zeta \in S \) satisfying \( d(\zeta, x) < c \)
and for all Borel sets \( A \)
\[
\sum_{i=1}^{k} p^{(i)}(\zeta, A) \geq c \beta(A).
\]

We now state the Laplace principle for the empirical measures.

Theorem 8.4.3. Let \( \{X_j, j \in \mathbb{N}_0\} \) be a Markov chain that takes values in the Polish
space \( S \) and has stationary transition probabilities. We denote by \( p(x, dy) \) the transition
probability function of the Markov chain. For \( \mu \in \mathcal{P}(S) \), we define
\[
I(\mu) \doteq \inf_{\{q \in T; \mu_q = \mu\}} \int_S R(q(x, \cdot)\|p(x, \cdot)) \mu(dx),
\]
where \( T \) denotes the set of all transition probability functions on \( S \). For each bounded
continuous function \( h \) mapping \( \mathcal{P}(S) \) into \( \mathbb{R} \), the following conclusions hold.

(a) Under Conditions 8.2.2 and 8.3.1, \( I \) has compact level sets.

(b) Under Conditions 8.2.2 and 8.3.1, for each compact subset \( K \) of \( S \) we have the
uniform Laplace principle upper bound
\[
\limsup_{n \to \infty} \sup_{x \in K} \frac{1}{n} \log E_x \{\exp[-n h(L^n)]\} \leq - \inf_{\mu \in \mathcal{P}(S)} \{I(\mu) + h(\mu)\}.
\]

(c) Under Condition 8.4.1, for each initial condition \( X_0 = x \in S \) we have the Laplace
principle lower bound
\[
\liminf_{n \to \infty} \frac{1}{n} \log E_x \{\exp[-n h(L^n)]\} \geq - \inf_{\mu \in \mathcal{P}(S)} \{I(\mu) + h(\mu)\}.
\]
8.5. Properties of the Rate Function

Under Conditions 8.4.1 and 8.4.2, for each compact subset $K$ of $\mathcal{S}$ we have the uniform Laplace principle lower bound

$$\liminf_{n \to \infty} \inf_{x \in K} \frac{1}{n} \log E_x \{ \exp[-n h(L^n)] \} \geq - \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{ I(\mu) + h(\mu) \}.$$  

(d) In summary, under Conditions 8.2.2, 8.3.1, and 8.4.1, for any initial condition $X_0 = x \in \mathcal{S}$ the sequence $\{L^n\}$ satisfies the Laplace principle on $\mathcal{P}(\mathcal{S})$ with rate function $I$. Under Conditions 8.2.2, 8.3.1, 8.4.1, and 8.4.2, the sequence $\{L^n\}$ satisfies the Laplace principle on $\mathcal{P}(\mathcal{S})$ with rate function $I$ uniformly on compacts.

The form of the rate function differs from the standard Donsker–Varadhan rate function given in [32, 34]. By a purely analytic argument that makes no hypotheses on $p(x, dy)$, one can prove that the two rate functions agree. In the context of the full Laplace principle in part (d), part (b) of Condition 8.4.1, which requires the existence of an invariant measure of $p(x, dy)$, is automatically valid under the other conditions that we assume (see Proposition 8.3.4).

If $\mathcal{S}$ is compact, then there are a number of simplifications in the theorem. Under Condition 8.3.1, $I(\mu)$ is a lower semicontinuous function of $\mu \in \mathcal{P}(\mathcal{S})$ [Proposition 8.5.2 (a)], and so we can conclude, without Condition 8.2.2, that $I$ has compact level sets. Using Remark 8.2.6, we can also prove part (b) of the theorem without Condition 8.2.2.

We next note that the initial condition $X_0 = x$ appearing in both the Laplace principle upper bound and lower bound can be replaced by an initial distribution $\mu$. Furthermore, these bounds hold uniformly over certain classes of initial distributions $\mu$. First, let $\Phi$ be any family of measures in $\mathcal{P}(\mathcal{S})$ for which there exists a compact subset $K$ of $\mathcal{S}$ such that $\inf_{\mu \in \Phi} \mu(K) > 0$. If, for example, $\Phi$ is relatively compact, then this condition holds by Prohorov’s Theorem. For any such family $\Phi$ the Laplace principle lower bound holds uniformly for $\mu \in \Phi$. Second, let $\Phi$ be any family of measures in $\mathcal{P}(\mathcal{S})$ satisfying

$$\sup_{\mu \in \Phi} \int_{\mathcal{S}} U \, d\mu < \infty,$$

where $U$ is the function appearing in Condition 8.2.2. It is easy to check that $U$ is a tightness function, and thus any such family $\Phi$ is relatively compact. The Laplace principle upper bound and lower bound both hold uniformly for $\mu$ in any such family $\Phi$. The proofs of these uniform bounds involve straightforward modifications of the proofs of the Laplace principle bounds in Theorem 8.4.3 and will be omitted.

The Laplace principle upper bound stated in part (b) of Theorem 8.4.3 has already been proved in the previous section. In the next section, we establish part (a) of the theorem together with other properties of $I$. The lower bounds stated in part (c) will be proved in Section 8.6.

8.5 Properties of the Rate Function

The main result of the present section is that the function $I$ defined in equation (8.25) has compact level sets. This is carried out in Proposition 8.5.2, which also establishes that $I$ is convex.
In preparation for the proof, we present a useful construction relating probability measures on \( \mathcal{S} \times \mathcal{S} \) with transition probability functions on \( \mathcal{S} \). We recall that for any probability measure \( \theta \) on \( \mathcal{S} \) and any transition probability function \( q(x, dy) \) on \( \mathcal{S} \), \( \theta \otimes q \) denotes the probability measure on \( \mathcal{S} \times \mathcal{S} \) given by

\[
(\theta \otimes q)(A \times B) = \int_{A \times B} \theta(dx) q(x, dy) = \int_A q(x, B) \theta(dx)
\]

for Borel sets \( A \) and \( B \). Now let \( \tau \) be a probability measure on \( \mathcal{S} \times \mathcal{S} \) with marginals \( \tau_1 \) and \( \tau_2 \). According to Corollary A.5.5, there exists a transition probability function \( q_\tau(x, dy) \) on \( \mathcal{S} \) such that

\[
\tau(A \times B) = \int_{A \times B} \tau_1(dx) q_\tau(x, dy) = \int_A q_\tau(x, B) \tau_1(dx)
\]

for all Borel sets \( A \) and \( B \). We write \( \tau = \tau_1 \otimes q_\tau \). Now suppose that both marginals \( \tau_1 \) and \( \tau_2 \) of \( \tau \) equal a given probability measure \( \mu \). Then \( q_\tau(x, dy) \) has \( \mu \) as an invariant measure. Indeed, for any Borel set \( B \)

\[
\mu(B) = \tau_2(B) = \tau(\mathcal{S} \times B) = \int_\mathcal{S} q_\tau(x, B) \tau_1(dx) = \int_\mathcal{S} q_\tau(x, B) \mu(dx) = \mu q_\tau(B).
\]

We summarize this discussion in part (a) of the next lemma. Part (b) provides an alternate representation for \( I \) which will be used in conjunction with part (c) to establish properties of this function. The three parts of the lemma hold without any conditions on \( p(x, dy) \).

**Lemma 8.5.1.** For \( \mu \in \mathcal{P}(\mathcal{S}) \) we define \( I(\mu) \) by (8.25) and

\[
J(\mu) \doteq \inf \{ R(\tau \| \tau_1 \otimes p) : \tau \in \mathcal{P}(\mathcal{S} \times \mathcal{S}), \tau_1 = \tau_2 = \mu \}.
\]

The following conclusions hold.

(a) If \( \tau \in \mathcal{P}(\mathcal{S} \times \mathcal{S}) \) satisfies \( \tau_1 = \mu \), then there exists a transition probability function \( q_\tau(x, dy) \) on \( \mathcal{S} \) such that \( \tau = \mu \otimes q_\tau \). If \( \tau \in \mathcal{P}(\mathcal{S} \times \mathcal{S}) \) satisfies \( \tau_1 = \tau_2 = \mu \), then in addition \( \mu \) is an invariant measure of \( q_\tau(x, dy) \).

(b) For all \( \mu \in \mathcal{P}(\mathcal{S}) \), \( I(\mu) = J(\mu) \).

(c) For all \( \mu \in \mathcal{P}(\mathcal{S}) \) the infima in the definitions of \( I(\mu) \) and \( J(\mu) \) are attained. Thus there exists a transition probability function \( q(x, dy) \) on \( \mathcal{S} \) with invariant measure \( \mu \) and a probability measure \( \tau \) on \( \mathcal{S} \times \mathcal{S} \) with both marginals equal to \( \mu \) such that

\[
I(\mu) = \int_\mathcal{S} R(q(x, \cdot) \| p(x, \cdot)) \mu(dx) = R(\tau \| \mu \otimes p).
\]

**Proof.** (a) As we discussed prior to the statement of the lemma, this follows from Corollary A.5.5.

(b) Part (f) of Lemma 1.4.3 implies that for any transition probability function \( q(x, dy) \)

\[
\int_\mathcal{S} R(q(x, \cdot) \| p(x, \cdot)) \mu(dx) = R(\mu \otimes q \| \mu \otimes p).
\]
8.5. PROPERTIES OF THE RATE FUNCTION

It follows that

\[ I(\mu) = \inf_{\{q \in \mathcal{P} \mid \mu \equiv \mu \}} R(\mu \otimes q \| \mu \otimes p). \]

A probability measure \( \tau \) on \( S \times S \) satisfies \( \tau_1 = \tau_2 = \mu \) if and only if \( \tau = \mu \otimes q \) for some transition probability function \( q(x, dy) \) on \( S \) having \( \mu \) as an invariant measure. Hence

\[ I(\mu) = \inf_{\{q \in \mathcal{P} \mid \mu \equiv \mu \}} R(\mu \otimes q \| \mu \otimes p) = \inf \{ R(\tau \| \tau_1 \otimes p) : \tau \in \mathcal{P}(S \times S), \tau_1 = \tau_2 = \mu \} = J(\mu). \]

(c) For \( \mu \in \mathcal{P}(S) \) we define

\[ A(\mu) \doteq \{ \tau \in \mathcal{P}(S \times S) : \tau_1 = \tau_2 = \mu \}; \]

\( A(\mu) \) is nonempty since it contains \( \mu \times \mu \). The tightness of \( \mu \) guarantees that for each \( \varepsilon > 0 \) there exists a compact set \( K \) such that \( \mu(K) \geq 1 - \varepsilon \) [Theorem A.2.3]. \( K \times K \) is a compact subset of \( S \times S \), and for each \( \tau \in A(\mu) \) \( \tau(K \times K) \geq 1 - 2\varepsilon \). Since \( A(\mu) \) is closed, Prohorov’s Theorem implies that \( A(\mu) \) is compact. The equality \( I(\mu) = J(\mu) \), proved in part (b), can be rewritten in terms of \( A(\mu) \) as

\[ I(\mu) = J(\mu) = \inf_{\tau \in A(\mu)} R(\tau \| \mu \otimes p). \]

Since the function mapping \( \tau \in \mathcal{P}(S \times S) \longmapsto R(\tau \| \mu \otimes p) \) is lower semicontinuous [Lemma 1.4.3 (b)], it attains its infimum over the compact set \( A(\mu) \). Hence there exists \( \tau \in A(\mu) \) such that \( I(\mu) = R(\tau \| \mu \otimes p) \). According to part (a) of the present lemma and part (f) of Lemma 1.4.3, there exists a transition probability function \( q(x, dy) \) with invariant measure \( \mu \) such that

\[ I(\mu) = R(\tau \| \mu \otimes p) = R(\mu \otimes q \| \mu \otimes p) = \int_S R(q(x, \cdot) \| p(x, \cdot)) \mu(dx). \]

This completes the proof of part (c). \( \blacksquare \)

The next proposition states a number of properties of \( I \).

**Proposition 8.5.2.** The function \( I \) defined in (8.25) has the following properties.

(a) Without any conditions on \( p(x, dy) \), \( I(\mu) \) is a convex function of \( \mu \in \mathcal{P}(S) \). Under Condition 8.3.1 on \( p(x, dy) \), \( I(\mu) \) is a lower semicontinuous function of \( \mu \in \mathcal{P}(S) \).

(b) Under Condition 8.2.2, for each \( M < \infty \) the level set \( \{ \mu \in \mathcal{P}(S) : I(\mu) \leq M \} \) is a relatively compact subset of \( \mathcal{P}(S) \). If in addition Condition 8.3.1 holds, then the level set is compact.

**Proof.** (a) For \( \mu \in \mathcal{P}(S) \) we define

\[ A(\mu) \doteq \{ \tau \in \mathcal{P}(S \times S) : \tau_1 = \tau_2 = \mu \}. \]

We first prove that \( I \) is convex. For \( i = 1, 2 \), given \( \mu^i \in \mathcal{P}(S) \) there exists \( \tau^i \in A(\mu^i) \) such that \( I(\mu^i) = R(\tau^i \| \mu^i \otimes p) \) [Lemma 8.5.1 (c)]. Then for any \( \lambda \in (0, 1) \)

\[ (\lambda \tau^1 + (1 - \lambda) \tau^2) = (\lambda \tau^1 + (1 - \lambda) \tau^2) \in A(\mu) \]

so that

\[ (\lambda \tau^1 + (1 - \lambda) \tau^2) = \lambda \mu^1 + (1 - \lambda) \mu^2, \]
and so

\[ I(\lambda \mu^1 + (1 - \lambda)\mu^2) \leq R(\lambda \tau^1 + (1 - \lambda)\tau^2 \| \lambda \mu^1 \otimes p + (1 - \lambda)\mu^2 \otimes p). \]

Since \( R(\cdot \| \cdot) \) is convex on \( \mathcal{P}(\mathcal{S} \times \mathcal{S}) \times \mathcal{P}(\mathcal{S} \times \mathcal{S}) \), for any \( \lambda \in (0, 1) \)

\[ I(\lambda \mu^1 + (1 - \lambda)\mu^2) \leq \lambda R(\tau^1 \| \mu^1 \otimes p) + (1 - \lambda)R(\tau^2 \| \mu^2 \otimes p) \]
\[ = \lambda I(\mu^1) + (1 - \lambda)I(\mu^2). \]

This proves the convexity of \( I \).

We now prove that \( I \) is lower semicontinuous. Let \( \{\mu_n, n \in \mathbb{N}\} \) be any sequence in \( \mathcal{P}(\mathcal{S}) \) converging weakly to \( \mu \). For each \( n \) there exists \( \tau_n \in A(\mu_n) \) such that \( I(\mu_n) = R(\tau_n \| \mu_n \otimes p) \) [Lemma 8.5.1 (c)]. It suffices to prove that any subsequence of \( \{\mu_n\} \) has a subsubsequence such that
\[
\liminf_{n \to \infty} I(\mu_n) \geq I(\mu).
\]

Since for each \( n \in \mathbb{N} \) the marginals \( (\tau_n)_1 \) and \( (\tau_n)_2 \) equal \( \mu_n \) and since \( \mu_n \Rightarrow \mu \), any subsequence of \( \{\tau_n\} \) is tight. Therefore, there exists a subsubsequence such that \( \tau_n \Rightarrow \tau \in A(\mu) \). In addition, \( \mu_n \otimes p \Rightarrow \mu \otimes p \) [Lemma 8.3.2]. The lower semicontinuity of \( R(\cdot \| \cdot) \) now yields
\[
\liminf_{n \to \infty} I(\mu_n) = \liminf_{n \to \infty} R(\tau_n \| \mu_n \otimes p) \geq R(\tau \| \mu \otimes p) \geq I(\mu).
\]

This proves the lower semicontinuity of \( I \).

(b) For any \( \mu \in \mathcal{P}(\mathcal{S}) \) for which \( I(\mu) < \infty \), we use part (c) of Lemma 8.5.1 to choose a transition probability function \( q(x, dy) \) on \( \mathcal{S} \) that has \( \mu \) as an invariant measure and satisfies
\[
\int_{\mathcal{S}} R(q(x, \cdot) \| p(x, \cdot)) \mu(dx) = I(\mu).
\]

Since \( \mu \) is an invariant measure for \( q \), it follows that for any bounded measurable function \( g \) mapping \( \mathcal{S} \) into \( \mathbb{R} \)
\[
\int_{\mathcal{S}} g(x) \mu(dx) = \int_{\mathcal{S} \times \mathcal{S}} g(y) q(x, dy) \mu(dx).
\]

By taking suitable limits on \( g \), we obtain this equality for the nonnegative function \( U \) appearing in Condition 8.2.2; thus
\[
\int_{\mathcal{S}} U(x) \mu(dx) = \int_{\mathcal{S} \times \mathcal{S}} U(y) q(x, dy) \mu(dx).
\]

We also recall (8.11), which for any \( x \in \mathcal{S} \) yields for the measure \( \gamma(\cdot) \triangleq q(x, \cdot) \)
\[
\int_{\mathcal{S}} U(y) q(x, dy) - \log \int_{\mathcal{S}} e^{U(y)} p(x, dy) \leq R(q(x, \cdot) \| p(x, \cdot)).
\]

Defining
\[ c(x) \triangleq U(x) - \int_{\mathcal{S}} e^{U(y)} p(x, dy), \]
we deduce that
\[
\int_S c(x) \mu(dx) = \int_S U(x) \mu(dx) - \int_S \left( \log \int_S e^{U(y)} p(x,dy) \right) \mu(dx) \\
= \int_{S \times S} U(y) q(x,dy) \mu(dx) - \int_S \left( \log \int_S e^{U(y)} p(x,dy) \right) \mu(dx) \\
\leq \int_S R(q(x,\cdot)\|p(x,\cdot)) \mu(dx) \\
= I(\mu).
\]

Thus for each \( M < \infty \) and for all \( \mu \in \mathcal{P}(S) \) satisfying \( I(\mu) \leq M \)
\[
\int_S c \, d\mu \leq I(\mu) \leq M.
\]

Part (b) of Lemma 8.2.4 and Prohorov’s Theorem imply that the level set \( \{ \mu \in \mathcal{P}(S) : I(\mu) \leq M \} \) is relatively compact. Since Condition 8.3.1 and part (a) of the present proposition imply that the level set is closed, Conditions 8.2.2 and 8.3.1 together imply that the level set is compact. This completes the proof of part (b).

It is worthwhile to note that the lower semicontinuity and compactness properties of \( I \) just proved are deterministic analogues of properties established in the proof of the Laplace principle upper bound. In the next section, we prove the Laplace principle lower bounds for the empirical measures as stated in Theorem 8.4.3.

### 8.6 Proofs of the Laplace Principle Lower Bounds

The purpose of this section is to prove the following proposition, in which we verify two Laplace principle lower bounds. This proposition gives part (c) of the Laplace principle stated in Theorem 8.4.3.

**Proposition 8.6.1.** We define \( I \) in (8.25). For each bounded continuous function \( h \) mapping \( \mathcal{P}(S) \) into \( \mathbb{R} \) the following conclusions hold.

(a) Under Condition 8.4.1, for each initial condition \( X_0 = x \in S \) we have the Laplace principle lower bound
\[
\liminf_{n \to \infty} \frac{1}{n} \log E_x \{ \exp[-n \, h(L^n)] \} \geq - \inf_{\mu \in \mathcal{P}(S)} \{ I(\mu) + h(\mu) \}.
\]

(b) Under Conditions 8.4.1 and 8.4.2, for each compact subset \( K \) of \( S \) we have the uniform Laplace principle lower bound
\[
\liminf_{n \to \infty} \inf_{x \in K} \frac{1}{n} \log E_x \{ \exp[-n \, h(L^n)] \} \geq - \inf_{\mu \in \mathcal{P}(S)} \{ I(\mu) + h(\mu) \}.
\]
The difficulties that we will encounter in the proof of this proposition are reminiscent of those encountered in the proofs of the Laplace principle lower bounds for the random walk models of Chapters 6 and 7. In order to prove a Laplace principle lower bound here, we must prove a corresponding upper limit for the minimal cost functions \( V^n(x) \). For a given \( \varepsilon > 0 \) there exists a measure \( \gamma \in \mathcal{P}(\mathcal{S}) \) satisfying

\[
I(\gamma) + h(\gamma) \leq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{ I(\mu) + h(\mu) \} + \varepsilon < \infty.
\]

We then choose a transition probability function \( q(x, dy) \) on \( \mathcal{S} \) such that \( q(x, dy) \) is minimizing in the definition of \( I(\gamma) \) [Lemma 8.5.1 (c)]; i.e.,

\[
I(\gamma) = \int_{\mathcal{S}} R(q(x, \cdot) \| p(x, \cdot)) \gamma(dx).
\]  

(8.26)

A basic question is how to choose an appropriate admissible control sequence \( \{\nu^n_j\} \) that will be used in proving the upper limit for \( V^n(x) \). Given a transition probability function \( q(x, dy) \) satisfying (8.26), a natural choice is \( \nu^n_j(dy|x, \mu) = q(x, dy) \). Use of this control sequence is equivalent to having the controlled process \( \{X^n_j\} \) evolve as a Markov chain with transition probability function \( q(x, dy) \). An essential step in the proof is to determine the limit of the controlled empirical measures \( \{\hat{L}^n\} \) as \( n \to \infty \). This is straightforward when the Markov chain associated with \( \gamma \) and \( q(x, dy) \) is ergodic. However, as we point out in the next paragraph, the ergodicity will not be true in general.

Formula (8.26) and the finiteness of \( I(\gamma) \) imply that \( \gamma \)-a.s. for \( x \in \mathcal{S}, q(x, \cdot) \) is absolutely continuous with respect to \( p(x, \cdot) \). Thus in a certain sense \( q \) cannot be too large compared to \( p \). However, there is no a priori restriction on \( q \) from below. In fact the density of \( q(x, \cdot) \) with respect to \( p(x, \cdot) \) may equal 0 on a set of positive \( p(x, \cdot) \)-measure for a set of \( x \) of positive \( \gamma \)-measure. In this case the Markov chain associated with \( \gamma \) and \( q(x, dy) \) is not necessarily ergodic.

In order to guarantee the ergodicity, we impose the strong hypotheses in Condition 8.4.1. Under this condition \( p(x, dy) \) has an invariant measure and the associated Markov chain is ergodic [Lemma 8.6.2 (a)]. We will prove under the same condition that if \( I(\gamma) < \infty \) and \( dq(x, \cdot)/dp(x, \cdot) > 0 \) \( \gamma \)-a.s., then the Markov chain associated with \( \gamma \) and \( q(x, dy) \) is also ergodic [Lemma 8.6.3 (c)]. Although the latter inequality is not available for an arbitrary \( \gamma \) satisfying \( I(\gamma) < \infty \), an approximation procedure analogous to that used in Chapters 6 and 7 is applicable. More precisely, for any \( \gamma \in \mathcal{P}(\mathcal{S}) \) satisfying \( I(\gamma) < \infty \) we will be able to find a measure \( \nu^* \) with the following properties: \( \nu^* \) is close to \( \gamma \), there exists a transition probability function \( q^*(x, dy) \) such that \( q^*(x, dy) \) has \( \nu^* \) as an invariant measure, \( dq^*(x, \cdot)/dp(x, \cdot) > 0 \) \( \nu^* \)-a.s., and

\[
I(\nu^*) \leq \int_{\mathcal{S}} R(q^*(x, \cdot) \| p(x, \cdot)) \nu^*(dx) \leq I(\gamma).
\]

The positivity of \( dq^*(x, \cdot)/dp(x, \cdot) \) will imply that the Markov chain associated with \( \nu^* \) and \( q^*(x, dy) \) is ergodic. In turn, these properties will enable us to prove the required upper limit for \( V^n(x) \).
As we have just discussed, we will prove the Laplace principle lower bounds by approximating $\gamma$ by an element of a special class $\mathcal{N}$ of measures. We next identify this class.

**Identification of the class $\mathcal{N}$.** $\mathcal{N}$ consists of all measures $\gamma \in \mathcal{P}(\mathcal{S})$ satisfying the following two conditions:

(a) $I(\gamma) < \infty$.

(b) There exists a transition probability function $q(x, dy)$ on $\mathcal{S}$ such that $\gamma$ is an invariant measure of $q(x, dy)$ and the Markov chain with initial distribution $\gamma$ and transition probability function $q(x, dy)$ is ergodic.

Condition 8.4.1 will allow us to approximate, by an element of $\mathcal{N}$, an arbitrary $\mu \in \mathcal{P}(\mathcal{S})$ satisfying $I(\mu) < \infty$ and eventually to prove the Laplace principle lower bound. The next lemma gives useful information that will enable us to carry this out.

**Lemma 8.6.2.** We assume Condition 8.4.1 and define $I$ in (8.25). The following conclusions hold.

(a) The transition probability function $p(x, dy)$ has a unique invariant measure $\mu^*$. The Markov chain having $\mu^*$ as its initial distribution and $p(x, dy)$ as its transition probability function is ergodic.

(b) Let $A$ be a Borel set such that $p^{(0)}(x_0, A) > 0$ for some $x_0 \in \mathcal{S}$. Then $\mu^*(A) > 0$.

(c) For $\mu \in \mathcal{P}(\mathcal{S})$, $I(\mu) < \infty$ implies $\mu \ll \mu^*$.

Before proving this lemma, let us note that part (c) will play a key role in the proof of Lemma 8.6.3, which is the heart of the proof of the Laplace principle lower bounds. Thus it seems likely that any alternative condition besides Condition 8.4.1 which yields part (c) would also be useful for proving the Laplace principle lower bounds. As an indication of the importance of the property stated in part (c), we note that when it does not hold, part (c) of Theorem 9.3.2 is not valid. The latter is a Laplace lower bound with respect to the $\tau$–topology; a topology often considered for this problem.

**Proof of Lemma 8.6.2.** (a) Part (b) of Condition 8.4.1 states that $p(x, dy)$ has an invariant measure, which we label $\mu^*$. We prove that $p(x, dy)$ is indecomposable; i.e., that there do not exist disjoint Borel sets $A_1$ and $A_2$ such that

$$p(x, A_1) = 1 \text{ for all } x \in A_1 \text{ and } p(y, A_2) = 1 \text{ for all } y \in A_2.$$  \hfill (8.27)

Theorem A.4.5 then implies that $\mu^*$ is unique and that the Markov chain is ergodic, completing the proof of part (a).

Let us suppose that there exist disjoint Borel sets $A_1$ and $A_2$ such that formula (8.27) holds. Then for any $\ell$ and $n$ in $\mathcal{N}$ we have $p^{(\ell)}(x, A_1) = 1$ for all $x \in A_1$ and $p^{(n)}(y, A_2) = 1$ for all $y \in A_2$. According to part (a) of Condition 8.4.1, the first of these equalities with $\ell = \ell_0$ guarantees that for each $y \in A_2$ there exists $j \geq n_0$ such that $p^{(j)}(y, A_1) > 0$. 

Since \( A_1 \) and \( A_2 \) are disjoint, for \( y \in A_2 \) this inequality is incompatible with the equality
\[ p^{(n)}(y, A_2) = 1 \] for any \( n \in \mathbb{N} \). This contradiction shows that \( p(x, dy) \) is indecomposable.

(b) We define a function \( \iota \) mapping \( \mathcal{S} \) into \( \mathbb{N} \) by
\[ \iota(\zeta) = \min\{j \in \mathbb{N} : j \geq n_0 \text{ and } p^{(j)}(\zeta, A) > 0\} . \]

Part (a) of Condition 8.4.1 guarantees that \( \iota(\zeta) \) is well defined. Furthermore, \( \iota(\zeta) \) is measurable. \( \mathcal{S} \) can be written as the disjoint union of Borel sets \( \bigcup_{j=n_0}^{\infty} \Sigma^{(j)} \), where
\[ \Sigma^{(j)} = \{\zeta \in \mathcal{S} : \iota(\zeta) = j\} . \]

Clearly, \( \mu^*(S^{(k)}) > 0 \) for some \( k \in \mathbb{N} \) satisfying \( k \geq n_0 \). In addition, \( p^{(k)}(\zeta, A) > 0 \) for all \( \zeta \in \Sigma^{(k)} \). Hence
\[ \mu^*(A) = \int_{\mathcal{S}} p^{(k)}(\zeta, A) \mu^*(d\zeta) \geq \int_{\Sigma^{(k)}} p^{(k)}(\zeta, A) \mu^*(d\zeta) > 0 . \]

(c) Let \( \mu \in \mathcal{P}(\mathcal{S}) \) satisfy \( I(\mu) < \infty \). By part (c) of Lemma 8.5.1 there exists a transition probability function \( q(x, dy) \) that has \( \mu \) as an invariant measure and satisfies
\[ \int_{\mathcal{S}} R(q(x, \cdot)\|p(x, \cdot)) \mu(dx) = I(\mu) < \infty . \]

The set \( \Delta = \{x \in \mathcal{S} : q(x, \cdot) \ll p(x, \cdot)\} \) satisfies \( \mu(\Delta) = 1 \). Since \( q(x, dy) \) has \( \mu \) as an invariant measure, the transition probability function
\[ \tilde{q}(x, \cdot) = \begin{cases} q(x, \cdot) & \text{if } x \in \Delta \\ p(x, \cdot) & \text{if } x \in \Delta^c \end{cases} \]
also has \( \mu \) as an invariant measure. We will prove that
\[ \tilde{q}^{(0)}(x, \cdot) \ll p^{(0)}(x, \cdot) \text{ for all } x \in \Delta . \]

Part (c) of the present lemma is an immediate consequence of this assertion. Indeed, suppose that \( \mu(A) > 0 \) for some Borel set \( A \). Iterating the equation \( \mu \tilde{q} = \mu \) yields
\[ \int_{\mathcal{S}} \tilde{q}^{(0)}(x, A) \mu(dx) = \mu(A) > 0 , \]
where \( \ell_0 \) is the number appearing in part (a) of Condition 8.4.1. Hence there exists a Borel set \( B \) such that \( \mu(B) > 0 \) and \( \tilde{q}^{(0)}(x, A) > 0 \) for all \( x \in B \). By formula (8.28), \( p^{(0)}(x_0, A) > 0 \) for some \( x_0 \in B \cap \Delta \). Part (b) implies that \( \mu^*(A) > 0 \). This shows that \( \mu \ll \mu^* \) and completes the proof of part (c).

We now prove (8.28). In fact, we will prove by induction that for each \( k \in \mathbb{N} \)
\[ \tilde{q}^{(k)}(x, \cdot) \ll p^{(k)}(x, \cdot) \text{ for all } x \in \Delta . \]

For \( k = 1 \) this assertion is immediate by the definition of \( \Delta \). For any \( j \in \mathbb{N} \), we now assume (8.29) for all \( k \in \{1, 2, \ldots, j\} \) and prove it for \( k = j + 1 \). For \( x \in \Delta \), let \( C \) be a Borel set such that
\[ p^{(j+1)}(x, C) = \int_{\mathcal{S}} p(y, C) p^{(j)}(x, dy) = 0 . \]
This implies that there exists a Borel set $\Gamma$ such that $p^{(j)}(x, \Gamma) = 1$ and $p(y, C) = 0$ for all $y \in \Gamma$. Since $x \in \Delta$, the inductive hypothesis implies that $\bar{q}^{(j)}(x, \Gamma) = 1$. It also implies that for $y \in \Gamma \cap \Delta$ $\bar{q}(y, C) = 0$. Hence, for any $x \in \Delta$

$$\bar{q}^{(j+1)}(x, C) = \int_S \bar{q}(y, C) \bar{q}^{(j)}(x, dy) = \int_{\Gamma} \bar{q}(y, C) \bar{q}^{(j)}(x, dy)$$

$$= \int_{\Gamma \cap \Delta} \bar{q}(y, C) \bar{q}^{(j)}(x, dy) + \int_{\Gamma \setminus \Delta} \bar{q}(y, C) \bar{q}^{(j)}(x, dy)$$

$$= \int_{\Gamma \cap \Delta} \bar{q}(y, C) \bar{q}^{(j)}(x, dy) + \int_{\Gamma \setminus \Delta} p(y, C) \bar{q}^{(j)}(x, dy) = 0.$$ 

This proves formula (8.29) for $k = j + 1$. The proof of the lemma is complete.

If $\ell_0 = 1$, then the proof of part (c) can be simplified. Indeed, in place of (8.28) we have $q(x, \cdot) \ll p(x, \cdot)$ for all $x \in \Delta$, and $q(x, dy)$ can be used in place of $\bar{q}(x, dy)$ in the paragraph following that formula. ■

The next lemma will play the same role in the proofs of the Laplace principle lower bounds for the empirical measures as the Approximation Procedure 6.4.3 played in the proofs of the Laplace principle lower bounds in Chapters 6 and 7.

**Lemma 8.6.3.** Assume Condition 8.4.1 and consider any $\gamma \in \mathcal{P}(S)$ satisfying $I(\gamma) < \infty$. Then given $\delta > 0$ there exists $\nu^* \in \mathcal{P}(S)$ with the following properties.

(a) $\|\nu^* - \gamma\|_v \leq \delta$.

(b) $\mu^* \ll \nu^*$ and $\nu^* \ll \mu^*$, where $\mu^*$ is the unique invariant measure of $p(x, dy)$ [Lemma 8.6.2 (a)].

(c) The measure $\nu^*$ belongs to $\mathcal{N}$; i.e., there exists a transition probability function $q^*(x, dy)$ on $S$ such that $\nu^*$ is an invariant measure of $q^*(x, dy)$ (in fact, the unique invariant measure) and the associated Markov chain is ergodic. In addition

$$I(\nu^*) \leq \int_S R(q^*(x, \cdot)||p(x, \cdot)) \nu^*(dx) \leq I(\gamma).$$  

(8.30)

**Proof.** (a) For $\kappa \in (0, 1)$, the probability measure $\gamma^\kappa \triangleq (1-\kappa)\gamma + \kappa \mu^*$ satisfies $\|\gamma^\kappa - \gamma\|_v = \kappa \|\mu^* - \gamma\|_v \leq 2\kappa$. Hence $\|\gamma^\kappa - \gamma\|_v \leq \delta$ for any $\kappa \in (0, \delta/2]$. In the statement of the lemma, we take $\nu^* \triangleq \gamma^\kappa$ for any $\kappa \in (0, \delta/2]$.

(b) Since $\kappa > 0$ and $\nu^*(A) \geq \kappa \mu^*(A)$ for all Borel sets $A$, it follows that $\mu^* \ll \nu^*$. Since $I(\gamma) < \infty$, part (c) of Lemma 8.6.2 implies that $\gamma \ll \mu^*$; thus $\nu^* \ll \mu^*$.

(c) Using part (c) of Lemma 8.5.1, we choose a transition probability function $q(x, dy)$ on $S$ that has $\gamma$ as an invariant measure and satisfies

$$\int_S R(q(x, \cdot)||p(x, \cdot)) \gamma(dx) = I(\gamma).$$

We then define probability measures $\lambda$, $\theta$, and $\lambda^*$ on $S \times S$ by

$$\lambda \triangleq \gamma \otimes q, \ \theta \triangleq \mu^* \otimes p, \ \text{and} \ \lambda^* \triangleq (1-\kappa)\lambda + \kappa \theta.$$
Since both of the marginals of $\lambda$ (resp., $\theta$) equal $\gamma$ (resp., $\mu^*$), both of the marginals of $\lambda^*$ equal $\nu^*$. Hence there exists a transition probability function $q^*(x, dy)$ on $\mathcal{S}$ such that $\lambda^* = \nu^* \otimes q^*$ and $\nu^*$ is an invariant measure of $q^*(x, dy)$ [Lemma 8.5.1 (a)].

We next verify formula (8.30), using part (f) of Lemma 1.4.3 and the fact that $R(\cdot \| \cdot)$ is convex and satisfies $R(\alpha \| \alpha) = 0$. We have

$$I(\nu^*) \leq \int_{\mathcal{S}} R(q^*(x, \cdot) \| p(x, \cdot)) \nu^*(dx) \leq R(\nu^* \otimes q^* \| \nu^* \otimes p) \leq R(\lambda^* \| \nu^* \otimes p) \leq (1 - \kappa) R(\gamma \otimes q \| \gamma \otimes p + \kappa \mu^* \otimes p) \leq (1 - \kappa) I(\gamma).$$

This proves formula (8.30).

In order to prove that $\nu^*$ is the unique invariant measure of $q^*(x, dy)$ and that the associated Markov chain is ergodic, we would like to apply part (a) of Lemma 8.6.2 with the pair $(\nu^*, q^*)$ replacing $(\mu^*, p)$. To do this, we must verify that Condition 8.4.1, assumed to hold for $p(x, dy)$, also holds for $q^*(x, dy)$. Since $\nu^*$ is an invariant measure of $q^*(x, dy)$, it suffices to show that the absolute continuity condition

$$\sum_{i=0}^{\infty} \frac{1}{2^i} p^{(i)}(x, dy) \ll \sum_{j=0}^{\infty} \frac{1}{2^j} p^{(j)}(\zeta, dy)$$

for all $x$ and $\zeta$ in $\mathcal{S}$ implies the absolute continuity condition

$$\sum_{i=0}^{\infty} \frac{1}{2^i} (q^*)^{(i)}(x, dy) \ll \sum_{j=0}^{\infty} \frac{1}{2^j} (q^*)^{(j)}(\zeta, dy)$$

for all $x$ and $\zeta$ in $\mathcal{S}$.

According to part (b) of the present lemma, we have $\nu^* \ll \mu^*$. Since $\nu^*(A) \geq \kappa \mu^*(A)$ for all Borel sets $A$, the Radon-Nikodym derivative $f(x) = \frac{d\nu^*}{d\mu^*}(x)$ satisfies $f(x) \in [\kappa, \infty)$ $\mu^*$-a.s. for $x \in \mathcal{S}$. For any Borel sets $A$ and $B$

$$\int_A q^*(x, B) f(x) \mu^*(dx) = \lambda^*(A \times B) \geq \kappa \theta(A \times B) = \kappa \int_A p(x, B) \mu^*(dx).$$

This implies that $\mu^*$-a.s. for $x \in \mathcal{S}$

$$q^*(x, B) \geq \frac{\kappa}{f(x)} p(x, B).$$

The formula

$$\int_{\mathcal{S}} R(q^*(x, \cdot) \| p(x, \cdot)) \nu^*(dx) \leq I(\gamma) < \infty,$$
proved in (8.31), implies that \( \nu^* \)-a.s. for \( x \in S \)
\[
q^*(x, \cdot) \ll p(x, \cdot).
\] (8.35)

Since \( \mu^* \) and \( \nu^* \) are mutually absolutely continuous (part (b)), there exists a Borel set \( C \) such that \( \mu^*(C) = 0 = \nu^*(C) \) and both (8.34) and (8.35) hold on the complement of \( C \). If we redefine \( q^*(x, dy) \) to equal \( p(x, dy) \) for \( x \in C \), then since \( \nu^*(C) = 0 \), \( \nu^* \)
remains an invariant measure of \( q^*(x, dy) \). In addition, (8.34) and (8.35) are then valid for all \( x \in S \), implying in turn that for each \( j \in \mathbb{N} \) and all \( x \in S \) the measures \( p^{(j)}(x, \cdot) \)
and \( q^{(j)}(x, \cdot) \) are mutually absolutely continuous. It follows that the absolute continuity condition (8.32) on \( p(x, dy) \) implies the absolute continuity condition (8.33) on \( q^*(x, dy) \).

Applying part (a) of Lemma 8.6.2 to the pair \((\nu^*, q^*)\) completes the proof of part (c) of the present lemma.

In order to prove the two Laplace principle lower bounds stated in Proposition 8.6.1, it suffices to consider only bounded, Lipschitz continuous functions \( h \). For the nonuniform lower bound in part (a), this is spelled out in Corollary 1.2.5. For the uniform lower bound in part (b), the fact that the rate function is independent of the initial point \( X_0 = x \) allows us easily to adapt the proof of that corollary to justify the restriction to bounded, Lipschitz continuous functions. While the restriction simplifies the proof here, it could be avoided if desired; see the proof of part (c) of Theorem 9.2.3.

**Proof of Proposition 8.6.1.** For a subset \( \Phi \) of \( S \) to be determined below, we will first prove for \( x \in \Phi \) and any bounded, Lipschitz continuous function \( h \) the lower bound
\[
\liminf_{n \to \infty} \frac{1}{n} \log E_x \{ \exp[-n h(L^n)] \} \geq - \inf_{\mu \in \mathcal{P}(S)} \{ I(\nu) + h(\nu) \}.
\] (8.36)

Afterwards, we will see how to convert this into a proof of the Laplace principle lower bound for all \( x \in S \).

These proofs will be carried out by working with
\[
W^n(x) \doteq - \frac{1}{n} \log E_x \{ \exp[-n h(L^n)] \},
\]
which according to Theorem 8.2.1 equals the minimal cost function
\[
V^n(x) \doteq \inf_{\{\nu^n_j\}} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot | X^n_j, L^n_j), p(X^n_j, \cdot)) + h(L^n) \right\}.
\]

Using this representation formula, we will prove (8.36) by showing the upper limit
\[
\limsup_{n \to \infty} W^n(x) \leq \inf_{\mu \in \mathcal{P}(S)} \{ I(\mu) + h(\mu) \}.
\] (8.37)

It suffices to show that any subsequence of \( \{W^n(x), n \in \mathbb{N}\} \) has a subsubsequence satisfying this upper limit. We will work with a fixed subsequence for the remainder of the proof, indexing it by \( n \in \mathbb{N} \).
Let $\varepsilon > 0$ be given and choose $\gamma \in \mathcal{P}(\mathcal{S})$ such that
\[ I(\gamma) + h(\gamma) \leq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{ I(\mu) + h(\mu) \} + \varepsilon < \infty. \]
Since $h$ is continuous and convergence in total variation norm implies weak convergence [Theorem A.3.1], Lemma 8.6.3 yields the existence of $\nu^* \in \mathcal{P}(\mathcal{S})$ and a transition probability function $q^*(x, dy)$ with the following properties: $\nu^*$ is an invariant measure of $q^*(x, dy)$, the Markov chain with initial distribution $\nu^*$ and transition probability function $q^*(x, dy)$ is ergodic, and
\[ h(\nu^*) \leq h(\gamma) + \varepsilon, \int_{\mathcal{S}} R(q^*(x, \cdot)\|p(x, \cdot)) \nu^*(dx) \leq I(\gamma) < \infty. \]
For each $j \in \{0, 1, \ldots, n - 1\}$ and $(x, \mu) \in \mathcal{S} \times \mathcal{M}_{j/n}(\mathcal{S})$ we define
\[ \nu^*_n(dy|x, \mu) = q^*(x, dy) \]
and consider two different choices for the distribution of $X^n_0$: $\delta_x$ and $\nu^*$. In the first case the corresponding measure on $(\tilde{\Omega}, \tilde{\mathcal{F}})$ is denoted by $\tilde{P}_x$. In the second case the controlled process $\{\tilde{X}^n_j, j = 0, 1, \ldots, n\}$ equals the first $(n+1)$ steps of the ergodic Markov chain with initial distribution $\nu^*$ and transition probability function $q^*(x, dy)$. The corresponding measure on $(\tilde{\Omega}, \tilde{\mathcal{F}})$ is denoted by $\tilde{P}^*$. We first study the convergence of the running costs
\[ \frac{1}{n} \sum_{j=0}^{n-1} \int_{\mathcal{S}} R(q^*(\tilde{X}^n_j, \cdot)\|p(\tilde{X}^n_j, \cdot)) = \frac{1}{n} \sum_{j=0}^{n-1} \int_{\mathcal{S}} R(q^*(\tilde{X}^n_j, \cdot)\|p(\tilde{X}^n_j, \cdot)) \cdot \nu^*(d\xi). \]
Define
\[ D^n \doteq \tilde{E}^* \left\{ \left| \frac{1}{n} \sum_{j=0}^{n-1} \int_{\mathcal{S}} R(q^*(\tilde{X}^n_j, \cdot)\|p(\tilde{X}^n_j, \cdot)) - \int_{\mathcal{S}} R(q^*(\xi, \cdot)\|p(\xi, \cdot)) \nu^*(d\xi) \right| \right\}, \]
where $\tilde{E}^*$ denotes expectation with respect to $\tilde{P}^*$, and
\[ D^n_x \doteq \tilde{E}_x \left\{ \left| \frac{1}{n} \sum_{j=0}^{n-1} \int_{\mathcal{S}} R(q^*(\tilde{X}^n_j, \cdot)\|p(\tilde{X}^n_j, \cdot)) - \int_{\mathcal{S}} R(q^*(\xi, \cdot)\|p(\xi, \cdot)) \nu^*(d\xi) \right| \right\}, \]
where $\tilde{E}_x$ denotes expectation with respect to $\tilde{P}_x$. Since the relative entropy is nonnegative and
\[ \tilde{E}^* \left\{ R(q^*(\tilde{X}^n_j, \cdot)\|p(\tilde{X}^n_j, \cdot)) \right\} = \int_{\mathcal{S}} R(q^*(\xi, \cdot)\|p(\xi, \cdot)) \nu^*(d\xi) \leq I(\gamma) < \infty, \]
we can apply the $L^1$ ergodic theorem [Theorem A.4.4], which implies that
\[ \lim_{n \to \infty} D^n = \lim_{n \to \infty} \int_{\mathcal{S}} D^n_x \nu^*(dx) = 0. \]
Hence by Chebyshev’s Inequality, for any $c > 0$
\[ \lim_{n \to \infty} \nu^* \{ x \in \mathcal{S} : D^n_x \geq c \} = 0. \]
The Borel–Cantelli Lemma guarantees that there exists a Borel set \( \Phi_1 \) and a subsubsequence of \( n \in \mathbb{N} \) such that \( \nu^*(\Phi_1) = 1 \) and such that whenever the initial condition \( \tilde{X}_0^n = x \) lies in \( \Phi_1 \)

\[
\lim_{n \to \infty} D^n_x = \lim_{n \to \infty} \tilde{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(q^*(\tilde{X}_j^n, \cdot)) ||p(\tilde{X}_j^n, \cdot)|| - \int_S R(q^*(\xi, \cdot)) ||p(\xi, \cdot)|| \nu^*(d\xi) \right\} = 0.
\]

We now consider the convergence of the corresponding subsequence of controlled empirical measures \( \{\tilde{L}^n\} \). According to Theorem A.6.1, \( \mathcal{S} \) admits an equivalent metric \( m(x, y) \) with the property that the space of bounded, uniformly continuous functions \( \mathcal{U}_b(S, m) \) is separable. Let \( \Xi \) be a countable dense subset of \( \mathcal{U}_b(S, m) \) and for \( g \in \Xi \) define

\[
A(g) = \left\{ \lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} g(\tilde{X}_j^n) = \int_S g \, d\nu^* \right\}.
\]

The pointwise ergodic theorem [Theorem A.4.4] implies that

\[
P^* \{ A(g) \} = \int_S P_x \{ A(g) \} \, \nu^*(dx) = 1,
\]

which in turn implies that there exists a Borel set \( \Phi_2(g) \) such that \( \nu^*(\Phi_2(g)) = 1 \) and \( \tilde{P}_x(A(g)) = 1 \) whenever the initial condition \( \tilde{X}_0^n = x \) lies in \( \Phi_2(g) \). Define \( \Phi_2 = \bigcap_{g \in \Xi} \Phi_2(g) \). Then whenever \( \tilde{X}_0^n = x \) lies in \( \Phi_2 \), we have with \( \tilde{P}_x \)-probability 1

\[
\lim_{n \to \infty} \int_S g \, d\tilde{L}^n = \lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} g(\tilde{X}_j^n) = \int_S g \, d\nu^*
\]

for all \( g \in \Xi \). Since \( \Xi \) is a dense subset of \( \mathcal{U}_b(S, m) \), it follows that whenever \( \tilde{X}_0^n = x \) lies in \( \Phi_2 \), \( \tilde{L}^n \implies \nu^* \) with \( \tilde{P}_x \)-probability 1. The continuity of \( h \) on \( \mathcal{P}(\mathcal{S}) \) then implies that

\[
\lim_{n \to \infty} h(\tilde{L}^n) = h(\nu^*).
\]

We now put these facts together. Define \( \Phi = \Phi_1 \cap \Phi_2 \). Then \( \nu^*(\Phi) = 1 \) and whenever \( \tilde{X}_0^n = x \) lies in \( \Phi \)

\[
\limsup_{n \to \infty} W^n(x) = \limsup_{n \to \infty} V^n(x)
\]

\[
\leq \lim_{n \to \infty} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(q^*(\tilde{X}_j^n, \cdot)) ||p(\tilde{X}_j^n, \cdot)|| + h(L^n) \right\}
\]

\[
= \int_S R(q^*(\xi, \cdot)) ||p(\xi, \cdot)|| \nu^*(d\xi) + h(\nu^*)
\]

\[
\leq I(\gamma) + h(\gamma) + \varepsilon
\]

\[
\leq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{ I(\mu) + h(\mu) \} + 2\varepsilon.
\]
Sending $\varepsilon \to 0$ yields the upper limit (8.37) whenever $x \in \Phi$.

A short argument will extend this upper limit to all $x \in \mathcal{S}$, thus proving the Laplace principle lower bound. This will be carried out using the Lipschitz continuity of $h$. Let $\ell_0$ be the number occurring in part (a) of Condition 8.4.1. Since $\mu^* \ll \nu^*$ [Lemma 8.6.3 (b)] and $\nu^*(\Phi) = 1$, $\mu^*(\Phi) = 1$. We claim that this implies

$$p^{(\ell_0)}(x, \Phi) = 1 \text{ for all } x \in \mathcal{S}. \quad (8.38)$$

If this claim were not true, then for some $x_0 \in \mathcal{S}$ we would have $p^{(\ell_0)}(x_0, \Phi^c) > 0$ and thus, by part (b) of Lemma 8.6.2, $\mu^*(\Phi^c) > 0$. Since this contradicts $\mu^*(\Phi) = 1$, the claim (8.38) is proved.

For $n \in \mathbb{N}$ we define

$$\tilde{L}^n = \frac{1}{n} \sum_{j=\ell_0}^{n+\ell_0-1} \delta_{X_j},$$

which also take values in $\mathcal{P}(\mathcal{S})$. Then uniformly with respect to $\omega$

$$\left\| \tilde{L}^n - L^n \right\|_v \leq \frac{2\ell_0}{n}.$$ 

Hence, if $M < \infty$ denotes the Lipschitz constant of $h$ with respect to the Lévy–Prohorov metric, then since $\mathcal{L}(L^n, \tilde{L}^n) \leq \|L^n - \tilde{L}^n\|_v$ [Theorem A.3.1] we have uniformly with respect to $\omega$

$$h(L^n) \leq h(\tilde{L}^n) + M \mathcal{L}(L^n, \tilde{L}^n) \leq h(\tilde{L}^n) + 2\ell_0M/n.$$ 

For each $x \in \mathcal{S}$ an application of the Markov property now gives for all $n > \ell_0$

$$E_x \{\exp[-n h(L^n)]\}
\begin{align*}
&= \int_{\mathcal{S}} E \{\exp[-n h(L^n)] | X_{\ell_0} = \zeta\} \, p^{(\ell_0)}(x, d\zeta) \\
&\geq \int_{\mathcal{S}} \exp[-2\ell_0M] E \{\exp[-n h(\tilde{L}^n)] | X_{\ell_0} = \zeta\} \, p^{(\ell_0)}(x, d\zeta) \\
&= \exp[-2\ell_0M] \int_{\Phi} E_\zeta \{\exp[-n h(L^n)]\} \, p^{(\ell_0)}(x, d\zeta). 
\end{align*} \quad (8.39)$$

Let $\varepsilon > 0$ be given. According to equation (8.37), for each $\zeta \in \Phi$ there exists $N(\zeta, \varepsilon) \in \mathbb{N}$ such that for all $n \geq N(\zeta, \varepsilon)$

$$W^n(\zeta) \leq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\} + \varepsilon.$$

We assume that $N(\zeta, \varepsilon)$ is the minimal positive integer with this property. Then the function mapping $\zeta \in \mathcal{S} \mapsto N(\zeta, \varepsilon)$ is measurable, and $\Phi$ can be written as the disjoint union of Borel sets $\bigcup_{i=1}^\infty \Phi^{(i)}$, where

$$\Phi^{(i)} \equiv \{\zeta \in \Phi : N(\zeta, \varepsilon) = i\}.$$
Since \( p^{(\Phi)}(x, \Phi) = 1 \), there exists \( i_0 \in \mathbb{N} \) such that \( p^{(\Phi)}(x, \Phi^{(i_0)}) > 0 \). Hence using (8.39), we have for all sufficiently large \( n \in \mathbb{N} \)
\[
\exp[-n W^n(x)] \geq \exp[-2\ell_0 M] \int_{\Phi} E_{\zeta} \{\exp[-n h(L^n)]\} p^{(\Phi)}(x, d\zeta)
\geq \exp[-2\ell_0 M] \int_{\Phi^{(i_0)}} E_{\zeta} \{\exp[-n h(L^n)]\} p^{(\Phi)}(x, d\zeta)
= \exp[-2\ell_0 M] \int_{\Phi^{(i_0)}} \exp[-n W^n(\zeta)] p^{(\Phi)}(x, d\zeta)
\geq \exp[-2\ell_0 M] \exp \left[ -n \left( \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\} + \varepsilon \right) \right] p^{(\Phi)}(x, \Phi^{(i_0)}).
\]

This gives
\[
\limsup_{n \to \infty} W^n(x) \leq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\} + \varepsilon.
\]

Sending \( \varepsilon \to 0 \), we have shown that for each \( x \in \mathcal{S} \) every subsequence of the original sequence \( \{W^n(x), n \in \mathbb{N}\} \) has a subsubsequence satisfying
\[
\limsup_{n \to \infty} W^n(x) \leq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\},
\]
which therefore must be satisfied by the entire sequence. This completes the proof of the Laplace principle lower bound
\[
\liminf_{n \to \infty} \frac{1}{n} \log E_x \{\exp[-n h(L^n)]\} \geq - \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}.
\]

Part (a) of Proposition 8.6.1 is proved.

Now suppose that in addition to Condition 8.4.1, Condition 8.4.2 also holds. We prove the uniform Laplace principle lower bound stated in part (b) of Proposition 8.6.1. For \( x \in \mathcal{S} \) and \( c > 0 \), \( B(x, c) \) denotes the open ball \( \{\zeta \in \mathcal{S} : d(\zeta, x) < c\} \). Let us assume that we have proved
\[
\limsup_{n \to \infty} \sup_{\zeta \in B(x, c)} W^n(x) \leq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}
\]
for any bounded, Lipschitz continuous function \( h \). This will be taken care of in the next paragraph. There exist finitely many points \( x_1, x_2, \ldots, x_r \) in \( \mathcal{S} \) such that the compact set \( K \) can be covered by \( \{B(x_i, c_i), i = 1, 2, \ldots, r\} \), where \( c_i \) is the radius associated with \( x_i \) in Condition 8.4.2. Hence the last display implies that
\[
\limsup_{n \to \infty} \sup_{\zeta \in K} W^n(\zeta) \leq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}.
\]

We now prove the upper limit (8.41). As in (8.39), for each \( \zeta \in B(x, c) \), each \( i \in \{1, 2, \ldots, k\} \), and all sufficiently large \( n \in \mathbb{N} \)
\[
E_{\zeta} \{\exp[-n h(L^n)]\} \geq \int_{\mathcal{S}} \exp[-2iM] E_{\zeta} \{\exp[-n h(L^n)]\} p^{(i)}(\zeta, dz)
\geq \int_{\mathcal{S}} \exp[-2kM] E_{\zeta} \{\exp[-n h(L^n)]\} p^{(i)}(\zeta, dz),
\]
where $M$ denotes the Lipschitz constant of $h$. Hence by Condition 8.4.2, for all sufficiently large $n \in \mathbb{N}$

$$
\inf_{\zeta \in B(x, \varepsilon)} E_\zeta\{\exp[-n \ h(L^n)]\} \\
\geq \exp[-2kM] \int_S E_z\{\exp[-n \ h(L^n)]\} \left( \frac{1}{k} \inf_{\zeta \in B(x, \varepsilon)} \sum_{i=1}^k p^{(i)}(\zeta, dz) \right) \\
\geq \frac{c}{k} \exp[-2kM] \int_S E_z\{\exp[-n \ h(L^n)]\} \beta(dz).
$$

In order to complete the proof, we must deal with uniformity properties of the Laplace principle with respect to the variable $z$. In the proof of the upper limit (8.40), this was done by introducing the sets $\Phi^{(i)}$. The analogous sets can be used in the same way to prove the desired upper limit (8.41). This finishes the proof of Proposition 8.6.1. ■

We have concluded our analysis of the empirical measures of a Markov chain whose transition probability function satisfies the Feller property as well as other conditions. In the next chapter we will present a number of extensions of this result. They are the Laplace principle for the empirical measures of a Markov chain whose transition probability function satisfies a weakened version of the Feller property and Laplace limits for the empirical measures of a Markov chain in the $\tau$-topology.
Chapter 9

Extensions of the Laplace Principle for the Empirical Measures of a Markov Chain

9.1 Introduction

In Theorem 8.4.3 we proved the Laplace principle for the empirical measures of a Markov chain. The techniques that we used are flexible enough that we can obtain, without too much extra work, two important extensions. These are carried out in the present chapter.

The first extension is given in Section 9.2. There we prove the Laplace principle for the empirical measures of a Markov chain when the Feller property is weakened to the property stated in Condition 9.2.1. Let \( \{X_j, j \in \mathbb{N}_0\} \) be a Markov chain taking values in a Polish space \( \mathcal{S} \) and having transition probability function \( p(x, dy) \). One of the two conditions needed to prove the Laplace principle upper bound and the compactness of the level sets of the rate function in Theorem 8.4.3 was that \( p(x, dy) \) satisfy the Feller property. This property states that the function mapping \( x \in \mathcal{S} \mapsto p(x, \cdot) \in \mathcal{P}(\mathcal{S}) \) is continuous in the weak topology on \( \mathcal{P}(\mathcal{S}) \). However, as the following elementary example shows, Markov chains arise naturally for which the Feller property does not hold. Let \( \{Y_j, j \in \mathbb{N}_0\} \) be a sequence of i.i.d. random variables taking values in \( \mathbb{R}^d \) and having common distribution \( \rho \). Let \( g \) be a measurable function mapping \( \mathbb{R}^d \) into \( \mathbb{R} \). For \( x \in \mathbb{R}^d \) and \( j \in \mathbb{N}_0 \), we define \( X_0 = x \) and

\[
X_{j+1} = g(X_j) + Y_j.
\]

The random variables \( \{X_j, j \in \mathbb{N}_0\} \) form a Markov chain with transition probability function \( p(x, dy) = \rho(dy - g(x)) \). Thus \( p(x, dy) \) satisfies the Feller property if and only if \( g \) is continuous. The set of points at which the function mapping \( x \mapsto p(x, \cdot) \) is discontinuous in the weak topology on \( \mathcal{P}(\mathbb{R}^d) \) equals the set of points at which \( g \) is discontinuous.

Condition 9.2.1 introduces a weakening of the Feller property under which we can still obtain the Laplace principle upper bound and the compactness of the level sets of the
rate function. The basic idea is that the set of points at which \( p(x, \cdot) \) is discontinuous should be negligible from a large deviation perspective. Condition 9.2.1 provides one set of conditions that are sufficient for this to happen. The weakening of the Feller property is satisfied by many stochastic processes that arise in applications and do not satisfy the ordinary Feller property; e.g., processes occurring in adaptive algorithms, in which the dynamics of the processes often involve indicator functions of certain events [7]. The Laplace principle is stated in Theorem 9.2.3. Since it dispenses with the global continuity of the transition probability function as expressed by the Feller property, Theorem 9.2.3 may be regarded as a Laplace principle for Markov chains with discontinuous statistics.

The paper [43] analyzes via the weak convergence approach an ergodic Markov chain that does not satisfy Condition 9.2.1. A variety of new phenomena arise, including a nonstandard form of the rate function and uniform large deviation results in the weak topology with no corresponding results in the \( \tau \)-topology.

It is instructive to compare the Laplace principle in Section 9.2 and in the previous chapter [Theorem 8.4.3] with the Laplace principles in Chapters 6 and 7. As the reader may recall, the rate function for the random walk model with discontinuous statistics studied in Chapter 7 differs markedly from the rate function for the random walk model with continuous statistics studied in Chapter 6. The reason is that for the processes studied in Chapter 7 the discontinuous statistics have a strong effect on their distributions. In contrast, in Section 9.2 we isolate conditions which guarantee that the rate functions for the empirical measures of Markov chains both with continuous statistics and with these particular discontinuous statistics are identical. Indeed, under the weakening of the Feller property given in Condition 9.2.1, the values of the transition probability function in a neighborhood of the set of discontinuities have, from the viewpoint of large deviations, a negligible effect on the distributions of the empirical measures. As a result, the form of the rate function is not affected. The weak convergence approach is a natural method for dealing with discontinuities of this type.

In Section 9.3 we consider another useful extension of the Laplace principle for the empirical measures of a Markov chain. Until this point in the book we have always topologized the space \( \mathcal{P}(\mathcal{S}) \) of probability measures on \( \mathcal{S} \) with the weak topology. A nice feature is that with this topology \( \mathcal{P}(\mathcal{S}) \) can be metrized as a Polish space. However, certain applications require that one consider Laplace limits for the empirical measures with respect to a stronger topology on \( \mathcal{P}(\mathcal{S}) \) known as the \( \tau \)-topology [68]. This is carried out in Theorem 9.3.2, in which we extend the class of functions \( h \) for which the Laplace principle in the previous chapter was proved. The extension is to functions that are bounded and continuous with respect to the \( \tau \)-topology on \( \mathcal{P}(\mathcal{S}) \). The proof is based on an approximation argument that allows us to modify the proof of the Laplace principle given in the previous chapter. References on the large deviation principle for the empirical measures of a Markov chain in the \( \tau \)-topology are given in Section 8.1.

In the next section we formulate and prove the Laplace principle for the empirical measures of a Markov chain whose transition probability function satisfies a weakened version of the Feller property.
9.2 Laplace Principle for the Empirical Measures of a Markov Chain with Discontinuous Statistics

Let \( \{X_j, j \in \mathbb{N}_0\} \) be a Markov chain that takes values in a Polish space \( \mathcal{S} \) and has stationary transition probabilities. We denote by \( p(x, dy) \) the transition probability function of the Markov chain and for \( n \in \mathbb{N} \) consider the empirical measures

\[
L^n \doteq \frac{1}{n} \sum_{j=0}^{n-1} \delta_{x_j}.
\]

The Laplace principle for the sequence of empirical measures is formulated in Theorem 9.2.3 under Condition 9.2.1, which is a weakening of the ordinary Feller property, and under other conditions assumed in the previous chapter. In order to state Condition 9.2.1, we define \( D_p \) to be the set of points at which the function mapping \( x \in \mathcal{S} \mapsto p(x, \cdot) \in \mathcal{P}(\mathcal{S}) \) is discontinuous in the topology of weak convergence on \( \mathcal{P}(\mathcal{S}) \). \( D_p \) is a Borel subset of \( \mathcal{S} \) [9, page 225]. The role of Condition 9.2.1 is to rule out the possibility of starting in a small neighborhood of \( D_p \) and returning to this set with a large probability. If this possibility can be avoided, then the Feller property can be exploited on the complement of \( D_p \). Although it is possible to weaken the condition, the resulting hypotheses are rather technical (see [40, Sect. 8.2]).

**Condition 9.2.1.** Let \( d(\cdot, \cdot) \) denote the metric on \( \mathcal{S} \) and let \( K \) be any compact subset of \( \mathcal{S} \). Then for each \( x \in K \) and each \( \delta > 0 \), there exist an open subset \( G \) of \( \mathcal{S} \) and \( \varepsilon > 0 \) such that \( D_p \subset G \) and \( p(\xi, G) < \delta \) whenever \( \xi \in \mathcal{S} \) satisfies \( d(\xi, x) < \varepsilon \).

Before stating the Laplace principle, we illustrate Condition 9.2.1 via an elementary example. More complicated and more realistic examples can easily be found.

**Example 9.2.2.** Let \( \{Y_j, j \in \mathbb{N}_0\} \) be a sequence of i.i.d. random variables taking values in \( \mathbb{R}^d \) and having common distribution \( \rho \). Let \( g \) be a Borel–measurable function mapping \( \mathbb{R}^d \) to \( \mathbb{R} \). For \( x \in \mathbb{R}^d \) and \( j \in \mathbb{N}_0 \), we define \( X_0 = x \) and

\[
X_{j+1} = g(X_j) + Y_j.
\]

As we pointed out in the introduction to this chapter, the random variables \( \{X_j, j \in \mathbb{N}_0\} \) form a Markov chain with transition probability function \( p(x, dy) = \rho(dy - g(x)) \). In addition, \( D_p \) equals the set of discontinuities of \( g \). Let us assume that \( \rho \) is absolutely continuous with respect to Lebesgue measure \( \lambda \) on \( \mathbb{R}^d \) and that \( \lambda(F) = 0 \), where \( F \) denotes the closure of \( D_p \). Hence, if we define for \( \eta > 0 \) the open set

\[
F^{(\eta)} = \{z \in \mathbb{R}^d : \|z - \zeta\| < \eta \text{ for some } \zeta \in F\},
\]

then since \( F^{(\eta)} \downarrow F \) as \( \eta \to 0 \), it follows that \( \lambda(F^{(\eta)}) \to \lambda(F) = 0 \). Given \( \delta > 0 \) there exists \( \theta > 0 \) such that whenever \( \lambda(A) < \theta \) for some Borel set \( A \), we have \( \rho(A) < \delta \). We now choose \( \eta > 0 \) such that \( \lambda(F^{(\eta)}) < \theta \). This implies that for all \( \zeta \in \mathbb{R}^d \)

\[
p(\zeta, F^{(\eta)}) = \rho(F^{(\eta)} - g(\zeta)) < \theta,
\]

which in turn implies that for all \( \xi \)

\[
p(\xi, F^{(\eta)}) = \rho(F^{(\eta)} - g(\xi)) < \delta.
\]

This verifies Condition 9.2.1 with \( G \doteq F^{(\eta)} \).
We next state the Laplace principle for the empirical measures, replacing the ordinary Feller property by Condition 9.2.1. The Laplace principle holds with the same rate function \(I\) as in the previous chapter. We recall that for \(\mu \in \mathcal{P}(\mathcal{S})\) this function is defined by

\[
I(\mu) \triangleq \inf_{\{\mu \in \mathcal{T} \mid q = \mu\}} \int_{\mathcal{S}} R(q(x, \cdot)) \|p(x, \cdot)\mu(dx),
\]

where \(\mathcal{T}\) denotes the set of all transition probability functions on \(\mathcal{S}\).

**Theorem 9.2.3.** For each bounded continuous function \(h\) mapping \(\mathcal{P}(\mathcal{S})\) into \(\mathbb{R}\), the following conclusions hold.

(a) Under Conditions 8.2.2 and 9.2.1, \(I\) has compact level sets.

(b) Under Conditions 8.2.2 and 9.2.1, for each compact subset \(K\) of \(\mathcal{S}\) we have the uniform Laplace principle upper bound

\[
\limsup_{n \to \infty} \sup_{x \in K} \frac{1}{n} \log E_x \{\exp[-n h(L^n)]\} \leq - \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}.
\]

(c) Under Condition 8.4.1, for each initial condition \(X_0 = x \in \mathcal{S}\) we have the Laplace principle lower bound

\[
\liminf_{n \to \infty} \frac{1}{n} \log E_x \{\exp[-n h(L^n)]\} \geq - \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}.
\]

Under Conditions 8.4.1 and 8.4.2, we have for each compact subset \(K\) of \(\mathcal{S}\) the uniform Laplace principle lower bound

\[
\liminf_{n \to \infty} \inf_{x \in K} \frac{1}{n} \log E_x \{\exp[-n h(L^n)]\} \geq - \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}.
\]

(d) In summary, under Conditions 8.2.2, 9.2.1, and 8.4.1, for any initial condition \(X_0 = x \in \mathcal{S}\) the sequence \(\{L^n\}\) satisfies the Laplace principle on \(\mathcal{P}(\mathcal{S})\) with rate function \(I\). Under Conditions 8.2.2, 9.2.1, 8.4.1, and 8.4.2, the sequence \(\{L^n\}\) satisfies the Laplace principle on \(\mathcal{P}(\mathcal{S})\) with rate function \(I\) uniformly on compacts.

Part (b) of Condition 8.4.1 requires the existence of an invariant measure of \(p(x, dy)\). In the context of the full Laplace principle in part (d) of this theorem, the existence of an invariant measure of \(p(x, dy)\) is automatic under the other conditions that we assume. This will be shown in Proposition 9.2.6.

The technique of proof of Theorem 9.2.3 is straightforward. A review of Section 8.6 shows that the Feller property is not needed in the proofs of the Laplace principle lower bounds, and hence these proofs remain unchanged. Since we will prove a Laplace principle upper bound with the same rate function as in Chapter 8, the lower bound of Chapter 8 cannot be improved. Concerning the Laplace principle upper bound, if \(p(x, dy)\) satisfies the Feller property, then \(D_p\) is empty. Condition 9.2.1 weakens this to the assumption that \(D_p\) is appropriately small relative to \(p(x, .)\). Under this condition we prove that if \(\{\nu^n\}\) is a convergent sequence of admissible control measures satisfying

\[
\liminf_{n \to \infty} R(\nu^n \| L^n \otimes p) < \infty
\]

and if the associated sequence of controlled empirical
measures \( \{\tilde{L}^n\} \) converges in distribution to \( \tilde{L} \), then \( \tilde{L}(D_p) = 0 \) w.p.1. In other words, w.p.1 the limit of the controlled empirical measures does not charge the set \( D_p \) of discontinuities of \( p(x, \cdot) \). This implies that \( D_p \) is unimportant as far as the large deviation analysis is concerned, and we can proceed to prove the Laplace principle upper bound essentially as in the previous chapter. Similar observations are used to prove that the rate function has the same form as in the previous chapter and that it has compact level sets.

We now turn to the proof of Theorem 9.2.3. We must prove parts (a) and (b), which correspond to parts (a) and (b) of the Laplace principle for Markov chains with continuous statistics given in Theorem 8.4.3. We first discuss part (b) of the theorem, postponing the proof of part (a) until the end of the section. In order to prove part (b), it suffices to show that for each bounded continuous function \( h \) mapping \( S \) into \( \mathbb{R} \) the functions

\[
W^n(x) = -\frac{1}{n} E_x \{ \exp[-n h(L^n)] \}
\]

satisfy for each compact set \( K \)

\[
\liminf_{n \to \infty} \inf_{x \in K} W^n(x) \geq \inf_{\mu \in \mathcal{P}(S)} \{ I(\mu) + h(\mu) \}.
\]

We will ease the discussion by focusing on the nonuniform version of this lower limit; namely,

\[
\liminf_{n \to \infty} W^n(x) \geq \inf_{\mu \in \mathcal{P}(S)} \{ I(\mu) + h(\mu) \} \tag{9.2}
\]

for each \( x \in S \). It is left to the reader to check that the uniform lower limit can be proved similarly.

It is helpful to recall how (9.2) was proved in Section 8.3, where the ordinary Feller property was assumed. The starting point was the stochastic control representation formula for \( W^n(x) \) given in Theorem 8.2.1:

\[
W^n(x) = V^n(x) \doteq \inf_{\{\nu^*_j\}} \tilde{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu_j^n(\cdot|\bar{X}_j^n, \bar{L}_j^n), p(\bar{X}_j^n, \cdot)) + h(\tilde{L}^n) \right\}.
\]

It sufficed to prove (9.2) when \( n \) is replaced by any subsequence along which the uniformly bounded functions \( W^n(x) \) converge.

Given \( \varepsilon > 0 \), for \( n \in \mathbb{N} \) we chose an admissible control sequence \( \{\nu^*_j\} \) satisfying

\[
V^n(x) + \varepsilon \geq \tilde{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu_j^n(\cdot|\bar{X}_j^n, \bar{L}_j^n), p(\bar{X}_j^n, \cdot)) + h(\tilde{L}^n) \right\}.
\]

(9.3)

Let \( \{\nu^n, n \in \mathbb{N}\} \) be the corresponding sequence of admissible control measures defined by

\[
\nu^n(dx \times dy) \doteq \frac{1}{n} \sum_{j=0}^{n-1} \delta_{\bar{X}_j^n}(dx) \times \nu_j^n(dy|\bar{X}_j^n, \bar{L}_j^n).
\]

Then according to Theorem 8.2.8, which uses only Condition 8.2.2 and not the Feller property, there exists a subsequence of \( \{(\nu^n, \tilde{L}^n)\} \) along which

\[
(\nu^n, \tilde{L}^n) \xrightarrow{D} (\nu, \tilde{L}).
\]
CHAPTER 9. EXTENSIONS OF LAPLACE PRINCIPLE

The pair \((\nu, \tilde{L})\) takes values in \(\mathcal{P}(\mathcal{S} \times \mathcal{S}) \times \mathcal{P}(\mathcal{S})\), and w.p.1 \(\tilde{L}\) equals the first marginal of \(\nu\). By using the Skorohod Representation Theorem, we assumed that the convergence occurs w.p.1. Parts (b) and (c) of Theorem 8.28 introduced a stochastic kernel \(\bar{q}(x, dy) \approx \nu(dy|x)\) satisfying w.p.1

\[
\nu(A \times B) = \int_A \bar{q}(x, B) \tilde{L}(dx) \quad \text{and} \quad \tilde{L}(B) = \int_S \bar{q}(x, B) \tilde{L}(dx)
\]

for Borel sets \(A\) and \(B\). The second of these equations expresses the fact that w.p.1 \(\tilde{L}\) is an invariant measure of \(q(x, dy)\). It also implies that w.p.1 the second marginal of \(\nu\) equals \(\tilde{L}\).

In order to prove the lower limit (9.2), we applied Corollary C.3.3 and Jensen’s Inequality to the right-hand side of formula (9.3), obtaining

\[
W^n(x) + \varepsilon = V^n(x) + \varepsilon \geq \bar{E}_x \{ R(\nu^n \| \tilde{L}^n \otimes p) + h(\tilde{L}^n) \}. \tag{9.4}
\]

The key step in proof of the lower limit was then (8.19), which we repeat here for easy reference. As noted in the paragraph leading up to equation (8.22), the last two lines of the following display are consequences of the definition of \(I\) and the fact that w.p.1 \(\tilde{L}\) is an invariant measure of \(\bar{q}(x, dy)\):

\[
\lim_{n \to \infty} W^n(x) + \varepsilon = \lim_{n \to \infty} V^n(x) + \varepsilon \geq \liminf_{n \to \infty} \bar{E}_x \{ R(\nu^n \| \tilde{L}^n \otimes p) + h(\tilde{L}^n) \} \geq \bar{E}_x \{ R(\tilde{L} \otimes \bar{q} \| \tilde{L} \otimes p) + h(\tilde{L}) \} \geq \bar{E}_x \{ R(\tilde{L} \otimes \bar{q} \| \tilde{L} \otimes p) + h(\tilde{L}) \}
\]

\[
\geq \bar{E}_x \{ I(\tilde{L}) + h(\tilde{L}) \} \geq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{ I(\mu) + h(\mu) \}.
\]

Since \(\varepsilon > 0\) is arbitrary and since this display is valid for some subsequence of any convergent subsequence of \(\{W^n(x), n \in \mathbb{N}\}\), the desired lower limit

\[
\liminf_{n \to \infty} W^n(x) \geq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{ I(\mu) + h(\mu) \}
\]

is proved for the entire sequence \(\{W^n(x), n \in \mathbb{N}\}\).

The Feller property is needed only once in (9.5). Namely, if the function mapping \(x \in \mathcal{S} \mapsto p(x, \cdot) \in \mathcal{P}(\mathcal{S})\) is continuous, then Lemma 8.3.2 and the probability−1 weak convergence \(\tilde{L}^n \Rightarrow \tilde{L}\) yield the probability−1 weak convergence \(\tilde{L}^n \otimes p \Rightarrow \tilde{L} \otimes p\). This in turn implies that w.p.1 \(\liminf_{n \to \infty} R(\nu^n \| \tilde{L}^n \otimes p) \geq R(\nu \| \tilde{L} \otimes p)\). Fatou’s Lemma then yields

\[
\liminf_{n \to \infty} \bar{E}_x \{ R(\nu^n \| \tilde{L}^n \otimes p) \} \geq \bar{E}_x \{ R(\nu \| \tilde{L} \otimes p) \}. \tag{9.6}
\]
9.2. DISCONTINUOUS STATISTICS

Clearly, in order to prove

\[ \liminf_{n \to \infty} W^n(x) \geq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{ I(\mu) + h(\mu) \}, \]

one can use any set of conditions besides the Feller property which will yield the probability-1 weak convergence \( \tilde{L}^n \otimes p \Rightarrow L \otimes p \). As we will see in the next proposition, a sufficient condition for this weak convergence is that w.p.1 \( \tilde{L}(D_p) = 0 \). Later on we will derive this equality from Condition 9.2.1 and from the probability-1 boundedness of the sequence \( \{ R(\nu^n \| \tilde{L}^n \otimes p) \} \).

**Proposition 9.2.4.** We assume Condition 9.2.1. If \( \{ \mu_n, n \in \mathbb{N} \} \) is a sequence of probability measures on \( \mathcal{S} \) converging weakly to \( \mu \) and \( \mu(D_p) = 0 \), then \( \mu_n \otimes p \Rightarrow \mu \otimes p \) on \( \mathcal{S} \times \mathcal{S} \).

**Proof.** According to Theorem A.3.14, one can prove the weak convergence by considering integrals with respect to the functions \( g(x) h(y) \), where \( g \) and \( h \) are arbitrary bounded continuous functions mapping \( \mathcal{S} \) into \( \mathbb{R} \). By Condition 9.2.1, the set of discontinuities of the bounded function mapping

\[ x \in \mathcal{S} \mapsto \left( \int_{\mathcal{S}} h(y) p(x, dy) \right) g(x) \in \mathbb{R} \]

is a subset of \( D_p \). Since \( \mu_n \Rightarrow \mu \) and \( \mu(D_p) = 0 \), the extension of the Continuous Mapping Theorem given in Theorem A.3.11 yields the limit

\[ \lim_{n \to \infty} \int_{\mathcal{S} \times \mathcal{S}} g(x) h(y)(\mu_n \otimes p)(dx \times dy) = \lim_{n \to \infty} \int_{\mathcal{S}} \left( \int_{\mathcal{S}} h(y) p(x, dy) \right) g(x) \mu_n(dx) \]

\[ = \int_{\mathcal{S}} \left( \int_{\mathcal{S}} h(y) p(x, dy) \right) g(x) \mu(dx) \]

\[ = \int_{\mathcal{S} \times \mathcal{S}} g(x) h(y)(\mu \otimes p)(d\xi \times dy). \]

This completes the proof. \( \blacksquare \)

The next proposition gives a condition that implies the equality \( \mu(D_p) = 0 \), which is a hypothesis in the proposition just proved. We will apply this in the proof of the Laplace principle upper bound under Condition 9.2.1 to show that w.p.1 \( \tilde{L}(D) = 0 \) and thus \( \tilde{L}^n \otimes p \Rightarrow L \otimes p \). As we will see afterwards, the boundedness condition \( \lim \inf_{n \to \infty} R(\tau_n \| \mu_n \otimes p) < \infty \) appearing in the next proposition is w.p.1 automatically satisfied in the proof of the Laplace principle upper bound, in which \( \tau_n \) will equal the admissible control measures \( \nu^n \) and \( \tau \) will equal the limiting quantity \( \nu \). If \( \sigma \) is a probability measure on \( \mathcal{S} \times \mathcal{S} \), then \( \sigma_1 \) and \( \sigma_2 \) denote, respectively, the first marginal and the second marginal of \( \sigma \).

**Proposition 9.2.5.** We assume Condition 9.2.1. Let \( \{ \mu_n, n \in \mathbb{N} \} \) be a sequence of probability measures on \( \mathcal{S} \) converging weakly to \( \mu \) and \( \{ \tau_n, n \in \mathbb{N} \} \) a sequence of probability measures on \( \mathcal{S} \times \mathcal{S} \) converging weakly to \( \tau \). Assume that for each \( n \in \mathbb{N} \) \( (\tau_n)_1 \) equals \( \mu_n \) and that both \( \tau_1 \) and \( \tau_2 \) equal \( \mu \). Then \( \mu(D_p) > 0 \) implies that

\[ \lim_{n \to \infty} R(\tau_n \| \mu_n \otimes p) = \infty. \]

By contraposition, it follows that if \( \lim \inf_{n \to \infty} R(\tau_n \| \mu_n \otimes p) < \infty \), then \( \mu(D_p) = 0 \).
Proof. By part (a) of Lemma 8.5.1, the equality \((\tau_n)_{1} = \mu_n\) implies that there exists a transition probability function \(q_n(x, dy)\) on \(S\) such that \(\tau_n = \mu_n \otimes q_n\). Set \(\alpha \doteq \mu(D_p) > 0\). The tightness of \(\mu\) [Theorem A.2.3] and Prohorov’s Theorem guarantee that there exists a compact set \(K\) such that
\[
\mu(K \cap D_p) \geq \alpha / 2 \quad \text{and} \quad \sup_{n \in \mathbb{N}} \mu_n(K^c) \leq \alpha / 8.
\]
In addition, if \(G\) is any open subset of \(S\) containing \(D_p\), then the weak convergence \(\tau_n \Rightarrow \tau\) implies that
\[
\liminf_{n \to \infty} \int_S q_n(x, G) \mu_n(dx) = \liminf_{n \to \infty} \tau_n(S \times G)
\]
\[
\geq \tau(S \times G) = \tau_2(G) = \mu(G) \geq \mu(K \cap D_p) \geq \alpha / 2.
\]
Hence for all sufficiently large \(n \in \mathbb{N}\)
\[
\int_K q_n(x, G) \mu_n(dx) = \int_S q_n(x, G) \mu_n(dx) - \int_{K^c} q_n(x, G) \mu_n(dx) \geq \alpha / 4 - \alpha / 8 = \alpha / 8.
\]
According to Condition 9.2.1, given \(\delta > 0\) there exists for any \(x \in K\) a number \(\varepsilon_x > 0\) and an open set \(G_x\) such that \(D_p \subset G_x\) and such that whenever \(\xi \in S\) satisfies \(d(\xi, x) < \varepsilon_x\), then \(p(\xi, G) < \delta\). We denote by \(B(x, \varepsilon_x)\) the open ball with center \(x\) and radius \(\varepsilon_x\). Since \(K\) is compact and \(K \subset \bigcup_{x \in K} B(x, \varepsilon_x)\), there exist finitely many points \(x_1, x_2, \ldots, x_r\) in \(K\) such that \(K \subset \bigcup_{i=1}^r B(x_i, \varepsilon_{x_i})\). The open set \(G \doteq \cap_{i=1}^r G_{x_i}\) has the property that \(p(\xi, G) \leq \delta\) for any \(\xi \in K\).

We now show that \(\lim_{n \to \infty} R(\tau_n \| \mu^n \otimes p) = \infty\). Part (g) of Lemma 1.4.3 implies that for each \(x \in S\)
\[
R(q_n(x, .)\| p(x, .)) \geq q_n(x, G) \log \frac{q_n(x, G)}{p(x, G)} - 1.
\]
Together with part (f) of Lemma 1.4.3 and Jensen’s Inequality, this yields for all sufficiently large \(n \in \mathbb{N}\)
\[
R(\tau_n \| \mu_n \otimes p)
\]
\[
= R(\mu_n \otimes q_n \| \mu_n \otimes p)
\]
\[
= \int_S R(q_n(x, .)\| p(x, .)) \mu_n(dx)
\]
\[
\geq \int_K R(q_n(x, .)\| p(x, .)) \mu_n(dx)
\]
\[
\geq \int_K \left[ q_n(x, G) \log \frac{q_n(x, G)}{p(x, G)} - 1 \right] \mu_n(dx)
\]
\[
\geq \int_K q_n(x, G) \log q_n(x, G) \mu_n(dx) - \int_K q_n(x, G) \log p(x, G) \mu_n(dx) - 1
\]
\[
\geq \left( \int_K q_n(x, G) \mu_n(dx) \right) \log \left( \frac{1}{\mu_n(K)} \right) \int_K q_n(x, G) \mu_n(dx)
\]
\[
- \int_K q_n(x, G) \log p(x, G) \mu_n(dx) - 1
\]
\[
\geq (\alpha / 8) \cdot \log(\alpha / 8) - \log \delta - 1.
\]
9.2. DISCONTINUOUS STATISTICS

Since $\delta > 0$ is arbitrary, we conclude that

$$\lim_{n \to \infty} R(\tau_n \| \mu_n \otimes p) = \infty.$$ 

This completes the proof. ■

We are now ready to prove the nonuniform version of the Laplace principle upper bound stated in part (b) of Theorem 9.2.3. Let $\{W^n(x), n \in \mathbb{N}\}$ be any convergent subsequence. Since $|W^n(x)| \leq \|h\|_{\infty}$, formula (9.4) guarantees that

$$\sup_{n \in \mathbb{N}} \mathbb{E}_x \{R(\nu^n \| \tilde{L}^n \otimes p)\} < \infty. \tag{9.7}$$

By Theorem 8.2.8 and the Skorohod Representation Theorem, there exist a probability space, a null set $\mathcal{N}$ in this space, and a subsequence of $\{(\nu^n, \tilde{L}^n)\}$ along which $(\nu^n, \tilde{L}^n) \xrightarrow{\text{w.p.1}} (\nu, \tilde{L})$ for all $\omega \in \mathcal{N}^c$. We are done once we show that w.p.1 $\tilde{L}^n \otimes p \xrightarrow{\text{w.p.1}} \tilde{L} \otimes p$. Indeed, the lower limit (9.6) is implied by this convergence, which was the only step in the proof of the Laplace principle in Theorem 8.4.3 that used the Feller property. In order to prove that $\tilde{L}^n \otimes p \xrightarrow{\text{w.p.1}} \tilde{L} \otimes p$, we apply Propositions 9.2.4 and 9.2.5 with $\mu_n \doteq \tilde{L}^n$, $\mu \doteq \tilde{L}$, $\tau_n \doteq \nu^n$, and $\tau \doteq \nu$. If we had $\tilde{L}(D_p) > 0$ on a subset of $\mathcal{N}^c$ of positive probability, then on this subset we would have

$$\lim_{n \to \infty} R(\nu^n \| \tilde{L}^n \otimes p) = \infty,$$

and Fatou’s Lemma would yield

$$\lim_{n \to \infty} \mathbb{E}_x \{R(\nu^n \| \tilde{L}^n \otimes p)\} = \infty.$$

This contradiction to (9.7) allows us to conclude that w.p.1 $\tilde{L}(D_p) = 0$. Proposition 9.2.4 then implies that $\tilde{L}^n \otimes p \xrightarrow{\text{w.p.1}} \tilde{L} \otimes p$. This completes the proof of the nonuniform version of the Laplace principle upper bound stated in part (b) of Theorem 9.2.3.

Proposition 8.3.4 showed that a transition probability function satisfying Condition 8.2.2 and the Feller property has an invariant measure. We now prove the same result when the Feller property is replaced by Condition 9.2.1.

**Proposition 9.2.6.** Under Conditions 8.2.2 and 9.2.1, $p(x, dy)$ has an invariant measure. In particular, if $\mathcal{S}$ is compact and $p(x, dy)$ satisfies Condition 9.2.1, then $p(x, dy)$ has an invariant measure.

**Proof.** It suffices to prove the first statement since if $\mathcal{S}$ is compact, then Condition 8.2.2 is satisfied by the constant function $U = 0$. In Theorem 8.2.8, for each $n \in \mathbb{N}$, $j \in \{0, 1, \ldots, n - 1\}$, $x \in \mathcal{S}$, and $\mu \in \mathcal{M}_{j/n}(\mathcal{S})$, we choose as the admissible control sequence $\nu^x_j(dy|x, \mu) \doteq p(x, dy)$. Then

$$\nu^n(dx \times dy) = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{x_j}(dx) \times p(\tilde{X}^n_j, dy) = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{x_j}(dx) \otimes p(x, dy) = \tilde{L}^n(dx) \otimes p(x, dy),$$

where $\tilde{L}^n(dx) = \prod_{j=0}^{n-1} \tilde{L}(dx_j)$. Therefore, $\nu^n(dx \times dy)$ is an invariant measure.
and so $R(\nu^n \| \tilde{L}^n \otimes p) = 0$. Since with $\nu^n(dy|x, \mu) \doteq p(x, dy)$ the bound (8.15) is trivially satisfied, Theorem 8.2.8 can be applied. There exists a subsequence of $\{ (\nu^n, \tilde{L}^n), n \in N \}$ which converges in distribution to a pair of stochastic kernels $(\nu, \tilde{L})$, where w.p.1 $\tilde{L}$ is both the first and second marginals of $\nu$. The Skorohod Representation Theorem allows us to assume that the convergence $(\nu^n, \tilde{L}^n) \to (\nu, \tilde{L})$ occurs w.p.1. As in the proof of the weak convergence $\tilde{L}^n \otimes p \Rightarrow \tilde{L} \otimes p$ just given, since $R(\nu^n \| \tilde{L}^n \otimes p) = 0$, we can apply Propositions 9.2.4 and 9.2.5 to the previous display and conclude that w.p.1 $\nu(dx \times dy) = \tilde{L}(dx) \otimes p(x, dy)$. With probability 1, since for all Borel sets $B$

\[
\tilde{L}(B) = \nu_2(B) = \int_S p(x, B) \tilde{L}(dx),
\]

$\tilde{L}$ is an invariant measure of $p(x, dy)$. ■

We end this section by proving the lower semicontinuity of $I$ under Conditions 8.2.2 and 9.2.1. Under Condition 8.2.2 alone, part (b) of Proposition 8.5.2 shows that for each $M < \infty$ the level set \{ $\mu \in \mathcal{P}(S) : I(\mu) \leq M$ \} is relatively compact. Hence, if we can prove the lower semicontinuity of $I$ under Condition 9.2.1, then it will follow that each level set is compact. This will yield part (a) of Theorem 9.2.3.

**Proposition 9.2.7.** Under Condition 9.2.1 the function $I(\mu)$ defined in (9.1) is a lower semicontinuous function of $\mu \in \mathcal{P}(S)$. It follows that under Conditions 8.2.2 and 9.2.1 $I$ has compact level sets.

**Proof.** For $\mu \in \mathcal{P}(S)$ we define

\[
A(\mu) \doteq \{ \tau \in \mathcal{P}(S \times S) : \tau_1 = \tau_2 = \mu \}.
\]

Let \{ $\mu_n, n \in N$ \} be any sequence of probability measures in $\mathcal{P}(S)$ converging weakly to $\mu$. For each $n \in N$ there exists $\tau_n \in A(\mu_n)$ such that $I(\mu_n) = R(\tau_n \| \mu_n \otimes p)$ [Lemma 8.5.1 (c)]. It suffices to prove that given any subsequence of \{ $\mu_n$ \} satisfying \( \liminf_{n \to \infty} I(\mu_n) < \infty \), there exists a subsequence such that

\[
\liminf_{n \to \infty} I(\mu_n) \geq I(\mu).
\]

Since for each $n$ (\( \tau_n \)) = (\( \tau_n \)) = $\mu_n$ and since $\mu_n \Rightarrow \mu$, the subsequence \{ $\tau_n$ \} is tight. Therefore, there exist a subsequence such that $\tau_n \Rightarrow \tau \in A(\mu)$. In addition

\[
\liminf_{n \to \infty} R(\tau_n \| \mu_n \otimes p) = \liminf_{n \to \infty} I(\mu_n) < \infty.
\]

Proposition 9.2.5 implies that $\mu(D_p) = 0$, and thus Proposition 9.2.4 allows us to conclude that $\mu_n \otimes p \Rightarrow \mu \otimes p$. The lower semicontinuity of $R(\cdot \| \cdot)$ now yields

\[
\liminf_{n \to \infty} I(\mu_n) = \liminf_{n \to \infty} R(\tau_n \| \mu_n \otimes p) \geq R(\tau \| \mu \otimes p) \geq I(\mu).
\]

This proves the lower semicontinuity of $I$. ■
9.3. **EMPIRICAL MEASURES IN THE \( \tau \)-TOPOLOGY**

This proof of the lower semicontinuity of \( I \) is a deterministic version of the proof of the Laplace principle upper bound. Indeed, the lower limit

\[
\liminf_{n \to \infty} R(\tau_n \| \mu_n \otimes p) \geq R(\tau \| \mu \otimes p)
\]

corresponds to lines two and three in (9.5), which was a key step in the proof of the upper bound. This completes the proof of the Laplace principle for the empirical measures stated in Theorem 9.2.3.

In the next section we consider the empirical measures of a Markov chain in the \( \tau \)-topology.

### 9.3 Laplace Limit for the Empirical Measures of a Markov Chain in the \( \tau \)-Topology

As in Chapter 8, let \( \{X_j, j \in \mathbb{N}_0\} \) be a Markov chain that takes values in a Polish space \( S \) and has transition probability function \( p(x, dy) \) and let \( \{L^n, n \in \mathbb{N}\} \) be the associated sequence of empirical measures. In this section we consider an extension of the class of functions \( h \) for which the Laplace limit

\[
\lim_{n \to \infty} \frac{1}{n} \log E_x \{\exp[-n h(L^n)]\} = - \inf_{\mu \in \mathcal{P}(S)} \{I(\mu) + h(\mu)\}
\]

can be proved. The class of functions that we wish to consider are those that are bounded and continuous with respect to the \( \tau \)-topology.

Let \( \{\theta_n, n \in \mathbb{N}\} \) be a sequence of probability measures on \( S \). We say that the sequence **converges in the \( \tau \)-topology** to \( \theta \), and write \( \theta_n \xrightarrow{\tau} \theta \), if for each bounded measurable function \( g \) mapping \( S \) into \( \mathbb{R} \)

\[
\lim_{n \to \infty} \int_S g \, d\theta_n = \int_S g \, d\theta.
\]

The \( \tau \)-topology on \( \mathcal{P}(S) \) is the topology generated by the basic open sets

\[
A = \left\{ \theta \in \mathcal{P}(S) : \left| \int_S g_i \, d\theta - \int_S g_i \, d\gamma \right| < \varepsilon, i = 1, 2, \ldots, k \right\},
\]

where \( \gamma \) is a probability measure on \( S \), \( \varepsilon \) is positive, \( k \) is a positive integer, and \( g_1, g_2, \ldots, g_k \) are bounded measurable functions mapping \( S \) into \( \mathbb{R} \). Clearly, if \( \theta_n \xrightarrow{\tau} \theta \), then \( \theta_n \Rightarrow \theta \), but the reverse implication is false.

The Laplace limit for functions \( h \) that are bounded and continuous in the \( \tau \)-topology is useful, for example, if one wants to consider functions of the form \( h(L^n) = \int_S g \, dL^n \) where \( g \) is bounded and measurable but not continuous. However, there are some subtleties associated with the \( \tau \)-topology which we would like to point out. In order to explain them, we introduce the notations \( \mathcal{B}_\mathbb{R}, \mathcal{B}_S, \) and \( \mathcal{B}_{\mathcal{P}(S)} \) for the Borel \( \sigma \)-fields of the respective spaces, where in the latter case \( \mathcal{P}(S) \) is equipped with the weak topology. We also define \( \mathcal{T}_{\mathcal{P}(S)} \) to be the Borel–\( \sigma \) field of \( \mathcal{P}(S) \) induced by the \( \tau \)-topology. First of all,
\( \mathcal{P}(S) \) equipped with the \( \tau \)-topology is neither a metrizable space nor a separable space. Furthermore, the function mapping \( y \in (S, \mathcal{B}_S) \mapsto \delta_y \in (\mathcal{P}(S), \mathcal{T}_{\mathcal{P}(S)}) \) need not be measurable. As a result, for \( A \in \mathcal{T}_{\mathcal{P}(S)} \) the set \( \{ \omega : L^n(\omega, \cdot) \in A \} \) need not be measurable in the underlying probability space. In other words, if the \( \sigma \)-field \( \mathcal{T}_{\mathcal{P}(S)} \) is used, then \( L^n \) need not be a random variable. What steps can one take to insure the measurability of \( L^n \)? Given \( g \) any bounded measurable function mapping \( S \) into \( \mathbb{R} \), let \( \mathcal{F}_{\mathcal{P}(S)} \subset \mathcal{T}_{\mathcal{P}(S)} \) be the smallest \( \sigma \)-field with respect to which each of the functions mapping \( \nu \in \mathcal{P}(S) \mapsto \int_S g \, d\nu \in \mathbb{R} \) is measurable. In the next paragraph we prove that the function mapping \( y \in (S, \mathcal{B}_S) \mapsto \delta_y \in (\mathcal{P}(S), \mathcal{F}_{\mathcal{P}(S)}) \) is measurable. It follows that if the \( \sigma \)-field \( \mathcal{F}_{\mathcal{P}(S)} \) is used, then as a mapping from the underlying probability space into \( \mathcal{P}(S) \), \( L^n \) is a random variable. It is important to note that \( \mathcal{F}_{\mathcal{P}(S)} \) equals the Borel \( \sigma \)-field \( \mathcal{B}_{\mathcal{P}(S)} \) corresponding to the weak topology on \( \mathcal{P}(S) \) [12, Lem. 2.1].

Let \( \Psi_b(S) \) denote the space of bounded measurable functions mapping \( S \) into \( \mathbb{R} \) and let \( \zeta \) denote the function mapping \( y \in (S, \mathcal{B}_S) \mapsto \delta_y \in (\mathcal{P}(S), \mathcal{F}_{\mathcal{P}(S)}) \). In order to prove that \( \zeta \) is measurable, we define for \( g \in \Psi_b(S) \) the mappings

\[
f_g : \nu \in (\mathcal{P}(S), \mathcal{F}_{\mathcal{P}(S)}) \mapsto \int_S g \, d\nu \in (\mathbb{R}, \mathcal{B}_{\mathbb{R}})
\]

and

\[
\xi_g : y \in (S, \mathcal{B}_S) \mapsto \int_S g \, d\delta_y \in (\mathbb{R}, \mathcal{B}_{\mathbb{R}}).
\]

These mappings are related by \( \xi_g = f_g \circ \zeta \). Since \( \xi_g(y) = g(y) \), \( \xi_g \) is measurable for each \( g \in \Psi_b(S) \), and since \( \mathcal{F}_{\mathcal{P}(S)} \) is the smallest \( \sigma \)-field with respect to which each \( f_g \) is measurable,

\[
\mathcal{F}_{\mathcal{P}(S)} = \sigma \left[ \bigcup_{g \in \Psi_b(S)} f_g^{-1}(\mathcal{B}_{\mathbb{R}}) \right].
\]

Hence

\[
\zeta^{-1}(\mathcal{F}_{\mathcal{P}(S)}) = \zeta^{-1} \left( \sigma \left[ \bigcup_{g \in \Psi_b(S)} f_g^{-1}(\mathcal{B}_{\mathbb{R}}) \right] \right)
\]

\[
= \sigma \left[ \bigcup_{g \in \Psi_b(S)} \zeta^{-1}(f_g^{-1}(\mathcal{B}_{\mathbb{R}})) \right] = \sigma \left[ \bigcup_{g \in \Psi_b(S)} \xi_g^{-1}(\mathcal{B}_{\mathbb{R}}) \right] \subset \mathcal{B}_S.
\]

We conclude that \( \zeta \) is measurable. Since the first equality in the last display uses the characterization of \( \mathcal{F}_{\mathcal{P}(S)} \) as a minimal \( \sigma \)-field, this calculation illuminates the difficulty that arises when \( \mathcal{T}_{\mathcal{P}(S)} \) is used in place of \( \mathcal{F}_{\mathcal{P}(S)} \).

We can now state the hypotheses on the function \( h \) appearing in the Laplace limit. We assume that \( h \) is bounded and continuous with respect to the \( \tau \)-topology and that \( h \) is measurable on \( (\mathcal{P}(S), \mathcal{F}_{\mathcal{P}(S)}) \). Thus \( h(L^n) \) defines a random variable. An example of a function satisfying these two hypotheses can be obtained as follows. For each \( i \in \mathbb{N} \) let \( h_i \) denote a measurable function with respect to the Borel \( \sigma \)-field \( \mathcal{B}_{\mathcal{P}(S)} \) corresponding to the weak topology and assume that \( h(\mu) = \lim_{i \to \infty} h_i(\mu) \) exists for all \( \mu \in \mathcal{P}(S) \) and that \( h \) is bounded and continuous with respect to the \( \tau \)-topology.
Let \( \{ \theta_n, n \in \mathbb{N} \} \) and \( \theta \) be probability measures on \( \mathcal{S} \). Because \( \mathcal{P}(\mathcal{S}) \) is separable with respect to the weak topology, there exists a sequence \( \{ f_j, j \in \mathbb{N} \} \) of bounded continuous functions with the following property. If \( \int_S f_j d\theta_n \to \int_S f_j d\theta \) for each \( j \in \mathbb{N} \), then \( \theta_n \to \theta \). The fact that weak convergence can be reduced to checking a countable collection of convergences in \( \mathcal{R} \) is quite convenient. In particular, the countability guarantees that null sets do not cause a problem. An aspect of the nonseparability of \( \mathcal{P}(\mathcal{S}) \) with respect to the \( \tau \)-topology is the absence of an analogous criterion for checking \( \theta_n \to \theta \).

However, in the proof of the Laplace limit in the \( \tau \)-topology this aspect of the nonseparability of \( \mathcal{P}(\mathcal{S}) \) is really no problem at all. To prove such a Laplace limit, obviously one must show that w.p.1 \( \tilde{L}^n \to \tilde{L} \), which implies that \( h(\tilde{L}^n) \to h(\tilde{L}) \). In fact, one can prove this limit by using weak topology limits together with some general measure-theoretic approximations. The key is to exploit the existence of a reference measure \( \mu^* \), which is the unique invariant measure of \( p(x, dy) \). As we will see in the proof of Lemma 9.3.3, the probability-1 convergence \( \tilde{L}^n \to \tilde{L} \) can be reduced to a convergence analysis on the separable space \( L^1(\mathcal{S}, \mu^*) \).

The Laplace principle for the empirical measures of a Markov chain in the weak topology is proved in Theorem 8.4.3. We will prove the analogous result in the \( \tau \)-topology under the following hypothesis, which includes the conditions of Theorem 8.4.3 together with an additional uniformity assumption. We recall from Proposition 8.3.4 and Lemma 8.6.2 that under the first three of these conditions \( p(x, dy) \) has a unique invariant measure \( \mu^* \). In applications the uniformity assumption that we make seems to be more easily satisfied than other uniformity assumptions made in this context [11, 23].

**Condition 9.3.1.** We assume Conditions 8.2.2, 8.3.1, 8.4.1, and 8.4.2. We also assume that if a sequence of Borel subsets \( \{ A_j, j \in \mathbb{N} \} \) of \( \mathcal{S} \) satisfies \( A_j \downarrow A \) as \( j \to \infty \), where \( A \) satisfies \( \mu^*(A) = 0 \), then \( \lim_{j \to \infty} p(x, A_j) = 0 \) uniformly for \( x \) in compact subsets of \( \mathcal{S} \).

Suppose that the number \( \ell_0 \) in Condition 8.4.1 equals 1. Part (b) of Lemma 8.6.2 proves that if \( \mu^*(A) = 0 \), then for all \( x \in \mathcal{S} \) \( p(x, A) = 0 \) and thus \( \lim_{j \to \infty} p(x, A_j) = 0 \) if \( A_j \downarrow A \). Condition 9.3.1 somewhat strengthens this implication by adding the uniformity.

Although \( \mathcal{P}(\mathcal{S}) \) equipped with the \( \tau \)-topology is neither a metrizable space nor a separable one, it is a Hausdorff topological space, and it certainly makes sense to consider a Laplace upper bound and lower bound in this context. We now state these bounds for the empirical measures in the \( \tau \)-topology. They hold with the same function \( I \) as in the corresponding Laplace principle bounds in Chapter 8 [Theorem 8.4.3]. We also prove that \( I \) has compact level sets in the \( \tau \)-topology.

**Theorem 9.3.2.** Let \( \{ X_j, j \in \mathbb{N}_0 \} \) be a Markov chain that takes values in the Polish space \( \mathcal{S} \) and has stationary transition probabilities. We assume that the transition probability function \( p(x, dy) \) of the Markov chain satisfies Condition 9.3.1. For \( \mu \in \mathcal{P}(\mathcal{S}) \), we define

\[
I(\mu) \doteq \inf_{\{ q \in \mathcal{T} \mid q = \mu \}} \int_{\mathcal{S}} R(q(x, \cdot)\| p(x, \cdot) ) \mu(dx),
\]

where \( \mathcal{T} \) denotes the set of all transition probability functions on \( \mathcal{S} \). Let \( h \) be any function that is bounded and continuous with respect to the \( \tau \)-topology and is measurable on \( (\mathcal{P}(\mathcal{S}), \mathcal{F}_{\mathcal{P}(\mathcal{S})}) \). The following conclusions hold.
(a) \( I \) has compact level sets in the \( \tau \)-topology.

(b) For each compact subset \( K \) of \( S \) we have the uniform upper bound

\[
\limsup_{n \to \infty} \sup_{x \in K} \frac{1}{n} \log E_x \{ \exp[-n h(L^n)] \} \leq - \inf_{\mu \in \mathcal{P}(S)} \{ I(\mu) + h(\mu) \}.
\]

(c) For each compact subset \( K \) of \( S \) we have the uniform lower bound

\[
\liminf_{n \to \infty} \inf_{x \in K} \frac{1}{n} \log E_x \{ \exp[-n h(L^n)] \} \geq - \inf_{\mu \in \mathcal{P}(S)} \{ I(\mu) + h(\mu) \}.
\]

In order to prove the Laplace bounds, we introduce

\[
W^n(x) \doteq - \frac{1}{n} \log E_x \{ \exp[-n h(L^n)] \}
\]

and use the same representation formula as in Chapter 8; namely,

\[
W^n(x) = V^n(x) \doteq \inf_{\{\nu_j^n\}} E_x \left\{ - \frac{1}{n} \sum_{j=0}^{n-1} R(\nu_j^n(\cdot|\tilde{X}_j^n, \tilde{L}_j^n))||p(\tilde{X}_j^n, \cdot) + h(\tilde{L}_j^n) \right\}.
\]

Aside from an approximation argument, the proofs are essentially the same as the proofs of the Laplace principle bounds in the weak topology. The only difference now is that \( h \) is not continuous with respect to this topology. The main new ingredient will be to show that w.p.1 \( \tilde{L}^n \to \tilde{L} \) implies \( h(\tilde{L}^n) \to h(\tilde{L}) \) when the sequence of running costs is bounded. This implication is a consequence of Lemma 9.3.3.

As in the proof of the Laplace principle bounds in the weak topology, it will be necessary to consider convergent sequences of admissible control measures and controlled empirical measures, together with their limits, when sequence of running costs is bounded. The following lemma is the key to extending the proofs of the Laplace principle bounds to the \( \tau \)-topology. The proof is adapted from [13]. Given an admissible control sequence \( \{\nu_j^n, j = 0, 1, \ldots, n - 1\} \), we recall the admissible control measure

\[
\nu^n(dx \times dy) \doteq \frac{1}{n} \sum_{j=0}^{n-1} \delta_{\tilde{X}_j^n}(dx) \times \nu_j^n(dy|\tilde{X}_j^n, \tilde{L}_j^n).
\]

The first marginal \((\nu^n)_1\) of \(\nu^n\) equals the controlled empirical measure

\[
\tilde{L}^n \doteq \frac{1}{n} \sum_{j=0}^{n-1} \delta_{\tilde{X}_j^n}.
\]

**Lemma 9.3.3.** We assume Condition 9.3.1. Let \((\bar{\Omega}, \bar{\mathcal{F}}, \bar{P}_x)\) be a probability space and let \(\nu\) and \(\bar{L}\) be respectively a stochastic kernel on \(S \times S\) given \(\bar{\Omega}\) and a stochastic kernel on \(S\) given \(\bar{\Omega}\). Consider a sequence of admissible control measures \(\{\nu^n, n \in \mathbb{N}\}\) and controlled empirical measures \(\{\tilde{L}^n, n \in \mathbb{N}\}\) defined on this probability space such that w.p.1 \(\limsup_{n \to \infty} R(\nu^n||\tilde{L}^n \otimes p) < \infty\), \(\nu^n \Rightarrow \nu\), \(\tilde{L}^n \Rightarrow \tilde{L}\), and \((\nu^n)_2 \Rightarrow \bar{L}\). Then there exists a subsequence of \(n \in \mathbb{N}\) such that w.p.1 \(\tilde{L}^n \Rightarrow \tilde{L}\) and \((\nu^n)_2 \Rightarrow \bar{L}\).
We first show how the lemma yields the Laplace bounds in the $\tau$-topology. The proof of the lemma is deferred until the end of the section as is the proof that $I$ has compact level sets in the $\tau$-topology [Proposition 9.3.5]. For simplicity we will prove that

$$\liminf_{n \to \infty} W^n(x) \geq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{ I(\mu) + h(\mu) \}$$

and

$$\limsup_{n \to \infty} W^n(x) \leq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{ I(\mu) + h(\mu) \}$$

for each starting point $x \in \mathcal{S}$. The analogous limits that are uniform with respect to $x$ in compact sets are obtained as in Sections 8.3 and 8.6.

**Proof of Laplace upper bound.** We prove

$$\liminf_{n \to \infty} W^n(x) \geq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{ I(\mu) + h(\mu) \}$$

as in Section 8.3, where the analogous weak topology upper bound is proved. Let $\varepsilon > 0$ be given as well as any subsequence of $n \in \mathbb{N}$ along which the uniformly bounded functions $W^n(x)$ converge. For each $n$ consider an admissible control sequence $\{\nu^n\}$ satisfying

$$V^n(x) + \varepsilon \geq E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R\left(\nu^n_j(\cdot|X^n_j, \bar{L}^n_j)\|p(X^n_j, \cdot)\right) + h(\bar{L}^n) \right\}$$

and denote by $\nu^n$ and $\bar{L}^n$ the corresponding admissible control measure and controlled empirical measure. Since $|W^n(x)| \leq \|h\|_\infty$, Theorem 8.2.8 can be applied. There exists a subsequence of $\{(\nu^n, \bar{L}^n)\}$ which converges in distribution to a pair of stochastic kernels $(\nu, \bar{L})$, where w.p.1 $\nu_1 = \nu_2 = \bar{L}$; thus $(\nu^n)_2 \overset{\mathcal{D}}{\rightarrow} \bar{L}$. By Corollary C.3.3 and Jensen’s Inequality

$$\frac{1}{n} \sum_{j=0}^{n-1} R\left(\nu^n_j(\cdot|X^n_j, \bar{L}^n_j)\|p(X^n_j, \cdot)\right)$$

$$= \frac{1}{n} \sum_{j=0}^{n-1} R\left(\delta_{X^n_j}(\cdot) \times \nu^n_j(\cdot|X^n_j, \bar{L}^n_j)\|\delta_{X^n_j}(\cdot) \times p(X^n_j, \cdot)\right) \geq R(\nu^n\|\bar{L}^n \otimes p).$$

This yields

$$\infty > \sup_{n \in \mathbb{N}} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R\left(\nu^n_j(\cdot|X^n_j, \bar{L}^n_j)\|p(X^n_j, \cdot)\right) \right\} \geq \sup_{n \in \mathbb{N}} E_x \{ R(\nu^n\|\bar{L}^n \otimes p) \},$$

which in turn implies the tightness of the nonnegative subsequence $\{ R(\nu^n\|\bar{L}^n \otimes p) \}$. It follows that there exists a random variable $Y$ such that along a subsequence

$$(\nu^n, \bar{L}^n, (\nu^n)_2, R(\nu^n\|\bar{L}^n \otimes p)) \overset{\mathcal{D}}{\rightarrow} (\nu, \bar{L}, Y).$$
We now apply the Skorohod Representation Theorem. By Fatou's Lemma \( \mathbb{E}_x \{ Y \} < \infty \) and so \( Y \) is finite w.p.1. We conclude that along some subsequence of \( n \in \mathbb{N} \) the hypotheses of Lemma 9.3.3 are satisfied.

Lemma 9.3.3 implies that along this subsequence \( \tilde{L}^n \to \tilde{L} \) w.p.1. Since \( p(x, dy) \) satisfies the Feller property [Condition 8.3.1], Lemma 8.3.2 implies that \( \tilde{L}^n \otimes p \Rightarrow \tilde{L} \otimes p \) w.p.1. By the lower semicontinuity of \( R(\| \cdot \|) \), the equalities \( \nu_1 = \nu_2 = \tilde{L} \), and part (b) of Lemma 8.5.1, it follows that w.p.1

\[
\liminf_{n \to \infty} R(\nu^n \| \tilde{L}^n \otimes p) \geq R(\nu \| \tilde{L} \otimes p) \geq I(\tilde{L}). 
\]

The boundedness and continuity of \( h \) with respect to the \( \tau \)-topology now yield

\[
\liminf_{n \to \infty} W^n(x) + \varepsilon = \liminf_{n \to \infty} V^n(x) + \varepsilon \\
\geq \liminf_{n \to \infty} \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu_j^\text{n}(\cdot|X_j^\text{n}, \tilde{L}_j^\text{n}) \| p(X_j^\text{n}, \cdot)) + h(\tilde{L}^n) \right\} \\
\geq \liminf_{n \to \infty} \mathbb{E}_x \{ R(\nu^n \| \tilde{L}^n \otimes p) + h(\tilde{L}^n) \} \\
\geq \mathbb{E}_x \{ I(\tilde{L}) + h(\tilde{L}) \} \\
\geq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{ I(\mu) + h(\mu) \}.
\]

If we send \( \varepsilon \to 0 \), then the usual argument by contradiction establishes the desired lower limit

\[
\liminf_{n \to \infty} W^n(x) \geq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{ I(\mu) + h(\mu) \}
\]

for the entire sequence \( \{W^n(x), n \in \mathbb{N}\} \).

**Proof of the Laplace lower bound.** In order to prove that

\[
\limsup_{n \to \infty} W^n(x) \leq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{ I(\mu) + h(\mu) \},
\]

it suffices to show that any subsequence of \( \{W^n(x), n \in \mathbb{N}\} \) has a subsequence satisfying this upper limit. We will work with a fixed subsequence for the remainder of the proof, indexing it by \( n \in \mathbb{N} \). Before proceeding, we recall the proof of the analogous result in Section 8.6, where \( h \) was assumed to be continuous with respect to the weak topology on \( \mathcal{P}(\mathcal{S}) \). There the lower bound was first proved for initial conditions \( x \) in the set \( \Phi \) by applying the ergodic theorem and was subsequently extended to all initial conditions \( x \in \mathcal{S} \). Corollary 1.2.5 considerably simplified the extension by allowing us to assume that \( h \) is Lipschitz continuous. However, because this corollary requires that the space in which the random variables take values be a Polish space, it cannot be applied in the present section.

Hence in order to prove the Laplace lower bound here, we will need a method that avoids the Lipschitz continuity of \( h \). Before doing this, it is worthwhile to note that the proof of the lower bound

\[
\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{E}_x \{ \exp[-n h(L^n)] \} \geq - \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{ I(\mu) + h(\mu) \}
\]
given in Section 8.6 for \( x \in \Phi \) carries over without change to the functions \( h \) that we consider in this section; i.e., functions that are bounded and continuous with respect to the \( \tau \)-topology and are measurable on \((\mathcal{P}(\mathcal{S}), \mathcal{F}_{\mathcal{P}(\mathcal{S})})\). Thus the only significant difficulty will occur in the extension to all initial conditions \( x \in \mathcal{S} \). While the technique that we will use here is also valid when \( h \) is bounded and continuous with respect to the weak topology, we prefer the somewhat easier argument of Section 8.6 for that case.

To simplify the present proof, we will assume that \( \ell_0 = 1 \) in Condition 8.4.1. As noted in (8.38), this implies that \( p(x, \Phi) = 1 \) for all \( x \in \mathcal{S} \). The proof when \( \ell_0 \neq 1 \) requires only simple notational modifications.

We start by proving a slight extension of the lower bound proved in Section 8.6. Let \( h \) be bounded and continuous with respect to the \( \tau \)-topology and measurable on \((\mathcal{P}(\mathcal{S}), \mathcal{F}_{\mathcal{P}(\mathcal{S})})\). We assume that we are given a sequence \( \{h_n\} \) of measurable functions on \((\mathcal{P}(\mathcal{S}), \mathcal{F}_{\mathcal{P}(\mathcal{S})})\) which satisfy \( h_n(\theta_n) \to h(\theta) \) whenever \( \theta_n \xrightarrow{\tau} \theta \) and which for some \( M < \infty \) satisfy \( \|h_n\|_{\infty} \leq M \) for all \( n \). The function

\[
W^n(x) = -\frac{1}{n} \log E_x\{\exp[-n h_n(L^n)]\}
\]

has the same representation formula as \( W^n(x) \) except that \( h \) is replaced by \( h_n \) [Theorem 8.2.1]. We claim that there exists a Borel subset \( \Phi \) of \( \mathcal{S} \) such that for all \( x \in \Phi \)

\[
\limsup_{n \to \infty} W^n(x) \leq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}.
\]

The proof is as follows. Let \( \varepsilon > 0 \) be given. Following the pattern of proof of the Laplace principle lower bound in Section 8.6, we choose \( \gamma \in \mathcal{P}(\mathcal{S}) \) such that

\[
I(\gamma) + h(\gamma) \leq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\} + \varepsilon < \infty.
\]

Since convergence in total variation norm implies convergence in the \( \tau \)-topology, Lemma 8.6.3 yields the existence of \( \nu^* \in \mathcal{P}(\mathcal{S}) \) and a transition probability function \( q^*(x, dy) \) with the following properties: \( \nu^* \) is an invariant measure of \( q^*(x, dy) \), the Markov chain with initial distribution \( \nu^* \) and transition probability function \( q^*(x, dy) \) is ergodic,

\[
h(\nu^*) \leq h(\gamma) + \varepsilon, \quad \int_\mathcal{S} R(q^*(x, \cdot)||p(x, \cdot)) \nu^*(dx) \leq I(\gamma) < \infty.
\]

For each \( j \in \{0, 1, \ldots, n-1\} \) and \( (x, \mu) \in \mathcal{S} \times \mathcal{M}_{j+n}(\mathcal{S}) \) we define \( \nu^n_j(dy|x, \mu) \equiv q^*(x, dy) \). Then the admissible control measure \( \nu^n \) has the form

\[
\nu^n(dx \times dy) = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{\tilde{X}_j^n}(dx) \otimes q^*(\tilde{X}_j^n, dy) = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{\tilde{X}_j^n}(dx) \otimes q^*(x, dy) = \tilde{L}^n(dx) \otimes q^*(x, dy).
\]

As in Section 8.6, the \( L^1 \) and pointwise ergodic theorems, Chebyshev’s Inequality, the Borel-Cantelli Lemma, and a separability argument imply that there exist a Borel set \( \Phi \) satisfying \( \nu^*(\Phi) = 1 \) and a subsequence of \( n \in \mathbb{N} \) such that whenever \( \tilde{X}_{n_0}^n = x \) lies in \( \Phi \) we have

\[
\lim_{n \to \infty} E_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(q^*(\tilde{X}_j^n, \cdot)||p(\tilde{X}_j^n, \cdot)) - \int_\mathcal{S} R(q^*(x, \cdot)||p(x, \cdot)) \nu^*(dx) \right\} = 0.
\]
together with the probability-1 limits
\[
\lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} R(q^*(\bar{X}_j^n, \cdot)||p(\bar{X}_j^n, \cdot)) = \int \mathcal{S} R(q^*(x, \cdot)||p(x, \cdot)) \nu^*(dx) < \infty,
\]
\[
\nu^n(dx \times dy) \Rightarrow \nu^*(dx) \otimes q^*(x, dy), \bar{L}^n \Rightarrow \nu^*, \text{ and } (\nu^n)_2 \Rightarrow \nu^*.
\]
Since for each \(\omega\)
\[
\frac{1}{n} \sum_{j=0}^{n-1} R(q^*(\bar{X}_j^n, \cdot)||p(\bar{X}_j^n, \cdot)) \geq R(\nu^n||\bar{L}^n \otimes p),
\]
we have verified the hypotheses of Lemma 9.3.3 whenever \(\bar{X}_0^n = x\) lies in \(\Phi\). The lemma implies that for such \(x\) \(\bar{L}^n \Rightarrow \nu^*\) w.p.1. When combined with the properties that \(h\) is bounded and continuous with respect to the \(\tau\)-topology, \(h_n(\theta_n) \to h(\theta)\) whenever \(\theta_n \to \theta\), and \(\|h_n\|_{\infty} \leq M\) for all \(n\), the Lebesgue Dominated Convergence Theorem implies that
\[
\lim_{n \to \infty} \bar{E}_x\{h_n(\bar{L}^n)\} = \bar{E}_x\{h(\nu^*)\}.
\]
It then follows exactly as in Section 8.6 that whenever \(\bar{X}_0^n = x\) lies in \(\Phi\),
\[
\limsup_{n \to \infty} \bar{W}^n(x) \leq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\},
\]
An examination of the proof shows that \(\Phi\) is independent of \(h\) and the sequence \(\{h_n\}\).

Let \(x\) and \(y\) be arbitrary points in \(\mathcal{S}\). To complete the proof of the Laplace lower bound, we consider for \(n \geq 2\)
\[
W^n(x, y) \doteq -\frac{1}{n} \log E_{x,y}\{\exp[-n h(L^n)]\},
\]
where \(E_{x,y}\) denotes expectation conditioned on \(X_0 = x\) and \(X_1 = y\). For \(\theta \in \mathcal{P}(\mathcal{S})\) we define
\[
\bar{L}^{n-1} \doteq \frac{1}{n-1} \sum_{j=1}^{n-1} \delta_{X_j} \text{ and } h_{n,x}(\theta) \doteq \frac{n}{n-1} h\left(\frac{1}{n} \delta_x + \frac{n-1}{n} \theta\right).
\]
The distribution of \(\bar{L}^{n-1}\) conditioned on \(X_0 = x\) and \(X_1 = y\) is the same as the distribution of \(L^{n-1}\) given \(X_0 = y\). Since
\[
L^n = \frac{1}{n} \delta_x + \frac{n-1}{n} \bar{L}^{n-1},
\]
we have
\[
W^n(x, y) = \frac{n-1}{n} \bar{W}^{n-1}_x(y),
\]
where
\[
\bar{W}^{n-1}_x(y) \doteq -\frac{1}{n-1} \log E_y\{\exp[-(n-1) h_{n,x}(L^{n-1})]\}.
\]
Since for each \(x \in \mathcal{S}\) the sequence \(\{h_{n,x}\}\) satisfies the conditions on \(\{h_n\}\) imposed in the previous paragraph, it follows that for each \(x \in \mathcal{S}\) and \(y \in \Phi\)
\[
\limsup_{n \to \infty} W^n(x, y) = \limsup_{n \to \infty} \bar{W}^{n-1}_x(y) \leq \inf_{\mu \in \mathcal{P}(\mathcal{S})} \{I(\mu) + h(\mu)\}.
\]
9.3. EMPIRICAL MEASURES IN THE $\tau$–TOPOLOGY

We recall that since $\ell_0 = 1$, $p(x, \Phi) = 1$ for all $x \in S$.

The proof will be concluded by relating this limit involving $W^n(x, y)$ to the desired limit involving $\tilde{W}^n(x)$. Let $\varepsilon > 0$ and $x \in S$ be given. As in the argument in Section 8.6 leading up to equation (8.40), there exists a Borel subset $\tilde{\Phi} \subseteq \Phi$ and $N \in \mathbb{N}$ such that $p(x, \tilde{\Phi}) > 0$ and for all $y \in \tilde{\Phi}$ and $n \geq N$

$$E_{x, y} \{\exp[-n h(L^n)]\} \geq \exp \left[ -n \left( \inf_{\mu \in \mathcal{P}(S)} \{I(\mu) + h(\mu)\} + \varepsilon \right) \right].$$

This implies that

$$E_x \{\exp[-n h(L^n)]\} = \int_S E_{x, y} \{\exp[-n h(L^n)]\} p(x, dy) \geq \int_{\tilde{\Phi}} E_{x, y} \{\exp[-n h(L^n)]\} p(x, dy) \geq p(x, \tilde{\Phi}) \exp \left[ -n \left( \inf_{\mu \in \mathcal{P}(S)} \{I(\mu) + h(\mu)\} + \varepsilon \right) \right],$$

and so we obtain

$$\limsup_{n \to \infty} W^n(x) \leq \inf_{\mu \in \mathcal{P}(S)} \{I(\mu) + h(\mu)\} + \varepsilon.$$ 

Sending $\varepsilon \to 0$, we see that for each $x \in S$ every subsequence of the original sequence $\{W^n(x), n \in \mathbb{N}\}$ has a subsubsequence satisfying this upper limit, which therefore must be satisfied by the entire sequence. □

We now turn to the proof of Lemma 9.3.3. A key tool is an approximation argument based on the following observation.

**Lemma 9.3.4.** Let $\theta$ be a probability measure on $S$. Then there exists a countable set $A \subseteq C_b(S)$ with the property that, for any bounded measurable function $g$ on $S$, there is a sequence $\{f_j, j \in \mathbb{N}\}$ in $A$ and $\Lambda < \infty$ such that $f_j \to g \theta$–a.s. and

$$\sup_{j \in \mathbb{N}} \|f_j\|_\infty \vee \|g\|_\infty \leq \Lambda.$$ 

**Comments on the Proof.** The separable Banach space $L^1(S, \theta)$ has a countable dense set of functions in $C_b(S)$. By an elementary argument one can produce from this countable dense subset the set $A$ having the properties in the lemma. □

**Proof of Lemma 9.3.3.** Under the hypotheses of the lemma there exists a $\tilde{P}_x$–null set $N_0$ such that for $\omega \notin N_0 \limsup_{n \to \infty} R(\nu^n \| \tilde{L}^n \otimes p) < \infty$, $\nu^n \Rightarrow \nu$, $\tilde{L}^n \Rightarrow \tilde{L}$, and $(\nu^n)_2 \Rightarrow \tilde{L}$. For $\omega \notin N_0$, since $\tilde{L}^n = (\nu^n)_1$, it follows that $\nu_1 = \nu_2 = \tilde{L}$. Our aim is to find another $\tilde{P}_x$–null set $N_1$ and a subsequence of $n \in \mathbb{N}$ and to prove the following: for all $\omega \notin N_0 \cup N_1$ and any bounded measurable function $g$

$$\lim_{n \to \infty} \int_S g \ d\tilde{L}^n = \int_S g \ d\tilde{L}$$.
and for all $\omega \not\in N_0$ and any bounded measurable function $g$
\[ \lim_{n \to \infty} \int_S g \, d(\nu^n)_2 = \int_S g \, d\bar{L}. \]

We first prove the latter limit. This will be carried out by using a sequence $\{f_j, j \in \mathbb{N}\} \subset \mathcal{A}$ of bounded continuous functions having the properties given in Lemma 9.3.4 with $\theta$ equal to the unique invariant measure $\mu^*$ of $p(x, dy)$ [Corollary 8.3.4, Lemma 8.6.2 (a)]. By hypothesis on $R(\nu^n || \bar{L}^n \otimes p)$ and formula (9.8), for some $M < \infty$
\[ M \geq \limsup_{n \to \infty} R(\nu^n || \bar{L}^n \otimes p) \geq \liminf_{n \to \infty} R(\nu^n || \bar{L}^n \otimes p) \geq R(\nu || \bar{L} \otimes p) \geq I(\bar{L}). \] (9.9)

We proved in part (c) of Lemma 8.6.2 that under Condition 8.4.1 $I(\mu) < \infty$ implies that $\mu \ll \mu^*$. Thus the last display implies that $\bar{L} \ll \mu^*$. Since $\lim_{j \to \infty} f_j = g$ $\mu^*$-a.s., it follows that $\lim_{j \to \infty} f_j = g$ $\bar{L}$-a.s. and hence, by the Lebesgue Dominated Convergence Theorem, that
\[ \lim_{j \to \infty} \int_S f_j \, d\bar{L} = \int_S g \, d\bar{L}. \]

For each $j \in \mathbb{N}$ $\int_S f_j \, d(\nu^n)_2 \rightarrow \int_S f_j \, d\bar{L}$ since by assumption $(\nu^n)_2 \Rightarrow \bar{L}$. Hence it remains only to verify that
\[ \limsup_{n \to \infty} \limsup_{j \to \infty} \int_S |f_j - g| \, d(\nu^n)_2 = 0. \] (9.10)

Given $\epsilon > 0$ we have
\[ \int_S |f_j - g| \, d(\nu^n)_2 = \int_{\{|f_j - g| > \epsilon\}} |f_j - g| \, d(\nu^n)_2 + \int_{\{|f_j - g| \leq \epsilon\}} |f_j - g| \, d(\nu^n)_2 \leq \int_{\{|f_j - g| > \epsilon\}} |f_j - g| \, d(\nu^n)_2 + \epsilon \leq 2\lambda \cdot (\nu^n)_2\{|f_j - g| > \epsilon\} + \epsilon, \]

and so (9.10) will be proved if we can show that
\[ \limsup_{j \to \infty} \limsup_{n \to \infty} (\nu^n)_2\{|f_j - g| > \epsilon\} = 0. \]

Since $(\nu^n)_1 = \bar{L}^n$, there exists a transition probability function $q^n(x, dy)$ such that $\nu^n = \bar{L}^n \otimes q^n$ [Lemma 8.5.1 (a)]. For sufficiently large $n$ the bound
\[ \infty > R(\nu^n || \bar{L}^n \otimes p) = R(\bar{L}^n \otimes q^n || \bar{L}^n \otimes p) = \int_S R(q^n(x, \cdot) || p(x, \cdot)) \, \bar{L}^n(dx) \]
implies that $q^n(x, \cdot) \ll p(x, \cdot) \bar{L}^n$-a.s. It follows that for all sufficiently large $n$ there exists a measurable version of the Radon–Nikodym derivative $\frac{dq^n(x, \cdot)}{dp(x, \cdot)}(y)$ on $\mathcal{S} \times \mathcal{S}$ [Theorem A.5.7] and that for arbitrary $c > 0$
\[ (\nu^n)_2\{|f_j - g| > \epsilon\} = \int_{\mathcal{S} \times \{|f_j - g| > \epsilon\}} \frac{dq^n(x, \cdot)}{dp(x, \cdot)}(y) \, p(x, dy) \, \bar{L}^n(dx) \]
\[ = \int_{\{(x, y) : \frac{dq^n(x, \cdot)}{dp(x, \cdot)}(y) \leq c\} \cap \{|f_j - g| > \epsilon\}} \frac{dq^n(x, \cdot)}{dp(x, \cdot)}(y) \, p(x, dy) \, \bar{L}^n(dx) \]
\[ + \int_{\{(x, y) : \frac{dq^n(x, \cdot)}{dp(x, \cdot)}(y) > c\} \cap \{|f_j - g| > \epsilon\}} \frac{dq^n(x, \cdot)}{dp(x, \cdot)}(y) \, p(x, dy) \, \bar{L}^n(dx). \]
9.3. EMPIRICAL MEASURES IN THE $\tau$–TOPOLOGY

The first term in the last expression is bounded above by

$$c \int_S p(x, \{ |f_j - g| > \varepsilon \}) \bar{L}^n(dx).$$

Since $\bar{L}^n \rightarrow \bar{L}$, the sequence $\{ \bar{L}^n \}$ is tight [Theorem A.3.15]. Hence the $\mu^*$–a.s. convergence of $f_j$ to $g$ and the uniformity hypothesis in Condition 9.3.1 guarantee that this first term converges to 0 as $n \rightarrow \infty$ and then $j \rightarrow \infty$.

For the second term we will use the inequality

$$ab \leq e^{\sigma a} + \frac{1}{\sigma} (b \log b - b + 1),$$

which is valid for $a \geq 0$, $b \geq 0$, and $\sigma \geq 1$ [see (1.18)]. Taking $a = 1$ and $b = r^n(x, y) \pm \frac{dn}(dp(x,y))(y)$ and using the fact that $r^n \log r^n - r^n + 1 \geq 0$, we obtain for all sufficiently large $n$

$$\int_{(S \times \{|f_j - g| > \varepsilon \}) \cap \{r^n > c\}} r^n(x, y) p(x, dy) \bar{L}^n(dx)$$

$$\leq \int_{(S \times \{|f_j - g| > \varepsilon \}) \cap \{r^n > c\}} e^{\sigma} p(x, dy) \bar{L}^n(dx)$$

$$+ \frac{1}{\sigma} \int_{(S \times \{|f_j - g| > \varepsilon \}) \cap \{r^n > c\}} (r^n(x, y) \log r^n(x, y) - r^n(x, y) + 1) p(x, dy) \bar{L}^n(dx)$$

$$\leq e^{\sigma} (\bar{L}^n \otimes p)\{(x, y) \in S^2 : r^n(x, y) > c\}$$

$$+ \frac{1}{\sigma} \int_S R(q^n(x, \cdot) \| p(x, \cdot)) \bar{L}^n(dx) - \frac{1}{\sigma} + \frac{1}{\sigma}$$

$$= e^{\sigma} (\bar{L}^n \otimes p)\{(x, y) \in S^2 : r^n(x, y) > c\} + \frac{1}{\sigma} R(\bar{L}^n \otimes q^n \| \bar{L}^n \otimes p)$$

$$= e^{\sigma} (\bar{L}^n \otimes p)\{(x, y) \in S^2 : r^n(x, y) > c\} + \frac{1}{\sigma} R(\nu^n \| \bar{L}^n \otimes p).$$

The last line is a consequence of part (f) of Lemma 1.4.3 and the equality $\nu^n = \bar{L}^n \otimes q^n$. Chebyshev’s Inequality now yields

$$\limsup_{j \rightarrow \infty} \limsup_{n \rightarrow \infty} \int_{(S \times \{|f_j - g| > \varepsilon \}) \cap \{r^n > c\}} r^n(x, y) p(x, dy) \bar{L}^n(dx)$$

$$\leq \limsup_{n \rightarrow \infty} \left( \frac{e^{\sigma}}{c} + \frac{1}{\sigma} R(\nu^n \| \bar{L}^n \otimes p) \right) \leq \frac{e^{\sigma}}{c} + \frac{M}{\sigma}.$$ 

Taking $c \rightarrow \infty$ and then $\sigma \rightarrow \infty$, we have shown that

$$\limsup_{j \rightarrow \infty} \limsup_{n \rightarrow \infty} (\nu^n)_2 \{|f_j - g| > \varepsilon \} = 0.$$ 

This establishes (9.10) and completes the proof that

$$\lim_{n \rightarrow \infty} \int_S g \ d(\nu^n)_2 = \int_S g \ d\bar{L}.$$
for all $\omega \notin N_0$ and any bounded measurable function $g$.

We finish the proof of the lemma by finding a $\bar{\mathcal{P}}_\varepsilon$-null set $N_1$ and a subsequence of $n \in \mathbb{N}$ with the property that

$$
\lim_{n \to \infty} \int_S g \, d\bar{L}^n = \int_S g \, d\bar{L}
$$

for all $\omega \notin N_0 \cup N_1$ and any bounded measurable function $g$. Since $\mathcal{S}$ is a Polish space, there exists a countable collection of Borel sets $\{A_i, i \in \mathbb{N}\}$ that generate the Borel $\sigma$-field of $\mathcal{S}$. Let $\mathcal{G}$ denote the countable field generated by $\{A_i\}$. As a first step we will prove (9.11) for any $g$ of the form $1_A$ and all $\omega \notin N_0 \cup N(A)$, where $A \in \mathcal{G}$ and $N(A)$ is a $\bar{\mathcal{P}}_\varepsilon$-null set. To do this, we appeal to Lemma 8.2.7, which shows that for any $\varepsilon > 0$

$$
\bar{\mathcal{P}}_\varepsilon \left\{ \left( \int_S 1_A \, d\bar{L} - \int_S 1_A \, d\nu^2 \right) \geq \varepsilon \right\} \leq \frac{16}{n^2}.
$$

By the Borel–Cantelli Lemma and the first part of the proof, there exist a subsequence of $n \in \mathbb{N}$ which is independent of $A$ (e.g., $n = k^2$ for $k \in \mathbb{N}$) and a $\bar{\mathcal{P}}_\varepsilon$-null set $N(A)$ such that for all $\omega \notin N_0 \cup N(A)$

$$
\int_S 1_A \, d\bar{L} - \lim_{n \to \infty} \int_S 1_A \, d\bar{L}^n = \lim_{n \to \infty} \left( \int_S 1_A \, d\nu^2 - \int_S 1_A \, d\bar{L}^n \right) = 0.
$$

Defining $N_1 = \cup_{A \in \mathcal{G}} N(A)$, we obtain this limit for all $\omega \notin N_0 \cup N_1$ and any $A \in \mathcal{G}$.

We now prove (9.11) for all $\omega \notin N_0 \cup N_1$ and any bounded measurable function $g$. Let $\mathcal{D}$ denote the collection of Borel sets $E$ satisfying for all $\omega \notin N_0 \cup N_1$

$$
\lim_{n \to \infty} \int_S 1_E \, d\bar{L}^n = \int_S 1_E \, d\bar{L}
$$

when $n \to \infty$ along the subsequence chosen in the previous paragraph. One easily verifies that $\mathcal{D}$ is a Dynkin class. Since $\mathcal{G} \subset \mathcal{D}$, the Dynkin Class Theorem [Theorem A.2.1] implies that $\mathcal{D}$ equals the Borel $\sigma$-field of $\mathcal{S}$. In other words, the last display holds for all $\omega \notin N_0 \cup N_1$ and any Borel set $E$. Since for any bounded measurable function $g$ there exists a simple function $f$ such that $\|g - f\|_\infty$ is as small as desired, we have proved (9.11) for all $\omega \notin N_0 \cup N_1$ and any bounded measurable function $g$. The proof of the lemma is complete. ■

We now prove that the function $I$ appearing in Theorem 9.3.2 has compact level sets in the $\tau$-topology. The proof is a deterministic version of the proof of the lemma just given.

**Proposition 9.3.5.** Under Condition 9.3.1 the function $I$ defined in Theorem 9.3.2 has compact level sets in the $\tau$-topology.

**Proof.** Given any $M < \infty$ and a sequence $\{\mu_n, n \in \mathbb{N}\}$ of probability measures on $\mathcal{S}$ satisfying $\sup_{n \in \mathbb{N}} I(\mu_n) \leq M$, we show that there exists a subsequence converging in the $\tau$-topology to some probability measure $\mu$ that satisfies $I(\mu) \leq M$. Since $I$ has compact
level sets in the weak topology [Proposition 8.5.2 (b)], there certainly exists a subsequence converging weakly to a probability measure \( \mu \) that satisfies \( I(\mu) \leq M \). We complete the proof by showing that there exists a subsequence of the weakly convergent subsequence of \( \{ \mu_n \} \) such that for any bounded measurable function \( g \)

\[
\lim_{n \to \infty} \int_S g \, d\mu_n = \int_S g \, d\mu.
\]

We fix such a \( g \) for the remainder of the proof.

For each \( n \) indexing this subsequence of \( \{ \mu_n \} \), there exists a probability measure \( \tau_n \) on \( S \times S \) such that \( (\tau_n)_1 = (\tau_n)_2 = \mu_n \) and \( I(\mu_n) = R(\tau_n \| \mu_n \otimes p) \) [Lemma 8.5.1 (c)]. Since \( \mu_n \Rightarrow \mu \), the sequences of both first and second marginals of \( \tau_n \) are tight and thus \( \{ \tau_n \} \) is tight. Hence there exists a subsequence converging weakly to some probability measure \( \tau \), which must satisfy \( \tau_1 = \tau_2 = \mu \). The formula

\[
M \geq \liminf_{n \to \infty} I(\mu_n) = \liminf_{n \to \infty} R(\tau_n \| \mu_n \otimes p) \geq R(\tau \| \mu \otimes p) \geq I(\mu)
\]

and part (c) of Lemma 8.6.2 imply that \( \mu \) is absolutely continuous with respect to the unique invariant measure \( \mu^* \) of \( p(x, dy) \) [Corollary 8.3.4, Lemma 8.6.2 (a)]. Now let \( \{ f_j, j \in \mathbb{N} \} \subset \mathcal{A} \) be a subset of bounded continuous functions having the properties given in Lemma 9.3.4 with \( \theta = \mu^* \). Since \( f_j \to g \) \( \mu^* \)-a.s., it follows that \( f_j \to g \) \( \mu \)-a.s.

The Lebesgue Dominated Convergence Theorem yields

\[
\lim_{n \to \infty} \int_S f_j \, d\mu_n = \int_S g \, d\mu,
\]

and since \( \mu_n \Rightarrow \mu \), for each \( j \in \mathbb{N} \) \( \int_S f_j \, d\mu_n \to \int_S f_j \, d\mu \). Hence it remains only to verify that

\[
\limsup_{j \to \infty} \limsup_{n \to \infty} \int_S |f_j - g| \, d\mu_n = 0,
\]

which as in the proof of Lemma 9.3.3 will follow if we can show that for arbitrary \( \varepsilon > 0 \)

\[
\limsup_{j \to \infty} \limsup_{n \to \infty} \mu_n \{|f_j - g| > \varepsilon\} = 0.
\]

(9.12)

Exactly as in the proof of this lemma, we obtain for arbitrary \( c > 0 \) and \( \sigma > 1 \) the bound

\[
\mu_n \{|f_j - g| > \varepsilon\} \leq c \int_S p(x, \{|f_j - g| > \varepsilon\}) \, d\mu_n(dx) + \frac{e^\sigma}{c} + \frac{1}{\sigma} R(\tau_n \| \mu_n \otimes p)
\]

\[
\leq c \int_S p(x, \{|f_j - g| > \varepsilon\}) \, d\mu_n(dx) + \frac{e^\sigma}{c} + \frac{M}{\sigma}.
\]

Since \( \mu_n \Rightarrow \mu \), \( \{ \mu_n \} \) is tight. Hence the \( \mu^* \)-a.s. convergence of \( f_j \) to \( g \) and Condition 9.3.1 imply that

\[
\limsup_{j \to \infty} \limsup_{n \to \infty} \int_S p(x, \{|f_j - g| > \varepsilon\}) \, d\mu_n(dx) = 0.
\]

Taking \( c \to \infty \) and then \( \sigma \to \infty \) completes the proof of (9.12) and thus the proof of the lemma. \( \blacksquare \)
We end this section by pointing out a related property of the relative entropy; namely, for any probability measure \( \theta \) on \( \mathcal{S} \), the relative entropy \( R(\cdot \| \theta) \) has compact level sets in the \( \tau \)-topology. We omit the proof, which is carried out by specializing the proof of the previous proposition to the case \( p(x, dy) = \mu^*(dy) = \theta(dy) \). For details see Lemma 2.7 in [13]. Another proof is given, for example, in Lemma 2.3 in [56]. This property of the relative entropy is needed in the proof of Proposition 4.5.1.

**Proposition 9.3.6.** Let \( \mathcal{S} \) be a Polish space. Then for each \( \theta \in \mathcal{P}(\mathcal{S}) \) and each \( M < \infty \) the set \( \{ \gamma \in \mathcal{P}(\mathcal{S}) : R(\gamma \| \theta) \leq M \} \) is a compact subset of \( \mathcal{P}(\mathcal{S}) \) in the \( \tau \)-topology.

With this proposition we have completed our extensions of the Laplace principle for the empirical measures of a Markov chain. Continuous-time extensions of the results in Chapter 6 will be given in the next chapter.
Chapter 10

Laplace Principle for Continuous–Time Markov Processes with Continuous Statistics

10.1 Introduction

In this final chapter we consider the continuous–time version of the random walk model analyzed in Chapter 6. We will extend to a class of continuous–time Markov processes including diffusions and jump processes the results already obtained for this discrete–time model. The basic idea is to work with sampled versions of the continuous–time processes, a procedure that will allow us to apply the Laplace principle stated in Theorem 6.7.5. When combined with an estimate showing that the continuous–time processes and the sampled processes are superexponentially close, this procedure will yield the Laplace principle for the continuous–time processes. It is interesting to note that the Laplace principle in this chapter does not require weak sense uniqueness of the processes under consideration. The main result is stated in Theorem 10.2.6 and is proved in Section 10.3. An alternative approach, applied to a class of diffusions in [13], is to derive and use representation formulas for the continuous–time processes themselves. A proof of the appropriate representation formula was sketched in Subsection 4.6.4.

10.2 Statement of the Laplace Principle

We will deal with families of random variables that satisfy Laplace limits with the normalizing sequence \( n \in \mathbb{N} \) replaced by a continuous parameter. We start by defining the concept of a Laplace principle in this setting. Such a Laplace principle is equivalent to a large deviation principle with the same rate function. One can prove this as in the discrete parameter setting treated in Section 1.2.

**Definition 10.2.1.** Let \( \{ Y_\varepsilon, \varepsilon > 0 \} \) be a family of random variables indexed by \( \varepsilon > 0 \) and taking values in a Polish space \( \mathcal{X} \) and let \( I \) be a rate function on \( \mathcal{X} \). The family
\( \{Y^\varepsilon, \varepsilon > 0\} \) is said to satisfy the Laplace principle on \( \mathcal{X} \) with rate function \( I \) if for all bounded continuous functions \( h \) mapping \( \mathcal{X} \) into \( \mathbb{R} \)

\[
\lim_{\varepsilon \to 0} \varepsilon \log E \left\{ \exp \left[ -\frac{1}{\varepsilon} h(Y^\varepsilon) \right] \right\} = -\inf_{x \in \mathcal{X}} \{ h(x) + I(x) \}.
\]

The continuous-time processes to be considered in this chapter are defined in terms of the following quantities:

1. A bounded, measurable, \( \mathbb{R}^d \)-valued function \( b(x) = (b_1(x), ..., b_d(x)) \) of \( x \in \mathbb{R}^d \).

2. A bounded, measurable, \( d \times d \) matrix-valued function \( a(x) = \{a_{ij}(x), i, j = 1, ..., d\} \) that is symmetric and nonnegative definite for each \( x \in \mathbb{R}^d \).

3. A family of measures \( \{\nu_x, x \in \mathbb{R}^d\} \) such that \( \nu_x(A) \) is a measurable function of \( x \in \mathbb{R}^d \) for each Borel subset \( A \) of \( \mathbb{R}^d \) and for each \( x \in \mathbb{R}^d \)

\[
\nu_x(\{0\}) = 0, \nu_x(\mathbb{R}^d) < \infty, \text{ and } \int_{\mathbb{R}^d} ||y||^2 \nu_x(dy) < \infty.
\]

For \( \varepsilon > 0 \), we define an operator \( \mathcal{L}^\varepsilon \) on twice continuously differentiable functions with compact support by

\[
\mathcal{L}^\varepsilon f(x) = \sum_{i=1}^d b_i(x) \frac{\partial f(x)}{\partial x_i} + \frac{\varepsilon}{2} \sum_{i,j=1}^d a_{ij}(x) \frac{\partial^2 f(x)}{\partial x_i \partial x_j} + \frac{1}{\varepsilon} \int_{\mathbb{R}^d} \left( f(x + \varepsilon y) - f(x) - \varepsilon \sum_{i=1}^d y_i \frac{\partial f(x)}{\partial x_i} \right) \nu_x(dy).
\] (10.1)

We would like to study the large deviation properties of Markov processes associated with such generators \( \mathcal{L}^\varepsilon \) in the limit as \( \varepsilon \to 0 \). Let \( \mathcal{D}([0,1] : \mathbb{R}^d) \) be the space of functions that map \([0,1]\) into \( \mathbb{R}^d \), are right continuous, and have left hand limits. \( \mathcal{D}([0,1] : \mathbb{R}^d) \) is metrized by the Skorohod metric, with respect to which it is complete and separable [Theorem A.6.5 (a)]. The pioneering work of Stroock and Varadhan [86] has shown that a Markov process associated with \( \mathcal{L}^\varepsilon \) can be characterized in terms of a martingale property. More precisely, under a broad class of additional conditions on \( b(x), a(x), \) and \( \nu_x \), there exists for each \( \varepsilon > 0 \) and \( x \in \mathbb{R}^d \) a Markov process \( Y^\varepsilon = \{Y^\varepsilon(t), t \in [0,1]\} \) that satisfies \( Y^\varepsilon(0) = x \), has sample paths in \( \mathcal{D}([0,1] : \mathbb{R}^d) \), and solves the martingale problem for \( \mathcal{L}^\varepsilon \).

Since a full explanation of this concept would involve numerous technicalities, instead we refer the reader to one of the excellent expositions of the topic contained in [50], [61], or [86]. For \( t \in [0,1] \) let \( \mathcal{F}^t \) denote a \( \sigma \)-field with respect to which the family \( \{Y^\varepsilon(s), 0 \leq s \leq t\} \) is measurable and let \( \mathcal{C}_0^\infty(\mathbb{R}^d) \) denote the set of \( C^\infty \) functions mapping \( \mathbb{R}^d \) into \( \mathbb{R} \) with compact support. One consequence of the martingale formulation to be used below is that for each \( f \in \mathcal{C}_0^\infty(\mathbb{R}^d) \) and \( s \in [0,1] \)

\[
f(Y^\varepsilon(t)) - \int_s^t (\mathcal{L}^\varepsilon f)(Y^\varepsilon(u)) \, du
\] (10.2)

is an \( \mathcal{F}^s \)-martingale for all \( t \in [s,1] \).
10.2. STATEMENT OF THE LAPLACE PRINCIPLE

The Markov processes $Y^\varepsilon$ that we consider include jump processes and diffusions. For example, if $a(x)$ is identically zero, then $Y^\varepsilon$ is a jump Markov process with local jump measures $\nu_x$. In order to describe the motion of the process, for simplicity let us suppose that for each $x \in \mathbb{R}^d$ $\nu_x(\mathbb{R}^d) > 0$. Then over any finite time interval $Y^\varepsilon$ has w.p.1 finitely many jumps. If $Y^\varepsilon(t) = x$, then up to terms of order $\Delta t$ as $\Delta t \to 0$, over the time interval $(t, t + \Delta t)$ $Y^\varepsilon$ jumps with probability $\varepsilon^{-1} \nu_x(\mathbb{R}^d) \Delta t$ and the value of the jump has distribution $[\nu_x(\mathbb{R}^d)]^{-1} \nu_x(\varepsilon^{-1} dy)$. Between jumps $Y^\varepsilon$ follows the trajectory of the dynamical system $\dot{x}(t) = b(x(t))$, where

$$b(x) = b(x) - \int_{\mathbb{R}^d} y \, \nu_x(dy).$$

On the other hand, if $\nu_x$ equals 0 for each $x \in \mathbb{R}^d$, then $\mathcal{L}^\varepsilon$ becomes a second order differential operator, and the associated Markov process $Y^\varepsilon$ is a diffusion with local drift vector $b(x)$ and scaled local diffusion matrix $\varepsilon a(x)$. In the general case where neither $a(x)$ nor $\nu_x$ is 0, the motion of $Y^\varepsilon$ has both a jump component and a diffusive component. Extensions to include other classes of processes such as reflected diffusions are also possible.

We return to the general development, assuming that for each $\varepsilon > 0$ and $x \in \mathbb{R}^d$ there exists a Markov process $Y^\varepsilon \equiv \{Y^\varepsilon(t), t \in [0, 1]\}$ that satisfies $Y^\varepsilon(0) = x$, has sample paths in $\mathcal{D}([0, 1] : \mathbb{R}^d)$, and solves the martingale problem for $\mathcal{L}^\varepsilon$. We denote by $P_x$ the induced measure on $\mathcal{D}([0, 1] : \mathbb{R}^d)$ and by $E_x$ the corresponding expectation operator.

In the Laplace principle analysis of $Y^\varepsilon$ a key role is played by the function of $x$ and $\alpha$ in $\mathbb{R}^d$ given by

$$H(x, \alpha) \equiv \langle b(x), \alpha \rangle + \frac{1}{2} \langle \alpha, a(x) \alpha \rangle + \int_{\mathbb{R}^d} \left[ \exp(\langle \alpha, y \rangle) - 1 - \langle \alpha, y \rangle \right] \nu_x(dy). \quad (10.3)$$

In fact, $H(x, \alpha)$ will play the same role that the cumulant generating function defined in equation (6.3) played in the analysis of the random walk model in Chapter 6. We impose the following condition on $H(x, \alpha)$, which is the analogue of Condition 6.2.1. The term "continuous statistics" in the title of the present chapter reflects the continuity of $H$ assumed in part (b) of Condition 10.2.2.

**Condition 10.2.2.**

(a) For each $\alpha \in \mathbb{R}^d$, $\sup_{x \in \mathbb{R}^d} H(x, \alpha) < \infty$.

(b) $H(x, \alpha)$ is a continuous function of $(x, \alpha) \in \mathbb{R}^d \times \mathbb{R}^d$.

Under Condition 10.2.2, part (d) of Proposition 10.3.2 shows that there exists a stochastic kernel $\mu(dy|x)$ on $\mathbb{R}^d$ given $\mathbb{R}^d$ such that for each $x \in \mathbb{R}^d$ $H(x, \alpha)$ is the cumulant generating function of the probability measure $\mu(\cdot|x)$; i.e., for all $\alpha \in \mathbb{R}^d$

$$H(x, \alpha) = \log \int_{\mathbb{R}^d} \exp(\langle \alpha, y \rangle) \mu(dy|x).$$

In addition, the function mapping $x \in \mathbb{R}^d \mapsto \mu(\cdot|x) \in \mathcal{P}(\mathbb{R}^d)$ is continuous in the topology of weak convergence on $\mathcal{P}(\mathbb{R}^d)$. Thus $H(x, \alpha)$ and $\mu(dy|x)$ satisfy Condition
6.2.1. As we will see in the proof of Lemma 10.2.3, one may identify \( \mu(dy|x) \) as the distribution at time 1 of a Markov process \( \{Y(t), t \in [0,1]\} \) satisfying \( Y(0) = 0 \) and having constant drift vector \( b(x) \), constant diffusion matrix \( a(x) \), and constant jump measure \( \nu_x \).

For \( x \) and \( \beta \) in \( \mathbb{R}^d \) we define the Legendre–Fenchel transform
\[
L(x, \beta) \doteq \sup_{\alpha \in \mathbb{R}^d} [\langle \alpha, \beta \rangle - H(x, \alpha)] .
\]
In the present chapter, this function will play the same role in the definition of the rate function as the Legendre–Fenchel transform defined in equation (6.4) played in the definition of the rate function in Chapter 6. In order to state and prove the Laplace principle lower bound in Chapter 6, we needed to characterize the relative interior of the effective domain of \( L(x, \cdot) \) in terms of the problem data, which in that chapter was the stochastic kernel \( \mu(dy|x) \). This characterization was carried out in part (d) of Lemma 6.2.3. We next give the analogous characterization in the continuous–time setting. This will lead to Condition 10.2.4, which is needed to prove the Laplace principle lower bound for the family \( \{Y^\varepsilon, \varepsilon > 0\} \). The reader may skip the following somewhat technical discussion and jump directly to the paragraph after Condition 10.2.4, where we point out two important cases in which the condition is satisfied. The sets \( T_x \) appearing in the condition are defined in equation (10.5).

Equations (10.1) and (10.3) separate out the drift term \( \langle b(x), \alpha \rangle \) from the local martingale terms. For the purposes of characterizing the relative interior of the effective domain of \( L(x, \cdot) \), it is more convenient to rewrite \( H(x, \alpha) \) in the form
\[
H(x, \alpha) = \langle \tilde{b}(x), \alpha \rangle + \frac{1}{2} \langle \alpha, a(x) \alpha \rangle + \int_{\mathbb{R}^d} (\exp\langle \alpha, y \rangle - 1) \nu_x(\,dy),
\]
where
\[
\tilde{b}(x) \doteq b(x) - \int_{\mathbb{R}^d} y \nu_x(\,dy).
\]
For \( A \) and \( B \) subsets of \( \mathbb{R}^d \), define
\[
A + B \doteq \{x \in \mathbb{R}^d : x = y + z, y \in A, z \in B\}.
\]
A subset \( C \) of \( \mathbb{R}^d \) is called a convex cone if it has the property that \( y \in C \) implies \( \lambda y \in C \) for all \( \lambda \in [0, \infty) \). For \( x \in \mathbb{R}^d \) we define
\[
Q_x \doteq \{w \in \mathbb{R}^d : w = a(x)u \text{ for some } u \in \mathbb{R}^d\} + S_{\nu_x},
\]
where \( S_{\nu_x} \) denotes the support of \( \nu_x \). Let \( \text{con} \, Q_x \) be the smallest convex cone that contains \( Q_x \) and define
\[
T_x = \{\tilde{b}(x)\} + \text{con} \, Q_x.
\]
Note that if \( \nu_x = 0 \), then
\[
T_x = \{w \in \mathbb{R}^d : w = b(x) + a(x)u \text{ for some } u \in \mathbb{R}^d \}.
\]
Also, if \( a(x) = 0 \) and \( b(x) = 0 \), then
\[
\tilde{b}(x) = -\int_{\mathbb{R}^d} y \nu_x(\,dy) \quad \text{and} \quad T_x = \{\tilde{b}(x)\} + \text{con} \, S_{\nu_x}.
\]
Lemma 10.2.3. For $x$ and $\beta$ in $\mathbb{R}^d$ we define $L(x, \beta)$ by equation (10.4) and $T_x$ by equation (10.5). Then the relative interior of the effective domain of $L(x, \cdot)$ equals the relative interior of $T_x$; in symbols, $\text{ri}(\text{dom } L(x, \cdot)) = \text{ri } T_x$.

**Proof.** We first consider the special case where $\nu_x = 0$ for all $x \in \mathbb{R}^d$. In this case the conclusion of the lemma can be verified by explicit calculation. Indeed

$$L(x, \beta) = \sup_{\alpha \in \mathbb{R}^d} \left\{ \langle \alpha, \beta - b(x) \rangle - \frac{1}{2} \langle \alpha, a(x) \alpha \rangle \right\},$$

which is the rate function associated via Cramér’s Theorem with a Gaussian probability measure on $\mathbb{R}^d$ with mean vector $b(x)$ and covariance matrix $a(x)$. Explicitly, if $\beta - b(x)$ lies in the range of $a(x)$, then

$$L(x, \beta) = \frac{1}{2} \langle \alpha, \beta - b(x) \rangle,$$

where $\alpha \in \mathbb{R}^d$ is any vector satisfying $a(x) \alpha = \beta - b(x)$. One easily checks that $L(x, \beta)$ in the last display is well defined. On the other hand, if $\beta - b(x)$ does not lie in the range of $a(x)$, then $L(x, \beta)$ equals $\infty$. It follows that in this special case

$$\text{dom } L(x, \cdot) = \left\{ \beta \in \mathbb{R}^d : \beta - b(x) = a(x) u \text{ for some } u \in \mathbb{R}^d \right\} = T_x.$$

Hence $\text{ri}(\text{dom } L(x, \cdot)) = \text{ri } T_x$, as claimed.

We now consider the case where $\nu_x \neq 0$ for at least one $x \in \mathbb{R}^d$. As proved in part (d) of Proposition 10.3.2, there exists a stochastic kernel $\mu(dy|x)$ on $\mathbb{R}^d$ given $\mathbb{R}^d$ such that for each $x$ and $\alpha$ in $\mathbb{R}^d$

$$H(x, \alpha) = \log \int_{\mathbb{R}^d} \exp \langle \alpha, y \rangle \mu(dy|x).$$

For fixed $x \in \mathbb{R}^d$ let $Z_1$ be a Gaussian random vector with cumulant generating function

$$\langle \tilde{b}(x), \alpha \rangle + \frac{1}{2} \langle \alpha, a(x) \alpha \rangle.$$

In addition, let $Z_2$ be a random vector independent of $Z_1$ with cumulant generating function

$$\int_{\mathbb{R}^d} \left( \exp \langle \alpha, y \rangle - 1 \right) \nu_x(dy)$$

and take a point $y$ in the support of $\nu_x$. Using the interpretation of $Z_2$ as the position of a jump process at time 1 and the bound $\nu_x(\mathbb{R}^d) < \infty$, we see that for all $\delta > 0$ and all $j \in \mathbb{N}$

$$P\{\|Z_2 - jy\| < j\delta \} > 0 \text{ and } P\{\|Z_2\| < \delta \} > 0.$$

Since $H(x, \alpha)$ is the cumulant generating function of $Z_1 + Z_2$, it follows that $\mu(dy|x)$ is the distribution of $Z_1 + Z_2$ [Theorem A.3.21]. Hence for each $x \in \mathbb{R}^d$ the convex hull of the support of $\mu(\cdot|x)$ equals $T_x$, and so the relative interiors of these sets are equal. The conclusion of the lemma now follows since by part (d) of Lemma 6.2.3 $\text{ri}(\text{dom } L(x, \cdot))$ equals $\text{ri}(\text{conv } S_{\mu(\cdot|x)})$.

Having characterized the relative interior of the effective domain of $L(x, \cdot)$, we now adopt the following counterpart of Condition 6.7.4.
Chapter 10. Continuous-Time Markov Processes

Condition 10.2.4.
(a) The sets $\text{int } T_x$ are independent of $x \in \mathbb{R}^d$.
(b) $0 \in \Sigma \equiv \text{int } T_x$.

Condition 10.2.4 is automatically satisfied if $T_x = \mathbb{R}^d$ for all $x \in \mathbb{R}^d$. This is certainly the case either if $a(x)$ is positive definite for all $x \in \mathbb{R}^d$ or if the support of $\nu_x$ contains 0 in its interior for all $x \in \mathbb{R}^d$.

In Chapter 6 we introduced another condition, Condition 6.3.2, which allowed us to prove the Laplace principle lower bound when the analogues of the sets $\text{int } T_x$ depended on $x \in \mathbb{R}^d$. In this case, Condition 6.3.1 did not hold. We next state the counterpart of Condition 6.3.2 in the continuous-time setting. This is particularly important since it allows the treatment of degenerate jump diffusions. A set of techniques for verifying Condition 10.2.5 can be obtained from Proposition 6.3.4.

Condition 10.2.5. Let any compact subset $\Delta$ of $\mathbb{R}^d$ and any number $\varepsilon \in (0, 1)$ be given. There are numbers $\eta = \eta(\Delta, \varepsilon) \in (0, 1)$ and $K = K(\Delta, \varepsilon) \in (0, \infty)$ so that whenever $\xi \in \Delta$, $v \in \Delta$, and $\gamma$ in $\mathbb{R}^d$ satisfy $\|\xi - v\| \leq \eta$, there is a $\tilde{\beta} \in \mathbb{R}^d$ such that

$$L(\xi, \tilde{\beta}) - L(v, \gamma) \leq \varepsilon [1 + L(v, \gamma)]$$

and

$$\|\tilde{\beta} - \gamma\| \leq K \|\xi - v\| [1 + L(v, \gamma)].$$

We now state the Laplace principle for the family $\{Y^\varepsilon, \varepsilon > 0\}$, assuming Condition 10.2.2 and, for the purposes of the lower bound, either Condition 10.2.4 or Condition 10.2.5.

Theorem 10.2.6. We assume Condition 10.2.2 and either Condition 10.2.4 or Condition 10.2.5. For each $\varepsilon > 0$ and each initial condition $x \in \mathbb{R}^d$ we assume that there exists a Markov process $Y^\varepsilon$ that satisfies $Y^\varepsilon(0) = x$, has sample paths in $\mathcal{D}([0, 1] : \mathbb{R}^d)$, and solves the martingale problem for the generator $\mathcal{L}^\varepsilon$ given in equation (10.1). For absolutely continuous functions $\varphi \in \mathcal{D}([0, 1] : \mathbb{R}^d)$ satisfying $\varphi(0) = x$, we define

$$I_x(\varphi) = \int_0^1 L(\varphi(t), \dot{\varphi}(t)) \, dt,$$

where $L$ is defined in equation (10.4). For all other $\varphi \in \mathcal{D}([0, 1] : \mathbb{R}^d)$, we set $I_x(\varphi) \equiv \infty$.

Then the family $\{Y^\varepsilon(t), \varepsilon > 0\}$ satisfies the Laplace principle on $\mathcal{D}([0, 1] : \mathbb{R}^d)$ with rate function $I_x$. In fact, the Laplace principle holds uniformly on compacts.

An important special case of this theorem is Schilder’s Theorem, which was mentioned in Example 1.1.3. In Schilder’s Theorem $\mathcal{L}^\varepsilon$ is given by

$$\mathcal{L}^\varepsilon f(x) = \frac{\varepsilon}{2} \sum_{i=1}^d \frac{\partial^2 f(x)}{\partial x_i^2}.$$
and $Y^\varepsilon$ equals $\sqrt{\varepsilon} w$, where $w$ denotes a standard Brownian motion taking values in $\mathbb{R}^d$. In this case $L(x, \beta) = \frac{1}{2} \| \beta \|^2$ for each $x$ and $\beta$ in $\mathbb{R}^d$ and

$$ I_x(\varphi) = \frac{1}{2} \int_0^1 \| \varphi(t) \|^2 \, dt $$

for all absolutely continuous functions $\varphi \in \mathcal{D}([0, 1] : \mathbb{R}^d)$ satisfying $\varphi(0) = 0$.

The next section is devoted to the proof of Theorem 10.2.6. We will focus on the nonuniform Laplace principle. The uniform version can be proved with minor modifications.

## 10.3 Proof of the Laplace Principle

We first show that the function $I_x$ defined in Theorem 10.2.6 has compact level sets in $\mathcal{D}([0, 1] : \mathbb{R}^d)$. This function has the same form as the rate function $I_x$ in the Laplace principle stated in Theorem 6.3.3 for the random walk model of Chapter 6. It was proved in Proposition 6.2.4 that $I_x$ has compact level sets in $C([0, 1] : \mathbb{R}^d)$. The proof relied on the three properties of $L(x, \beta)$ summarized in Remark 6.2.5. These, in turn, are consequences of the convexity of $H(x, \cdot)$ for each $x \in \mathbb{R}^d$ and the fact that $H(x, \alpha)$ is a continuous function of $(x, \alpha) \in \mathbb{R}^d \times \mathbb{R}^d$. The latter two properties of $H(x, \alpha)$ are valid in the present setting because of Condition 10.2.2. It follows that the function $I_x$ defined in Theorem 10.2.6 has compact level sets in $C([0, 1] : \mathbb{R}^d)$. Since $C([0, 1] : \mathbb{R}^d)$ is a closed subset of $\mathcal{D}([0, 1] : \mathbb{R}^d)$ [Theorem A.6.5 (d)] and

$$ I_x(\varphi) = \infty \text{ for } \varphi \in \mathcal{D}([0, 1] : \mathbb{R}^d) \setminus C([0, 1] : \mathbb{R}^d), $$

we conclude that $I_x$ has compact level sets in $\mathcal{D}([0, 1] : \mathbb{R}^d)$. Since for any $\varphi \in \mathcal{D}([0, 1] : \mathbb{R}^d)$ $I_x(\varphi) \in [0, \infty]$, it follows that $I_x$ is a rate function on $\mathcal{D}([0, 1] : \mathbb{R}^d)$.

Let $h$ be any bounded continuous function mapping $\mathcal{D}([0, 1] : \mathbb{R}^d)$ into $\mathbb{R}$. In order to complete the proof of Theorem 10.2.6, it suffices to prove the Laplace limit

$$ \lim_{\varepsilon \to 0} \varepsilon \log E_x \left\{ \exp \left[ -\frac{1}{\varepsilon} h(Y^\varepsilon) \right] \right\} = \inf_{\varphi \in \mathcal{D}([0, 1] : \mathbb{R}^d)} \{ h(\varphi) + I_x(\varphi) \} $$

when $\varepsilon \to 0$ along any sequence $\{ \varepsilon_n, n \in \mathbb{N} \}$ in $(0, 1)$ converging to 0. Fixing such a sequence, we will prove this limit by applying the Laplace principle stated in Theorem 6.7.5. Specifically, by sampling the continuous–time Markov processes $Y^\varepsilon$ at a sequence of times depending on $\varepsilon_n$, we will define a sequence of piecewise linear processes $\{ Z^n, n \in \mathbb{N} \}$ to which Theorem 6.7.5 can be applied. Then we will show that as $n \to \infty$ this sequence is superexponentially close to $\{ Y^\varepsilon, n \in \mathbb{N} \}$. This estimate will allow us to deduce the Laplace principle for $\{ Y^\varepsilon \}$ with normalizing sequence $\{ 1/\varepsilon_n \}$. In this way we will derive the Laplace principle for the original family $\{ Y^\varepsilon, \varepsilon > 0 \}$.

For each $n \in \mathbb{N}$ let $c_n = [1/\varepsilon_n]$, which denotes the largest integer less than or equal to $1/\varepsilon_n$. We then define the process $Z^n = \{ Z^n(t), t \in [0, 1] \}$ by

$$ Z^n(t) = Y^\varepsilon(j/c_n) + c_n \left( t - \frac{j}{c_n} \right) \left[ Y^\varepsilon((j + 1)/c_n) - Y^\varepsilon(j/c_n) \right] \quad (10.6) $$
for $t \in [j/c_n, (j + 1)/c_n]$, $j = 0, 1, ..., c_n - 1$. This process is the piecewise linear interpolation of the sampled sequence $\{Y^\varepsilon_n(j/c_n), j = 0, 1, ..., c_n\}$, which is a Markov chain with stationary transition probabilities. $Z^n$ will be equal in distribution to the process $X^n = \{X^n(t), t \in [0, 1]\}$ defined in equations (6.68) and (6.69) and studied in Section 6.7 if in the definition of $X^n$ we take $\{v_j^n(x), x \in \mathbb{R}^d, j \in \mathbb{N}_0\}$ to be an i.i.d. sequence of random vector fields having the common distribution

$$
\mu^n(dy|x) = P_x\{c_n[Y^\varepsilon_n(1/c_n) - x] \in dy\};
$$

(10.7)

$\mu^n(dy|x)$ is a stochastic kernel on $\mathbb{R}^d$ given $\mathbb{R}^d$. For $n \in \mathbb{N}$ and $x$ and $\alpha$ in $\mathbb{R}^d$, we define

$$
H^n(x, \alpha) = \log \int_{\mathbb{R}^d} \exp(\alpha, y) \mu^n(dy|x) = \log E_x\{\exp(\alpha, c_n[Y^\varepsilon_n(1/c_n) - x])\}.
$$

(10.8)

Part (b) of Proposition 10.3.2 shows that these functions are well defined.

We next apply Theorem 6.7.5 to derive the Laplace limit for the sequence $\{Z^n, n \in \mathbb{N}\}$. This Laplace limit is stated in the following proposition first with the normalizing sequence $\{c_n\}$ and then with the asymptotically equivalent normalizing sequence $\{1/\varepsilon_n\}$.

**Proposition 10.3.1.** We assume Condition 10.2.2 and either Condition 10.2.4 or Condition 10.2.5. Let $\{\varepsilon_n\}$ be any sequence in $(0, 1)$ converging to 0 and define $c_n \equiv [1/\varepsilon_n]$. For $n \in \mathbb{N}$ and $x \in \mathbb{R}^d$ define the piecewise linear process $Z^n$ by formula (10.6), where $Y^\varepsilon_n$ denotes the Markov process specified in Theorem 10.2.6. Then for any bounded continuous function $h$ mapping $C([0, 1] : \mathbb{R}^d)$ into $\mathbb{R}$, we have the Laplace limits

$$
\lim_{n \to \infty} \frac{1}{c_n} \log E_x\{\exp[-c_n h(Z^n)]\} = -\inf_{\varphi \in C([0, 1] : \mathbb{R}^d)} \{I_x(\varphi) + h(\varphi)\}
$$

and

$$
\lim_{n \to \infty} \varepsilon_n \log E_x\{\exp\left[-\frac{1}{\varepsilon_n} h(Z^n)\right]\} = -\inf_{\varphi \in C([0, 1] : \mathbb{R}^d)} \{I_x(\varphi) + h(\varphi)\},
$$

where $I_x(\varphi)$ is defined in Theorem 10.2.6.

**Proof.** We prove the first limit in the statement of the proposition by applying Theorem 6.7.5, for which the hypotheses must be checked. Parts (a) and (b) of Condition 6.7.2 are verified in parts (b) and (c) of the next proposition. Condition 10.2.2 coupled with part (d) of the next proposition verifies Condition 6.2.1 while Conditions 10.2.4 and 10.2.5 correspond, respectively, to Conditions 6.7.4 and 6.3.2. Thus the hypotheses of Theorem 6.7.5 are satisfied. Since

$$
\left| \log E_x\{\exp[-c_n h(Z^n)]\} - \log E_x\{\exp\left[-\frac{1}{\varepsilon_n} h(Z^n)\right]\} \right| \leq \left( \frac{1}{\varepsilon_n} - c_n \right) \|h\|_{\infty} \leq \|h\|_{\infty}
$$

and $\lim_{n \to \infty} c_n \varepsilon_n = 1$, the first limit in the statement of the proposition implies the second limit. 

The proof just given relies on the next proposition. A measurable function $f$ mapping $(0, \infty)$ into $\mathbb{R} \cup \{\infty\}$ is called **superlinear** if $\lim_{c \to \infty} f(c)/c = \infty$. 

10.3. PROOF OF THE LAPLACE PRINCIPLE

Proposition 10.3.2. We assume Condition 10.2.2. For $x$ and $\alpha$ in $\mathbb{R}^d$ and $n \in \mathbb{N}$, we define $H(x, \alpha)$ by (10.3) and $H^n(x, \alpha)$ by (10.8). The following conclusions hold.

(a) There exists a superlinear function $\zeta$ mapping $(0, \infty)$ into $\mathbb{R} \cup \{\infty\}$ such that for any $\varepsilon > 0$, $\delta > 0$, $s \in [0, 1]$, and $t \in (s, 1]$

$$\sup_{x \in \mathbb{R}^d} P_x \left\{ \sup_{s \leq \sigma \leq t} \|Y^\varepsilon(\sigma) - Y^\varepsilon(s)\| \geq \delta \right\} \leq 2d \exp \left( - \left[ \frac{(t-s)}{\varepsilon} \zeta \left( \delta / \sqrt{d(t-s)} \right) \right] \right).$$

(b) For each $\alpha \in \mathbb{R}^d$

$$\sup_{n \in \mathbb{N}} \sup_{x \in \mathbb{R}^d} H^n(x, \alpha) < \infty.$$

(c) For each $\alpha \in \mathbb{R}^d$ and each compact subset $K$ of $\mathbb{R}^d$

$$\lim_{n \to \infty} \sup_{x \in K} |H^n(x, \alpha) - H(x, \alpha)| = 0.$$

(d) For each $x \in \mathbb{R}^d$ the sequence of probability measures $\{\mu^n(dy|x), n \in \mathbb{N}\}$ defined in equation (10.7) converge weakly to a probability measure $\mu(dy|x)$ on $\mathbb{R}^d$, and for each $\alpha \in \mathbb{R}^d$

$$H(x, \alpha) = \log \int_{\mathbb{R}^d} \exp(\alpha, y) \mu(dy|x).$$

The family $\{\mu(dy|x), x \in \mathbb{R}^d\}$ defines a stochastic kernel on $\mathbb{R}^d$ given $\mathbb{R}^d$. In addition, the function mapping $x \in \mathbb{R}^d \mapsto \mu(\cdot|x) \in \mathcal{P}(\mathbb{R}^d)$ is continuous in the topology of weak convergence on $\mathcal{P}(\mathbb{R}^d)$.

The proof of the proposition requires a number of technical estimates that are based on an exponential martingale $N_\varepsilon(s; t)$ introduced in the next lemma. The use of such martingales in this context is standard since the work of Stroock and Varadhan [86]. The proposition is proved afterwards. $\tilde{N}_{\varepsilon, n}(t)$ is introduced because the function $H^n(x, \alpha)$ appearing in the proposition can be readily expressed in terms of it. We denote by $d(\cdot, \cdot)$ the Euclidean distance in $\mathbb{R}^d$ and by $B(x, \delta)$ the open ball in $\mathbb{R}^d$ with center $x$ and radius $\delta$.

Lemma 10.3.3. We assume Condition 10.2.2. For $\varepsilon > 0$, $\alpha \in \mathbb{R}^d$, $s \in [0, 1]$, $t \in [s, 1]$, $n \in \mathbb{N}$, $x \in \mathbb{R}^d$, and $\delta > 0$ we define

$$N_{\alpha, n}^\varepsilon(s; t) \equiv \exp \left( \frac{1}{\varepsilon} \left\{ \langle \alpha, Y^\varepsilon(t) - Y^\varepsilon(s) \rangle - \int_s^t H(Y^\varepsilon(u), \alpha) \, du \right\} \right),$$

$$\tilde{N}_{\alpha, n}(t) \equiv \exp \left( c_n \left\{ \langle \alpha, Y_{\varepsilon, n}(t) \rangle - \int_0^t H(Y_{\varepsilon, n}(u), \alpha) \, du \right\} \right),$$

and

$$\tau_{n, x, \delta} \equiv \inf \left\{ t \in [0, 1] : \inf_{u \in [0, t]} d(Y_{\varepsilon, n}(u), [B(x, \delta)]^c) = 0 \right\} \wedge 1.$$

The following conclusions hold.
(a) $N^\varepsilon_\alpha(s; t)$ is an $\mathcal{F}_t$–martingale for $t \in [s, 1]$. Hence

$$E_x\{N^\varepsilon_\alpha(s; t)\} = E_x\{N^\varepsilon_\alpha(s; s)\} = 1.$$ 

(b) $N^\varepsilon_\alpha(0; t \wedge \tau^n_{x, \delta})$ is an $\mathcal{F}_t^n$–martingale for $t \in [0, 1]$. Hence

$$E_x\{N^\varepsilon_\alpha(0; t \wedge \tau^n_{x, \delta})\} = E_x\{N^\varepsilon_\alpha(0; 0)\} = 1.$$

(c) For any $\alpha \in \mathbb{R}^d$

$$\sup_{n \in \mathbb{N}} \sup_{x \in \mathbb{R}^d} E_x\left\{ \tilde{N}^\varepsilon_{\alpha, x}(1/c_n) \right\} \leq 1 \quad \text{and} \quad \sup_{n \in \mathbb{N}} \sup_{x \in \mathbb{R}^d} E_x\left\{ \tilde{N}^\varepsilon_{\alpha, x}((1/c_n) \wedge \tau^n_{x, \delta}) \right\} \leq 1.$$

(d) There exists $\gamma_1 \in (0, \infty)$ such that for any $x$ and $\alpha \in \mathbb{R}^d$

$$H(x, \alpha) \geq \langle \alpha, b(x) \rangle \geq -\gamma_1 \|\alpha\|.$$

(e) For any $\alpha \in \mathbb{R}^d$ there exists $\gamma_2 \in \mathbb{R}$ such that

$$\sup_{n \in \mathbb{N}} \sup_{x \in \mathbb{R}^d} E_x\left\{ \tilde{N}^\varepsilon_{\alpha, x}(1/c_n)^2 \right\} \leq e^{\gamma_2} \quad \text{and} \quad \sup_{n \in \mathbb{N}} \sup_{x \in \mathbb{R}^d} E_x\left\{ \tilde{N}^\varepsilon_{\alpha, x}((1/c_n) \wedge \tau^n_{x, \delta})^2 \right\} \leq e^{\gamma_2}.$$

(f) For any $\alpha \in \mathbb{R}^d$ and compact subset $K$ of $\mathbb{R}^d$, there exists $M \in \mathbb{N}$ such that

$$\inf_{n \geq M} \inf_{x \in K} E_x\left\{ \tilde{N}^\varepsilon_{\alpha, x}((1/c_n) \wedge \tau^n_{x, \delta}) \right\} \geq \exp(-\gamma_3(\delta)),$$

where $\gamma_3(\delta) \to 0$ as $\delta \to 0$.

**Proof.** (a) Since $Y^\varepsilon$ solves the martingale problem for $\mathcal{L}^\varepsilon$, for each $f \in C^\infty_0(\mathbb{R}^d)$ and $s \in [0, 1]$

$$f(Y^\varepsilon(t)) - \int_s^t (\mathcal{L}^\varepsilon f)(Y^\varepsilon(u)) \, du$$

is an $\mathcal{F}_t$–martingale for all $t \in [s, 1]$. In order to prove part (a), we follow the proof of the equivalence of martingales given in parts (i) and (v) of Theorem 4.2.1 in [86]. All of the steps but one apply immediately in our setting. The only step that must be checked is the uniform integrability proved on page 88. In our settings this is a consequence of part (a) of Condition 10.2.2 and part (d) of the present lemma. An analogous calculation is given in the proof of part (e) of the present lemma.

(b) Since $Y^{\varepsilon_n}$ has sample paths in $D([0, 1] : \mathbb{R}^d)$ and $[B(x, \delta)]^\varepsilon$ is closed, part (a) of Proposition 2.1.5 in [50] implies that $\tau^n_{x, \delta}$ is an $\{\mathcal{F}_t^n\}$–stopping time. The Optional Sampling Theorem [50, Thm. 2.2.13] then guarantees that $N^{\varepsilon_n}(0; t \wedge \tau^n_{x, \delta})$ is an $\mathcal{F}_t^n$–martingale for $t \in [0, 1]$.

(c) With $Y^{\varepsilon_n}(0) = x$, defining $p_n = 1/c_n \varepsilon_n \in [1, \infty)$ gives $[\tilde{N}^{\varepsilon_n}(t)]^{p_n} = N^{\varepsilon_n}(0; t)$. Hence by Hölder’s Inequality and part (a)

$$\sup_{n \in \mathbb{N}} \sup_{x \in \mathbb{R}^d} E_x\{\tilde{N}^{\varepsilon_n}(t)\} \leq \sup_{n \in \mathbb{N}} \sup_{x \in \mathbb{R}^d} (E_x\{N^{\varepsilon_n}(0; t)\})^{1/p_n} = 1.$$
This yields the first inequality in part (c). The second inequality in part (c) follows similarly from part (b).

(d) Since \( a(x) \) is a nonnegative definite matrix and by convexity \( e^z - 1 - z \geq 0 \) for all \( z \in \mathbb{R} \), \( H(x, \alpha) \geq \langle b(x), \alpha \rangle \) for each \( x \) and \( \alpha \) in \( \mathbb{R}^d \). Since \( b(x) \) is bounded, there exists \( \gamma_1 \in (0, \infty) \) such that \( \langle b(x), \alpha \rangle \geq -\gamma_1 \| \alpha \| \).

(e) In order to prove the first inequality, we write \( [\tilde{N}_{\alpha, x}^n(1/c_n)]^2 \) in terms of \( \tilde{N}_{2\alpha, x}^n(1/c_n) \), obtaining

\[
[\tilde{N}_{\alpha, x}^n(1/c_n)]^2 = \tilde{N}_{2\alpha, x}^n(1/c_n) \exp \left( c_n \int_0^{1/c_n} H(Y^n(u), 2\alpha) \, du - 2c_n \int_0^{1/c_n} H(Y^n(u), \alpha) \, du \right) \leq \tilde{N}_{2\alpha, x}^n(1/c_n) \exp(\gamma_2),
\]

where

\[
\gamma_2 = \sup_{x \in \mathbb{R}^d} H(x, 2\alpha) - 2 \inf_{x \in \mathbb{R}^d} H(x, \alpha).
\]

Part (a) of Condition 10.2.2 and part (d) of the present lemma guarantee that \( \gamma_2 \) is finite, and the first inequality in part (c) yields

\[
\sup_n \sup_{x \in \mathbb{R}^d} E_x \left\{ \left[ \tilde{N}_{\alpha, x}^n(1/c_n) \right]^2 \right\} \leq \sup_n \sup_{x \in \mathbb{R}^d} E_x \left\{ \tilde{N}_{2\alpha, x}^n(1/c_n) \right\} \exp(\gamma_2) \leq \exp(\gamma_2).
\]

This gives the first inequality in part (e). The second inequality in part (e) is proved similarly using the second inequality in part (c).

(f) For \( n \in \mathbb{N} \) we define \( \bar{p}_n = c_n \varepsilon_n/(1 - \varepsilon_n) \in [1, \infty) \) and

\[
\Gamma_{x, \delta}^n = c_n \int_0^{(1/c_n) \wedge \tau_{x, \delta}^n} H(Y^n(u), \alpha) \, du - \frac{c_n}{1 - \varepsilon_n} \int_0^{(1/c_n) \wedge \tau_{x, \delta}^n} H(Y^n(u), (1 - \varepsilon_n)\alpha) \, du.
\]

For \( x \) in any compact subset \( K \) of \( \mathbb{R}^d \), part (b) and Hölder’s Inequality imply that

\[
1 = \left( E_x \left\{ \tilde{N}_{(1-\varepsilon_n)\alpha, x}^n(0; (1/c_n) \wedge \tau_{x, \delta}^n) \right\} \right)^{\bar{p}_n} \leq E_x \left\{ \tilde{N}_{(1-\varepsilon_n)\alpha, x}^n(0; (1/c_n) \wedge \tau_{x, \delta}^n) \right\} \right)^{\bar{p}_n} = E_x \left\{ \tilde{N}_{\alpha, x}^n(1/c_n \wedge \tau_{x, \delta}^n) \exp(\Gamma_{x, \delta}^n) \right\}.
\]

Since \( \varepsilon_n \to 0 \), there exists \( M \in \mathbb{N} \) such that \( \varepsilon_n \leq (1/2) \wedge \delta \) for all \( n \geq M \). Hence by the definition of \( \tau_{x, \delta}^n \)

\[
\sup_{n \geq M} \sup_{x \in K} \Gamma_{x, \delta}^n \leq \sup_{n \geq M} \sup_{x \in K} \sup_{y: \|y - x\| \leq \delta} \left| H(y, \alpha) - \frac{1}{1 - \varepsilon_n} H(y, (1 - \varepsilon_n)\alpha) \right| \leq \gamma_3(\delta),
\]

where \( \gamma_3(\delta) \) equals

\[
3 \sup_{x \in K} \sup_{y: \|y - x\| \leq \delta} \left| H(y, \beta) - H(x, \alpha) \right| \leq \delta, \|\beta - \alpha\| \leq \delta \right) + 2\delta \sup_{x \in K} |H(x, \alpha)|.
\]
Thus
\[
\inf \inf_{n \geq M} E_x \left\{ \tilde{N}^{\varepsilon}_n (\varepsilon_n (1/c_n) \wedge \tau^{n}_{x,d}) \right\} \geq \exp[-\gamma_3(\delta)].
\]

Part (a) of Condition 10.2.2 and part (d) of the present lemma imply that \(\sup_{x \in K} |H(x, \alpha)| < \infty\). Hence by the uniform continuity of \(H(\cdot, \cdot)\) on compact subsets of \(\mathbb{R}^d \times \mathbb{R}^d\) [Condition 10.2.2 (b)], \(\gamma_3(\delta) \to 0\) as \(\delta \to 0\). This completes the proof of part (f). 

The proof of Proposition 10.3.2 can now be carried out.

**Proof of Proposition 10.3.2.** (a) For \(i \in \{1, 2, \ldots, d\}\) let \(u^i\) denote the unit vector in the \(i\)’th coordinate direction of \(\mathbb{R}^d\) and let \(r\) be any positive real number. Then for any \(\varepsilon > 0, \delta > 0, s \in [0, 1], t \in (s, 1]\), and each choice of sign \(\pm\)

\[
\sup_{x \in \mathbb{R}^d} P_x \left\{ \sup_{s \leq \sigma \leq t} \left\langle \pm u^i, Y^\varepsilon(\sigma) - Y^\varepsilon(s) \right\rangle \geq \frac{\delta}{\sqrt{d}} \right\}
\]

\[
= \sup_{x \in \mathbb{R}^d} P_x \left\{ \sup_{s \leq \sigma \leq t} \left\langle \pm r u^i, Y^\varepsilon(\sigma) - Y^\varepsilon(s) \right\rangle \geq \frac{r \delta}{\sqrt{d}} \right\}
\]

\[
\leq \sup_{x \in \mathbb{R}^d} P_x \left\{ \sup_{s \leq \sigma \leq t} \exp \left[ \frac{1}{\varepsilon} \left( \left\langle \pm r u^i, Y^\varepsilon(\sigma) - Y^\varepsilon(s) \right\rangle - \int_s^\sigma H(Y^\varepsilon(\sigma), \pm r u^i) d\sigma \right) \right] \right\}
\]

\[
\geq \exp \left[ \frac{1}{\varepsilon} \left( \frac{r \delta}{\sqrt{d}} - (t - s) \tilde{H}(r) \right) \right],
\]

where

\(\tilde{H}(r) = \max_{\pm i = 1, 2, \ldots, d} \max_{x \in \mathbb{R}^d} H(x, \pm r u^i)\).

By part (a) of Condition 10.2.2 \(\tilde{H}(r)\) is finite for each \(r > 0\). For \(c \in (0, \infty)\) we define

\(\zeta(c) = \sup_{r > 0} \{rc - \tilde{H}(r)\}\).

Then for \(M \in (0, \infty)\) and all sufficiently large \(c\)

\[
\frac{1}{c} \zeta(c) \geq M - \frac{1}{c} \tilde{H}(M) \geq \frac{M}{2}.
\]

Since \(M\) is arbitrary, \(\zeta\) is superlinear. We rewrite (10.9) in terms of the exponential martingale \(\mathcal{N}^\varepsilon_{\pm ru^i}(s; \sigma)\) appearing in Lemma 10.3.3 (a) and use the submartingale inequality [50, Prop. 2.2.16] to obtain

\[
\sup_{x \in \mathbb{R}^d} P_x \left\{ \sup_{s \leq \sigma \leq t} \left\langle \pm u^i, Y^\varepsilon(\sigma) - Y^\varepsilon(s) \right\rangle \geq \frac{\delta}{\sqrt{d}} \right\}
\]

\[
\leq \sup_{x \in \mathbb{R}^d} P_x \left\{ \sup_{s \leq \sigma \leq t} \mathcal{N}^\varepsilon_{\pm ru^i}(s; \sigma) \geq \exp \left[ \frac{1}{\varepsilon} \left( \frac{r \delta}{\sqrt{d}} - (t - s) \tilde{H}(r) \right) \right] \right\}
\]

\[
\leq \exp \left[ \frac{1}{\varepsilon} \left( \frac{r \delta}{\sqrt{d}} - (t - s) \tilde{H}(r) \right) \right] \sup_{x \in \mathbb{R}^d} E_x \{ \mathcal{N}^\varepsilon_{\pm ru^i}(s; t) \}
\]

\[
= \exp \left[ \frac{1}{\varepsilon} \left( \frac{r \delta}{\sqrt{d}} - (t - s) \tilde{H}(r) \right) \right].
\]
10.3. PROOF OF THE LAPLACE PRINCIPLE

Since \( r \) is an arbitrary positive number, it follows that

\[
\sup_{x \in \mathbb{R}^d} P_x \left\{ \sup_{s \leq r \leq t} \langle \pm u^i, Y^\varepsilon(s) - Y^\varepsilon(s) \rangle \geq \frac{\delta}{\sqrt{d}} \right\} \\
\leq \exp \left[ -\frac{t-s}{\varepsilon} \sup_{r > 0} \left( \frac{r \delta}{\sqrt{d}(t-s)} - \bar{H}(r) \right) \right] \\
= \exp \left[ - \left( \frac{(t-s)}{\varepsilon} \zeta \left( \delta / \sqrt{d}(t-s) \right) \right) \right].
\]

Combining these bounds for the \( 2d \) choices of vectors \( \{\pm u^i, i = 1, 2, \ldots, d\} \) completes the proof of part (a) of the lemma.

(b) In terms of the quantity \( \hat{N}_{\alpha,x}^\varepsilon \) introduced in Lemma 10.3.3

\[
\exp[H^n(x, \alpha)] = E_x \left\{ \exp(\alpha, c_n[Y^\varepsilon_n(1/c_n) - x]) \right\} \\
= E_x \left\{ \hat{N}_{\alpha,x}^\varepsilon(1/c_n) \exp \left( c_n \int_0^{1/c_n} H(Y^\varepsilon_n(u), \alpha) \, du \right) \right\}.
\]

Part (c) of Lemma 10.3.3 and part (a) of Condition 10.2.2 yield

\[
\sup_{n \in \mathbb{N}} \sup_{x \in \mathbb{R}^d} \exp[H^n(x, \alpha)] \leq \sup_{n \in \mathbb{N}} \sup_{x \in \mathbb{R}^d} E_x \left\{ \hat{N}_{\alpha,x}^\varepsilon(1/c_n) \right\} \exp \left( \sup_{y \in \mathbb{R}^d} H(y, \alpha) \right) \\
\leq \exp \left( \sup_{y \in \mathbb{R}^d} H(y, \alpha) \right) < \infty.
\]

(c) According to part (b) of Condition 10.2.2, for each \( \alpha \in \mathbb{R}^d \) \( H(x, \alpha) \) is a uniformly continuous function of \( x \) in any compact subset of \( \mathbb{R}^d \). Thus, given \( \theta > 0 \) there exists \( \delta > 0 \) such that

\[
\sup_{x \in K} \sup_{\{y: \|y-x\| < \delta\}} |H(y, \alpha) - H(x, \alpha)| < \theta. \tag{10.10}
\]

For \( n \in \mathbb{N} \) we define

\[
A(\varepsilon_n, x, \delta) \equiv \left\{ \sup_{0 \leq u \leq 1/c_n} \|Y^\varepsilon_n(u) - x\| < \delta \right\}.
\]

Part (a) of the present proposition gives

\[
\sup_{x \in \mathbb{R}^d} P_x \{[A(\varepsilon_n, x, \delta)]^c\} \leq 2d \exp \left[ -\frac{1}{c_n \varepsilon_n} \zeta(\delta c_n / \sqrt{d}) \right] \leq 2d \exp \left[ -\zeta(\delta c_n / \sqrt{d}) \right]. \tag{10.11}
\]

For \( x \in K \) and \( \alpha \in \mathbb{R}^d \)

\[
\exp[H^n(x, \alpha) - H(x, \alpha)] \\
= E_x \left\{ \exp(\langle \alpha, c_n[Y^\varepsilon_n(1/c_n) - x] \rangle - H(x, \alpha)) \right\} \\
= E_x \left\{ \hat{N}_{\alpha,x}^\varepsilon(1/c_n) \exp \left( c_n \int_0^{1/c_n} [H(Y^\varepsilon_n(u), \alpha) - H(x, \alpha)] \, du \right) \right\}.
\]
while the definition of $\tau_{x,\delta}^n$ and parts (f) and (c) of Lemma 10.3.3 imply that there exists $M \in \mathbb{N}$ such that

$$\inf_{n \geq M} \inf_{x \in K} E_x \left\{ \tilde{N}_{\alpha,x}^n (1/c_n) \wedge \tau_{x,\delta}^n \exp \left( c_n \int_0^{(1/c_n) \wedge \tau_{x,\delta}^n} [H(Y^{\varepsilon_n}(u), \alpha) - H(x, \alpha)] \, du \right) \right\} \geq \exp(-\gamma_3(\delta) - \theta)$$

and

$$\sup_{n \in \mathbb{N}} \sup_{x \in K} E_x \left\{ \tilde{N}_{\alpha,x}^n (1/c_n) \wedge \tau_{x,\delta}^n \exp \left( c_n \int_0^{(1/c_n) \wedge \tau_{x,\delta}^n} [H(Y^{\varepsilon_n}(u), \alpha) - H(x, \alpha)] \, du \right) \right\} \leq \exp(\theta).$$

In the next paragraph we will prove that

$$\limsup_{n \to \infty} \sup_{x \in K} \left| E_x \left\{ \tilde{N}_{\alpha,x}^n (1/c_n) \exp \left( c_n \int_0^{1/c_n} [H(Y^{\varepsilon_n}(u), \alpha) - H(x, \alpha)] \, du \right) \right\} \right| = 0.$$ 

The last four displays will then imply that

$$\liminf_{n \to \infty} \inf_{x \in K} [H^n(x, \alpha) - H(x, \alpha)] \geq -\gamma_3(\delta) - \theta$$

and

$$\limsup_{n \to \infty} \sup_{x \in K} [H^n(x, \alpha) - H(x, \alpha)] \leq \theta.$$ 

Since $\gamma_3(\delta) \to 0$, sending $\delta \to 0$ and then $\theta \to 0$ will complete the proof of part (c).

Let $\gamma_4$ denote the finite number $[\sup_{y \in \mathbb{R}^d} H(y, \alpha) - \inf_{x \in K} H(x, \alpha)]$. Since $A(\varepsilon_n, x, \delta) = \{\tau_{x,\delta}^n > 1/c_n\}$, the Cauchy–Schwarz Inequality, part (e) of Lemma 10.3.3, and (10.11) imply that

$$\sup_{x \in K} \left| E_x \left\{ \tilde{N}_{\alpha,x}^n (1/c_n) \exp \left( c_n \int_0^{1/c_n} [H(Y^{\varepsilon_n}(u), \alpha) - H(x, \alpha)] \, du \right) \right\} \right| \leq \sup_{x \in K} E_x \left\{ 1_{[A(\varepsilon_n, x, \delta)]^c} \tilde{N}_{\alpha,x}^n (1/c_n) \exp \left( c_n \int_0^{1/c_n} [H(Y^{\varepsilon_n}(u), \alpha) - H(x, \alpha)] \, du \right) \right\}$$

$$\leq \exp(\gamma_4) \left( \sup_{x \in K} E_x \left\{ 1_{[A(\varepsilon_n, x, \delta)]^c} \tilde{N}_{\alpha,x}^n (1/c_n) \right\} + \sup_{x \in K} E_x \left\{ 1_{[A(\varepsilon_n, x, \delta)]^c} \tilde{N}_{\alpha,x}^n ((1/c_n) \wedge \tau_{x,\delta}^n) \right\} \right)$$
10.3. PROOF OF THE LAPLACE PRINCIPLE

\[
\leq \exp(\gamma_4) \left( \sup_{x \in K} \left[ E_x \left\{ \left. 1_{(A(\varepsilon_n, x, \delta))} \right\} \right] \right)^{1/2} \left[ E_x \left\{ \left. \tilde{N}_{\alpha,x} \left( \frac{1}{c_n} \right)^2 \right\} \right] \right)^{1/2} \\
+ \sup_{x \in K} \left[ E_x \left\{ \left. 1_{(A(\varepsilon_n, x, \delta))} \right\} \right] \right)^{1/2} \left[ E_x \left\{ \left. \tilde{N}_{\alpha,x} \left( \frac{1}{c_n} \wedge \tau^n_{x,\delta} \right)^2 \right\} \right] \right)^{1/2} \\
\leq 2\sqrt{2d} \exp \left( \gamma_4 - \frac{1}{2} \zeta \left( \frac{\delta c_n}{\sqrt{d}} \right) + \frac{1}{2} \gamma_2 \right).
\]

If we send \( n \to \infty \), then \( c_n = [1/\varepsilon_n] \to \infty \), and since \( \zeta \) is superlinear, it follows that \( \zeta(\delta c_n/\sqrt{d}) \to \infty \) and thus that

\[
\lim_{n \to \infty} \sup_{x \in K} \left| E_x \left\{ \tilde{N}_{\alpha,x} \left( \frac{1}{c_n} \right) \exp \left( c_n \int_0^{1/c_n} [H(Y^{\varepsilon_n}(u), \alpha) - H(x, \alpha)] \, du \right) \right\} - E_x \left\{ \tilde{N}_{\alpha,x} \left( \frac{1}{c_n} \wedge \tau^n_{x,\delta} \right) \exp \left( c_n \int_0^{1/c_n \wedge \tau^n_{x,\delta}} [H(Y^{\varepsilon_n}(u), \alpha) - H(x, \alpha)] \, du \right) \right\} \right| = 0.
\]

This is what we want to prove.

(d) Part (b) of Condition 10.2.2 states that \( H(x, \alpha) \) is continuous for \( (x, \alpha) \in \mathbb{R}^d \times \mathbb{R}^d \). Part (c) of the present lemma implies that for each \( x \) and \( \alpha \) in \( \mathbb{R}^d \)

\[
\lim_{n \to \infty} H^n(x, \alpha) = H(x, \alpha).
\]

Thus for each \( x \) the sequence of moment generating functions of \( \{\mu^n(\cdot|x), n \in \mathbb{N}\} \), which is the sequence \( \{\exp[H^n(x, \alpha)]\} \), converges for each \( \alpha \in \mathbb{R}^d \) to \( \exp[H(x, \alpha)] \). Since the latter is a continuous function of \( \alpha \in \mathbb{R}^d \), we conclude from part (a) of the continuity theorem for moment generating functions [Theorem A.3.20] that \( \mu^n(\cdot|x) \Rightarrow \mu(\cdot|x) \), where \( \mu(\cdot|x) \) is a probability measure on \( \mathbb{R}^d \) satisfying for each \( \alpha \in \mathbb{R}^d \)

\[
H(x, \alpha) = \log \int_{\mathbb{R}^d} \exp(\alpha, y) \mu(dy|x).
\]

Theorem A.5.3 implies that the family \( \{\mu(dy|x), x \in \mathbb{R}^d\} \) defines a stochastic kernel on \( \mathbb{R}^d \) given \( \mathbb{R}^d \). Finally, the continuity of the function mapping \( x \in \mathbb{R}^d \mapsto \mu(\cdot|x) \in \mathcal{P}(\mathbb{R}^d) \) follows from Condition 10.2.2 and part (b) of the continuity theorem for moment generating functions. The proof of Proposition 10.3.2 is complete. 

We denote by \( \rho(\cdot, \cdot) \) the Skorohod metric on \( \mathcal{D}([0,1] : \mathbb{R}^d) \). The next lemma shows that the sequences \( \{Y^{\varepsilon_n}\} \) and \( \{Z^n\} \) are superexponentially close in this metric. This is not surprising because \( Z^n \) is the piecewise linear interpolation of \( Y^{\varepsilon_n} \) with interpolating points \( j/n \). After the proof, we will use this fact to derive the Laplace limit for \( \{Y^{\varepsilon_n}\} \) with normalizing sequence \( \{1/\varepsilon_n\} \).

Lemma 10.3.4. We assume Condition 10.2.2. Let \( \{\varepsilon_n\} \) be any sequence in \( (0,1) \) converging to 0 and define \( c_n = [1/\varepsilon_n] \). For \( n \in \mathbb{N} \) and \( x \in \mathbb{R}^d \) let \( Y^{\varepsilon_n} \) denote the Markov process specified in Theorem 10.2.6 and define the piecewise linear process \( Z^n \) by formula (10.6). Then for any \( \delta > 0 \)

\[
\lim_{n \to \infty} \sup_{x \in \mathbb{R}^d} \varepsilon_n \log P_x \{\rho(Y^{\varepsilon_n}, Z^n) > \delta\} = -\infty.
\]
Proof. By part (b) of Theorem A.6.5, for any two functions \( \varphi \) and \( \psi \) in \( \mathcal{D}([0,1]: \mathbb{R}^d) \)

\[
\rho(\varphi, \psi) \leq \sup_{t \in [0,1]} \|\varphi(t) - \psi(t)\|.
\]

Hence it suffices to prove that

\[
\limsup_{n \to \infty} \sup_{x \in \mathbb{R}^d} \varepsilon_n \log P_x \left\{ \sup_{0 \leq t \leq 1} \|Y^{\varepsilon_n}(t) - Z^n(t)\| > \delta \right\} = -\infty.
\]

For any \( x \in \mathbb{R}^d \) and \( \delta > 0 \) part (a) of Proposition 10.3.2 implies that

\[
P_x \left\{ \sup_{0 \leq t \leq 1} \|Y^{\varepsilon_n}(t) - Z^n(t)\| > \delta \right\}
\]

\[
\leq \sum_{j=0}^{c_n-1} P_x \left\{ \sup_{t \in [j/c_n,(j+1)/c_n]} \|Y^{\varepsilon_n}(t) - Z^n(t)\| > \delta \right\}
\]

\[
\leq \sum_{j=0}^{c_n-1} P_x \left\{ \sup_{t \in [j/c_n,(j+1)/c_n]} \|Y^{\varepsilon_n}(t) - Y^{\varepsilon_n}(j/c_n)\| > \delta/2 \right\}
\]

\[
\leq 2dc_n \exp \left[ -\frac{1}{c_n \varepsilon_n} \zeta \left( \delta c_n / 2\sqrt{d} \right) \right].
\]

Since \( c_n \varepsilon_n \leq 1 \) and \( \zeta \) is superlinear, it follows that

\[
\limsup_{n \to \infty} \sup_{x \in \mathbb{R}^d} \varepsilon_n \log P_x \left\{ \sup_{0 \leq t \leq 1} \|Y^{\varepsilon_n}(t) - Z^n(t)\| > \delta \right\}
\]

\[
\leq -\lim_{n \to \infty} \inf_{c_n \varepsilon_n} \frac{1}{c_n \varepsilon_n} \zeta \left( \delta c_n / 2\sqrt{d} \right) = -\infty.
\]

This completes the proof. \( \blacksquare \)

Let \( h \) be any bounded continuous function mapping \( \mathcal{D}([0,1]: \mathbb{R}^d) \) into \( \mathbb{R} \). At the beginning of this section we showed that the function \( I_x \) in Theorem 10.2.6 has compact level sets. Hence in order to complete the proof of the theorem, it remains to prove the Laplace limit

\[
\lim_{n \to \infty} \varepsilon_n \log E_x \left\{ \exp \left[ -\frac{1}{\varepsilon_n} h(Y^{\varepsilon_n}) \right] \right\} = -\inf_{\varphi \in \mathcal{D}([0,1]: \mathbb{R}^d)} \{ I_x(\varphi) + h(\varphi) \}.
\]

Since when relativized to \( \mathcal{C}([0,1]: \mathbb{R}^d) \) the Skorohod topology coincides with the uniform topology [Theorem A.6.5 (d)], \( h \) is also a bounded continuous function mapping \( \mathcal{C}([0,1]: \mathbb{R}^d) \) into \( \mathbb{R} \). Hence by Proposition 10.3.1

\[
\lim_{n \to \infty} \varepsilon_n \log E_x \left\{ \exp \left[ -\frac{1}{\varepsilon_n} h(Z^n) \right] \right\} = -\inf_{\varphi \in \mathcal{C}([0,1]: \mathbb{R}^d)} \{ I_x(\varphi) + h(\varphi) \}.
\]

The infimum of \( I_x + h \) over \( \mathcal{C}([0,1]: \mathbb{R}^d) \) equals the infimum of \( I_x + h \) over \( \mathcal{D}([0,1]: \mathbb{R}^d) \) since \( I_x(\varphi) = \infty \) on \( \mathcal{D}([0,1]: \mathbb{R}^d) \) \( \setminus \mathcal{C}([0,1]: \mathbb{R}^d) \). By Lemma 10.3.4, which shows that the sequences \( \{Y^{\varepsilon_n}, n \in \mathbb{N}\} \) and \( \{Z^n, n \in \mathbb{N}\} \) are superexponentially close, and by a trivial modification of Theorem 1.3.3, which involves replacing the normalizing sequence \( n \in \mathbb{N} \) by \( \{1/\varepsilon_n\} \), the Laplace limit for \( \{Z^n\} \) implies the Laplace limit for \( \{Y^{\varepsilon_n}\} \). The proof of Theorem 10.2.6 is complete. \( \blacksquare \)
Appendix A

Background Material

A.1 Introduction

In Section A.2 we state the Dynkin Class Theorem followed by two results concerning measures on a complete, separable metric space \( X \) or, as it is customarily designated, a Polish space. In Section A.3 we summarize what we need of the theory of weak convergence of probability measures. Although a number of the results in these two sections are valid without the assumption of completeness or separability, for ease of exposition we will not point out which ones. Excellent references for Sections A.2 and A.3 are the texts of Billingsley [9] and of Ethier and Kurtz [50]. In Section A.4 we collect some facts from probability theory. Section A.5 summarizes what is needed about stochastic kernels while Section A.6 contains some miscellaneous information from analysis.

We follow the same conventions as throughout the book. Thus the \( \sigma \)-field on a Polish space \( X \) is always the Borel \( \sigma \)-field. By a measure on \( X \) we mean a Borel measure on \( X \). A function mapping a Polish space into a Polish space is called measurable if it is measurable with respect to the underlying Borel \( \sigma \)-fields.

A.2 Measure Theory

Let \( \Gamma \) be a set. A collection \( \mathcal{D} \) of subsets of \( \Gamma \) is called a dynamkin class if \( \Gamma \in \mathcal{D} \), if \( A \) and \( B \) in \( \mathcal{D} \) and \( A \subset B \) imply \( B \setminus A \in \mathcal{D} \), and if \( \{ A_i, i \in \mathbb{N} \} \subset \mathcal{D} \) and \( A_1 \subset A_2 \subset \ldots \) imply \( \bigcup_{i \in \mathbb{N}} A_i \in \mathcal{D} \). The first theorem is called the Dynkin Class Theorem.

**Theorem A.2.1** [50, page 497]. Let \( \mathcal{H} \) be a collection of subsets of \( \Gamma \) such that \( A \) and \( B \) in \( \mathcal{H} \) imply \( A \cap B \in \mathcal{H} \). If \( \mathcal{D} \) is a Dynkin class and \( \mathcal{H} \subset \mathcal{D} \), then \( \sigma(\mathcal{H}) \subset \mathcal{D} \).

Let \( X \) be a Polish space with metric \( d(x, y) \) and Borel \( \sigma \)-field \( \mathcal{B}_X \). A class \( \mathcal{E} \) of bounded measurable functions mapping \( X \) into \( \mathbb{R} \) is said to be measure–determining for \( X \) if whenever \( \gamma \) and \( \theta \) are finite measures on \( X \) satisfying

\[
\int_X f \, d\gamma = \int_X f \, d\theta \quad \text{for all } f \in \mathcal{E},
\]

then \( \gamma = \theta \). The next result presents two measure–determining classes of functions.

301
Theorem A.2.2. The following classes are measure-determining for $\mathcal{X}$.

(a) $\mathcal{E} \doteq \{ f : f = 1_F, F \subset \mathcal{X} \text{ closed} \}$.

(b) The class of all bounded, uniformly continuous functions mapping $\mathcal{X}$ into $\mathbb{R}$.

Proof. (a) Let $\gamma$ and $\theta$ be finite measures on $\mathcal{X}$. The collection

$$\mathcal{D} \doteq \{ A \in \mathcal{B}_X : \gamma(A) = \theta(A) \}$$

is a Dynkin class which by hypothesis contains the closed sets. The Dynkin Class Theorem implies that $\mathcal{D} = \mathcal{B}_X$ and thus that $\mathcal{E}$ is measure-determining for $\mathcal{X}$.

(b) Given $F$ a closed set, we define for $n \in \mathbb{N}$ and $x \in \mathcal{X}$

$$d(x, F) \doteq \inf_{y \in F} d(x, y) \text{ and } f_n(x) \doteq \varphi(n \, d(x, F)), \quad (A.1)$$

where

$$\varphi(t) \doteq \begin{cases} 1 - t & \text{ if } t \in [0, 1] \\ 0 & \text{ if } t \geq 1. \end{cases}$$

Then $f_n$ is bounded and uniformly continuous, $f_n \geq 0$, and $f_n \downarrow 1_F$ as $n \to \infty$. Since for all $n$

$$\int_{\mathcal{X}} f_n \, d\gamma = \int_{\mathcal{X}} f_n \, d\theta,$$

the Lebesgue Dominated Convergence Theorem implies that $\gamma(F) = \theta(F)$. Part (a) of the theorem now completes the proof. ■

A probability measure $\gamma$ on $\mathcal{X}$ is said to be tight if for each $\varepsilon > 0$ there exists a compact subset $K$ of $\mathcal{X}$ such that $\gamma(K) \geq 1 - \varepsilon$.

Theorem A.2.3 [9, page 10]. Every probability measure on $\mathcal{X}$ is tight.

A.3 Weak Convergence of Probability Measures

Throughout this section $\mathcal{X}$ is a Polish space with metric $d(x, y)$. $\mathcal{M}(\mathcal{X})$ denotes the set of subprobability measures on $\mathcal{X}$; i.e., the set of measures $\gamma$ for which $\gamma(\mathcal{X}) \leq 1$. $\mathcal{P}(\mathcal{X})$ denotes the subset of $\mathcal{M}(\mathcal{X})$ consisting of probability measures and $\mathcal{C}_b(\mathcal{X})$ denotes the space of bounded continuous functions mapping $\mathcal{X}$ into $\mathbb{R}$. Let $\{ \theta_n, n \in \mathbb{N} \}$ be a sequence in $\mathcal{M}(\mathcal{X})$. We say that $\{ \theta_n \}$ converges weakly to $\theta$, and write $\theta_n \rightharpoonup \theta$, if for each $g \in \mathcal{C}_b(\mathcal{X})$

$$\lim_{n \to \infty} \int_{\mathcal{X}} g \, d\theta_n = \int_{\mathcal{X}} g \, d\theta.$$

$\mathcal{P}(\mathcal{X})$ is made into a topological space by taking as the basic open neighborhoods of $\gamma \in \mathcal{P}(\mathcal{X})$ the sets of the form

$$\left\{ \theta \in \mathcal{P}(\mathcal{X}) : \left| \int_{\mathcal{X}} g_i \, d\theta - \int_{\mathcal{X}} g_i \, d\gamma \right| < \varepsilon, i = 1, 2, \ldots, k \right\},$$
where \( \varepsilon \) is positive, \( k \) is a positive integer, and \( g_1, g_2, \ldots, g_k \) are in \( C_0(\mathcal{X}) \). The resulting topology is called the **topology of weak convergence** or simply the **weak topology**. \( \mathcal{M}(\mathcal{X}) \) is topologized in a similar way.

In order to introduce a metric on \( \mathcal{P}(\mathcal{X}) \), we define for \( A \subset \mathcal{X} \) and \( \varepsilon > 0 \)

\[
A^{(\varepsilon)} = \{ x \in \mathcal{X} : d(x, A) < \varepsilon \}.
\]

For \( \gamma \) and \( \theta \) in \( \mathcal{P}(\mathcal{X}) \) we then define

\[
\mathcal{L}(\gamma, \theta) \doteq \inf \{ \varepsilon > 0 : \gamma(F) \leq \theta(F^{(\varepsilon)}) + \varepsilon \text{ for all closed subsets } F \text{ of } \mathcal{X} \}.
\]

\( \mathcal{L}(\gamma, \theta) \) defines a metric on \( \mathcal{P}(\mathcal{X}) \) known as the **Lévy–Prohorov metric** [50, page 96]. Given probability measures \( \alpha \) and \( \beta \) on \( \mathcal{X} \), the **total variation norm** is defined by

\[
\| \alpha - \beta \|_v \doteq (\alpha - \beta)^+(\mathcal{X}) + (\alpha - \beta)^-(\mathcal{X}),
\]

where \( (\alpha - \beta)^+ \) and \( (\alpha - \beta)^- \) are the measures appearing in the Jordan decomposition of \( \alpha - \beta \). For any bounded measurable function \( h \) mapping \( \mathcal{X} \) into \( \mathbb{R} \)

\[
\left| \int_{\mathcal{X}} h \, d\alpha - \int_{\mathcal{X}} h \, d\beta \right| \leq \| h \|_{\infty} \| \alpha - \beta \|_v.
\]

The following result gives a simple bound on \( \mathcal{L}(\gamma, \theta) \).

**Theorem A.3.1.** For all \( \gamma \) and \( \theta \) in \( \mathcal{P}(\mathcal{X}) \)

\[
\mathcal{L}(\gamma, \theta) \leq \| \gamma - \theta \|_v \wedge 1.
\]

**Proof.** Let \( \varepsilon \doteq \| \gamma - \theta \|_v \). Since for any closed set \( F \)

\[
\gamma(F) = \theta(F) + \gamma(F) - \theta(F) \leq \theta(F^{(\varepsilon)}) + \| \gamma - \theta \|_v = \theta(F^{\varepsilon}) + \varepsilon,
\]

it follows that \( \mathcal{L}(\gamma, \theta) \leq \| \gamma - \theta \|_v \). The inequality \( \mathcal{L}(\gamma, \theta) \leq 1 \) follows from the definition of \( \mathcal{L} \). \( \blacksquare \)

As we state in the next theorem, the Lévy–Prohorov metric is compatible with the weak topology and with respect to it \( \mathcal{P}(\mathcal{X}) \) is Polish.

**Theorem A.3.2** [50, pages 101 and 108]. Let \( \{ \theta_n \} \) be a sequence in \( \mathcal{P}(\mathcal{X}) \). Then \( \theta_n \to \theta \in \mathcal{P}(\mathcal{X}) \) if and only if \( \mathcal{L}(\theta_n, \theta) \to 0 \). Furthermore, with respect to the Lévy–Prohorov metric \( \mathcal{P}(\mathcal{X}) \) is complete and separable.

We next introduce a metric on \( \mathcal{M}(\mathcal{X}) \) having properties analogous to those of \( \mathcal{L}(\cdot, \cdot) \). For \( \gamma \) and \( \theta \) in \( \mathcal{M}(\mathcal{X}) \) we define

\[
m(\gamma, \theta) \doteq [\gamma(\mathcal{X}) \wedge \theta(\mathcal{X})] \cdot \mathcal{L} \left( \frac{\gamma}{\gamma(\mathcal{X})}, \frac{\theta}{\theta(\mathcal{X})} \right) + |\gamma(\mathcal{X}) - \theta(\mathcal{X})|.
\]

The convention is that if \( \gamma \) or \( \theta \) equals the zero measure on \( \mathcal{X} \), then the first term in this definition is 0. Properties of \( m \) are given in the next theorem. The straightforward proof is omitted.
Theorem A.3.3. The quantity $m(\gamma, \theta)$ defines a metric on $\mathcal{M}(\mathcal{X})$ which is compatible with the weak topology. With respect to this metric $\mathcal{M}(\mathcal{X})$ is complete and separable.

In the remainder of this section we present additional results from the theory of weak convergence of probability measures on $\mathcal{X}$. A number of these results give properties of a sequence of probability measures on $\mathcal{X}$ converging weakly to a probability measure. Although we shall not bother to restate them, all of these results are valid, without change, in the context of weak convergence of subprobability measures on $\mathcal{X}$. This observation is needed in Chapter 7. The proofs use the following simple normalization device. If $\{\theta_n, n \in \mathbb{N}\}$ is a sequence of nonzero subprobability measures on $\mathcal{X}$ converging weakly to $\theta$, then either $\theta = 0$ or $\theta_n(\mathcal{X}) \to \theta(\mathcal{X}) > 0$ and for all sufficiently large $n$ the sequence of probability measures $\left\{\frac{1}{\theta_n(\mathcal{X})} \theta_n(dx)\right\}$ converges weakly to the probability measure $\frac{1}{\theta(\mathcal{X})} \theta(dx)$.

The first result, known as the Portmanteau Theorem, gives a number of useful conditions that are equivalent to weak convergence. For $\theta \in \mathcal{P}(\mathcal{X})$, a Borel set $A$ whose boundary $\partial A$ satisfies $\theta(\partial A) = 0$ is called a $\theta$–continuity set.

Theorem A.3.4 [9, page 11]. Let $\{\theta_n\}$ and $\theta$ be probability measures on $\mathcal{X}$. The following five conditions are equivalent.

(a) $\theta_n \rightharpoonup \theta$.

(b) $\lim_{n \to \infty} \int_{\mathcal{X}} g \, d\theta_n = \int_{\mathcal{X}} g \, d\theta$ for all bounded, uniformly continuous functions $g$ mapping $\mathcal{X}$ into $R$.

(c) $\limsup_{n \to \infty} \theta_n(F) \leq \theta(F)$ for all closed subsets $F$ of $\mathcal{X}$.

(d) $\liminf_{n \to \infty} \theta_n(G) \geq \theta(G)$ for all open subsets $G$ of $\mathcal{X}$.

(e) $\lim_{n \to \infty} \theta_n(A) = \theta(A)$ for all $\theta$–continuity sets $A$.

Remark A.3.5. The proof that (b) implies (c) uses the bounded, uniformly continuous functions $f_n$ defined in equation (A.1). Since these functions are also Lipschitz continuous, we can augment the Portmanteau Theorem with the following additional equivalent condition: $\theta_n \rightharpoonup \theta$ if and only if

$$\lim_{n \to \infty} \int_{\mathcal{X}} g \, d\theta_n = \int_{\mathcal{X}} g \, d\theta$$

for all bounded, Lipschitz continuous functions $g$ mapping $\mathcal{X}$ into $R$.

It is useful to introduce a notion of convergence of random variables that is equivalent to weak convergence of probability measures. Let $\{Y_n, n \in \mathbb{N}\}$ and $Y$ be random variables taking values in $\mathcal{X}$. We say that the sequence $\{Y_n\}$ converges in distribution to $Y$, and write $Y_n \xrightarrow{D} Y$, if the sequence of distributions of $\{Y_n\}$ converge weakly on $\mathcal{X}$ to the distribution of $Y$. In other words, $Y_n \xrightarrow{D} Y$ if and only if for any $g \in \mathcal{C}_b(\mathcal{X})$

$$\lim_{n \to \infty} E\{g(Y_n)\} = E\{g(Y)\}.$$

The notion of convergence in distribution does not require that the random variables $Y_n$ and $Y$ be defined on a common probability space. Nevertheless, we use the same symbol $E$ to denote the various expectation operators. The following is a simple consequence of the definition of convergence in distribution.
A.3. WEAK CONVERGENCE OF PROBABILITY MEASURES

Theorem A.3.6. If \( Y_n \overset{D}{\to} Y \) and \( f \in \mathcal{C}_b(\mathcal{X}) \), then \( f(Y_n) \overset{D}{\to} f(Y) \).

Let \( \{Y_n, n \in \mathbb{N}\} \) be a sequence of real-valued random variables that are defined on a sequence of probability spaces \( \{(\Omega_n, \mathcal{F}_n, P_n), n \in \mathbb{N}\} \). We say that the sequence \( \{Y_n\} \) converges in probability to 0, and write \( Y_n \overset{P}{\to} 0 \), if for every \( \varepsilon > 0 \)
\[
\lim_{n \to \infty} P_n\{|Y_n| \geq \varepsilon\} = 0.
\]
The proof of the following well known result is omitted.

Theorem A.3.7. \( Y_n \overset{P}{\to} 0 \) if and only if \( Y_n \overset{D}{\to} 0 \).

The next theorem shows that convergence in distribution is preserved if a sequence of random variables is replaced by another suitably close sequence.

Theorem A.3.8 [9, page 25]. We assume that for each \( n \in \mathbb{N} \) the random variables \( X_n \) and \( Y_n \) are defined on the same probability space and take values in \( \mathcal{X} \). Let \( Y \) be another random variables taking values in \( \mathcal{X} \) such that \( X_n \overset{D}{\to} Y \). If for each \( \varepsilon > 0 \)
\[
\lim_{n \to \infty} P\{d(X_n, Y_n) \geq \varepsilon\} = 0,
\]
then \( Y_n \overset{D}{\to} Y \).

As we have seen numerous times in the text, the Skorohod Representation Theorem is an extremely useful device when calculating limits of expectations of random variables that converge in distribution. The reader is referred to the second paragraph of Section 2.5, where we discuss the convention to be followed when the Skorohod Representation Theorem is applied.

Theorem A.3.9 [50, page 102]. Let \( \{Y_n, n \in \mathbb{N}\} \) and \( Y \) be random variables taking values in \( \mathcal{X} \) and assume that \( Y_n \overset{D}{\to} Y \). Then there exists a probability space \( (\Omega, \mathcal{F}, \overline{P}) \) on which \( \mathcal{X} \)-valued random variables \( \{\overline{Y}_n, n \in \mathbb{N}\} \) and \( \overline{Y} \) are defined with the following properties: each \( \overline{Y}_n \) has the same distribution as \( Y_n \), \( \overline{Y} \) has the same distribution as \( Y \), and the sequence \( \{\overline{Y}_n\} \) converges to \( \overline{Y} \) w.p.1.

Let \( \psi \) be a measurable function mapping \( \mathcal{X} \) into a Polish space \( \mathcal{Y} \). Then each probability measure \( \theta \) on \( \mathcal{X} \) gives rise to a probability measure \( \theta \circ \psi^{-1} \) on \( \mathcal{Y} \). For Borel subsets \( A \) of \( \mathcal{Y} \) this measure is defined by \( (\theta \circ \psi^{-1})(A) \triangleq \theta(\psi^{-1}(A)) \). A natural question is under what conditions on \( \psi \) does \( \theta_n \Longrightarrow \theta \) on \( \mathcal{X} \) imply that \( \theta_n \circ \psi^{-1} \Longrightarrow \theta \circ \psi^{-1} \) on \( \mathcal{Y} \). This obviously holds if \( \psi \) is continuous, since then \( g \circ \psi \) is continuous on \( \mathcal{X} \) whenever \( g \) is continuous on \( \mathcal{Y} \) and therefore
\[
\lim_{n \to \infty} \int_{\mathcal{Y}} g d(\theta_n \circ \psi^{-1}) = \lim_{n \to \infty} \int_{\mathcal{X}} g \circ \psi d\theta_n = \int_{\mathcal{X}} g \circ \psi d\theta = \int_{\mathcal{Y}} g d(\theta \circ \psi^{-1}).
\]
This fact is easy to interpret in terms of random variables. Thus, if \( Y_n \) and \( Y \) have the respective distributions \( \theta_n \) and \( \theta \), then \( \theta_n \circ \psi^{-1} \Longrightarrow \theta \circ \psi^{-1} \) if and only if \( \psi(Y_n) \overset{D}{\to} \psi(Y) \).
If ψ is discontinuous, then one might expect the same limit to hold if in some sense the discontinuities of ψ are not seen by θ_n and θ. Part (a) of the next theorem shows that such a condition can be given in terms of θ alone. We will refer to part (a) as the **Continuous Mapping Theorem.** The proof is taken from [50, page 103] and uses the Skorohod Representation Theorem, which makes the result almost transparent. Part (b) is an analogous result that involves sequences of measurable functions. The proof of part (b) uses Prohorov’s Theorem, which is stated in Theorem A.3.15. We denote by D_ψ the set of discontinuities of ψ. This is a Borel subset of X [9, page 225].

**Theorem A.3.10** [9, pages 30 and 34]. Let ψ be a measurable function mapping X into a Polish space Y. Let {θ_n, n ∈ N} be a sequence in P(X) converging weakly to θ. The following conclusions hold.

(a) If θ(D_ψ) = 0, then \( \theta_n \circ \psi^{-1} \xrightarrow{w} \theta \circ \psi^{-1} \) on Y.

(b) Let \( \{\psi_n, n \in N\} \) be a sequence of measurable functions mapping X into Y which converges to ψ uniformly on compact subsets of Y. If in addition ψ is continuous, then \( \theta_n \circ (\psi_n)^{-1} \xrightarrow{w} \theta \circ \psi^{-1} \) on Y.

**Proof.** (a) Since \( \theta_n \xrightarrow{w} \theta \), the Skorohod Representation Theorem guarantees the existence of a probability space \((\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})\) on which X-valued random variables \(\{\hat{Y}_n, n \in N\}\) and \(\hat{Y}\) are defined with the following properties: each \(\hat{Y}_n\) has the distribution \(\theta_n\), \(\hat{Y}\) has the distribution \(\theta\), and the sequence \(\{\hat{Y}_n\}\) converges to \(\hat{Y}\) w.p.1. The assumption \(\theta(D_\psi) = 0\) translates into the statement that \(\hat{P}(\hat{Y} \in D_\psi) = 0\), which implies that \(\psi(\hat{Y}_n) \to \psi(\hat{Y})\) w.p.1. Hence for any \(g \in C_b(Y)\) the Lebesgue Dominated Convergence Theorem gives the limit

\[
\lim_{n \to \infty} \int_Y g d(\theta_n \circ \psi^{-1}) = \lim_{n \to \infty} \int_\Omega g(\psi(\hat{Y}_n)) d\hat{P} = \int_\Omega g(\psi(\hat{Y})) d\hat{P} = \int_Y g d(\theta \circ \psi^{-1}).
\]

(b) We show that for any bounded, uniformly continuous function \(g\) mapping Y into \(\mathbb{R}\)

\[
\lim_{n \to \infty} \int_X g \circ \psi_n d\theta_n = \int_X g \circ \psi d\theta.
\]

Let \(\varepsilon > 0\) be given. According to Prohorov’s Theorem, there exists a compact set \(K\) such that \(\sup_{n \in N} \theta_n(K^c) \leq \varepsilon\). By the continuity of \(g \circ \psi\), the weak convergence \(\theta_n \xrightarrow{w} \theta\), the uniform continuity of \(g\), and the uniform convergence on \(K\) of \(\psi_n\) to \(\psi\), there exists \(N \in N\) such that for all \(n \geq N\)

\[
\left| \int_X g \circ \psi d\theta - \int_X g \circ \psi_n d\theta_n \right| \leq \varepsilon \quad \text{and} \quad \sup_{x \in K} |g(\psi(x)) - g(\psi_n(x))| \leq \varepsilon.
\]

Hence for all \(n \geq N\)

\[
\left| \int_X g \circ \psi d\theta - \int_X g \circ \psi_n d\theta_n \right| \\
\leq \left| \int_X g \circ \psi d\theta - \int_X g \circ \psi d\theta_n \right| + \left| \int_X g \circ \psi d\theta_n - \int_X g \circ \psi_n d\theta_n \right| \\
\leq \varepsilon + 2\|g\|_\infty \theta_n(K^c) + \sup_{x \in K} |g(\psi(x)) - g(\psi_n(x))| \\
\leq 2\|g\|_\infty \varepsilon + 2\varepsilon.
\]
Since $\varepsilon > 0$ is arbitrary, the proof is done. ■

The Continuous Mapping Theorem has the following consequence. We recall that if $\psi$ is a measurable function mapping $\mathcal{X}$ into a Polish space $\mathcal{Y}$, then $D_\psi$ denotes the set of discontinuities of $\psi$.

**Theorem A.3.11.** Let $\psi$ be a bounded measurable function mapping $\mathcal{X}$ into $\mathbb{R}$. Let $\{\theta_n, n \in \mathbb{N}\}$ be a sequence in $\mathcal{P}(\mathcal{X})$ converging weakly to $\theta$. If $\theta(D_\psi) = 0$, then

$$
\lim_{n \to \infty} \int_{\mathcal{X}} \psi \, d\theta_n = \int_{\mathcal{X}} \psi \, d\theta.
$$

**Proof.** If $\|\psi\|_\infty \leq M < \infty$, then let $g \in C_b(\mathbb{R})$ satisfy $g(y) = y$ for all $y \in \mathbb{R}$ satisfying $|y| \leq M$. Since $g(\psi(x)) = \psi(x)$ for all $x \in \mathcal{X}$, the theorem follows from part (a) of Theorem A.3.10. ■

The concept of weak convergence is defined in terms of the convergence of integrals of bounded continuous functions. The next theorem proves a limit property involving integrals of nonnegative, lower semicontinuous functions. This result is used so frequently in the book that it deserves two separate proofs, the first analytic and the second probabilistic.

**Theorem A.3.12.** Let $g$ be a nonnegative, lower semicontinuous function mapping $\mathcal{X}$ into $\mathbb{R} \cup \{\infty\}$. Let $\{\theta_n\}$ be a sequence in $\mathcal{P}(\mathcal{X})$ converging weakly to $\theta$. Then

$$
\liminf_{n \to \infty} \int_{\mathcal{X}} g \, d\theta_n \geq \int_{\mathcal{X}} g \, d\theta.
$$

Thus the function mapping $\theta \in \mathcal{P}(\mathcal{X}) \longmapsto \int_{\mathcal{X}} g \, d\theta$ is lower semicontinuous.

**Proof 1.** The first proof appeals to Lemma 1.3.5, which constructs a sequence $\{g_j, j \in \mathbb{N}\} \subseteq C_b(\mathcal{X})$ with the property that $g_j \uparrow g$. Thus for each $j \in \mathbb{N}$

$$
\liminf_{n \to \infty} \int_{\mathcal{X}} g \, d\theta_n \geq \liminf_{n \to \infty} \int_{\mathcal{X}} g_j \, d\theta_n = \int_{\mathcal{X}} g_j \, d\theta.
$$

The proof is completed by sending $j \to \infty$ and using the Monotone Convergence Theorem. ■

**Proof 2.** Since $\theta_n \Rightarrow \theta$, the Skorohod Representation Theorem guarantees the existence of a probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$ on which $\mathcal{X}$-valued random variables $\{\hat{Y}_n, n \in \mathbb{N}\}$ and $\hat{Y}$ are defined with the following properties: each $\hat{Y}_n$ has the distribution $\theta_n$, $\hat{Y}$ has the distribution $\theta$, and $\hat{Y}_n \to \hat{Y}$ w.p.1. Thus w.p.1

$$
\liminf_{n \to \infty} g(\hat{Y}_n) \geq g(\hat{Y}),
$$

and Fatou’s Lemma yields

$$
\liminf_{n \to \infty} \int_{\mathcal{X}} g \, d\theta_n = \liminf_{n \to \infty} \int_{\hat{\Omega}} g(\hat{Y}_n) \, d\hat{P} \geq \int_{\hat{\Omega}} g(\hat{Y}) \, d\hat{P} = \int_{\mathcal{X}} g \, d\theta.
$$
This completes the proof. ■

The next theorem is obtained by combining Theorem A.3.10 with Theorem A.3.12.

**Theorem A.3.13.** Let \( \psi \) be a measurable function mapping \( \mathcal{X} \) into a Polish space \( \mathcal{Y} \) and \( g \) a nonnegative, lower semicontinuous function mapping \( \mathcal{Y} \) into \( \mathbb{R} \cup \{\infty\} \). Let \( \{\theta_n\} \) be a sequence in \( \mathcal{P}(\mathcal{X}) \) converging weakly to \( \theta \). The following conclusions hold.

(a) If \( \theta(D_\psi) = 0 \), then

\[
\liminf_{n \to \infty} \int_{\mathcal{X}} g \circ \psi \, d\theta_n \geq \int_{\mathcal{X}} g \circ \psi \, d\theta.
\]

(b) Let \( \{\psi_n, n \in \mathbb{N}\} \) be a sequence of measurable functions mapping \( \mathcal{X} \) into \( \mathcal{Y} \) which converges to \( \psi \) uniformly on compact subsets of \( \mathcal{Y} \). If in addition \( \psi \) is continuous, then

\[
\liminf_{n \to \infty} \int_{\mathcal{X}} g \circ \psi_n \, d\theta_n \geq \int_{\mathcal{X}} g \circ \psi \, d\theta.
\]

**Proof.** By part (a) of Theorem A.3.10 \( \theta_n \circ \psi^{-1} \implies \theta \circ \psi^{-1} \) on \( \mathcal{Y} \), and thus by Theorem A.3.12

\[
\liminf_{n \to \infty} \int_{\mathcal{X}} g \circ \psi \, d\theta_n = \liminf_{n \to \infty} \int_{\mathcal{Y}} g \circ (\theta_n \circ \psi^{-1})
\geq \int_{\mathcal{X}} g \circ (\theta \circ \psi^{-1}) = \int_{\mathcal{X}} g \circ \psi \, d\theta.
\]

This completes the proof of part (a). The proof of part (b) follows similarly from part (b) of Theorem A.3.10 and Theorem A.3.12. ■

We need a result in order to deal with weak convergence on a product of Polish spaces. A subset \( \Gamma \) of \( \mathcal{C}_b(\mathcal{X}) \) is said to be **convergence-determining** for \( \mathcal{X} \) if the following holds. Whenever \( \{\theta_n, n \in \mathbb{N}\} \) and \( \theta \) are probability measures on \( \mathcal{X} \) satisfying

\[
\lim_{n \to \infty} \int_{\mathcal{X}} f \, d\theta_n = \int_{\mathcal{X}} f \, d\theta \quad \text{for all } f \in \Gamma,
\]

then \( \theta_n \implies \theta \). By definition \( \mathcal{C}_b(\mathcal{X}) \) itself is convergence-determining for \( \mathcal{X} \).

**Theorem A.3.14** [50, page 115]. Let \( \mathcal{X} \) and \( \mathcal{Y} \) be Polish spaces and define

\[
\Gamma \doteq \{ f : f(x, y) = g(x) h(y) \text{ for } g \in \mathcal{C}_b(\mathcal{X}) \text{ and } h \in \mathcal{C}_b(\mathcal{Y}) \}.
\]

Then \( \Gamma \) is convergence-determining for the product space \( \mathcal{X} \times \mathcal{Y} \).

We next state **Prohorov's Theorem**, which characterizes relatively compact subsets of \( \mathcal{P}(\mathcal{X}) \). It is one of the main results in the theory. A family \( \Phi \) of probability measures on \( \mathcal{X} \) is said to be **tight** if for each \( \varepsilon > 0 \) there exists a compact set \( K \) such that

\[
\inf_{\gamma \in \Phi} \gamma(K) \geq 1 - \varepsilon.
\]
Theorem A.3.15 [50, page 103]. A family of probability measures on $\mathcal{X}$ is relatively compact with respect to weak convergence if and only if it is tight. In particular, if $\theta_n \rightarrow \theta$, then $\{\theta_n\}$ is tight.

If $\mathcal{X}$ is compact, then $\mathcal{P}(\mathcal{X})$ is tight. Prohorov’s Theorem yields the following useful fact.

Corollary A.3.16. If $\mathcal{X}$ is a compact Polish space, then $\mathcal{P}(\mathcal{X})$ is compact as is $\mathcal{M}(\mathcal{X})$.

To prove the compactness of $\mathcal{M}(\mathcal{X})$, let $\{\theta_n, n \in \mathbb{N}\}$ be any sequence in $\mathcal{M}(\mathcal{X})$. If $\lim_{n \to \infty} \theta_n(\mathcal{X}) = 0$, then $\theta_n \rightarrow 0$. Otherwise, there exists a subsequence and $\alpha > 0$ such that $0 < \theta_n(\mathcal{X}) \rightarrow \alpha$. Since $\mathcal{P}(\mathcal{X})$ is compact, $\left\{\frac{1}{\theta_n(\mathcal{X})}\theta_n(dx)\right\}$ has a subsequence converging weakly to some $\rho \in \mathcal{P}(\mathcal{X})$. It follows that $\theta_n \rightarrow \alpha \rho$ and thus that $\mathcal{M}(\mathcal{X})$ is compact.

Several times in the text we have made use of the concept of a tightness function. Let us recall the definition. A measurable function $g$ mapping $\mathcal{X}$ into $\mathbb{R} \cup \{\infty\}$ is called a tightness function on $\mathcal{X}$ if the following two conditions hold:

(a) $\inf_{x \in \mathcal{X}} g(x) > -\infty$.
(b) For each $M < \infty$, the level set $Z_g(M) \equiv \{x \in \mathcal{X} : g(x) \leq M\}$ is a relatively compact subset of $\mathcal{X}$.

The next theorem states two properties of a tightness function. Part (a) is well known.

Theorem A.3.17. Let $g$ be a tightness function on $\mathcal{X}$. We define a function $G(\theta)$ mapping $\mathcal{P}(\mathcal{S})$ into $\mathbb{R} \cup \{\infty\}$ by

$$G(\theta) \equiv \int_{\mathcal{X}} g \, d\theta.$$ 

The following conclusions hold.

(a) For each $M < \infty$, the set $Z_G(M) \equiv \{\theta \in \mathcal{P}(\mathcal{X}) : G(\theta) \leq M\}$ is tight.
(b) $G$ is a tightness function on $\mathcal{P}(\mathcal{X})$.

Proof. (a) Without loss of generality, we assume that $\inf_{x \in \mathcal{X}} g(x) = 0$. Let $\varepsilon > 0$ be given. For any $\Lambda \in (0, \infty)$, since $(\text{cl } Z_g(\Lambda))^c \subset (Z_g(\Lambda))^c$, we have $g(x) > \Lambda$ for all $x \in (\text{cl } Z_g(\Lambda))^c$. Thus for any $M < \infty$, $\theta \in Z_G(M)$, and $\Lambda \in (0, \infty)

$$M \geq G(\theta) = \int_{\mathcal{X}} g \, d\theta \geq \int_{(\text{cl } Z_g(\Lambda))^c} g \, d\theta \geq \Lambda \theta\{(\text{cl } Z_g(\Lambda))^c\}.$$

It follows that

$$\sup_{\theta \in Z_G(M)} \theta\{(\text{cl } Z_g(\Lambda))^c\} \leq \frac{M}{\Lambda}.$$

If $\Lambda$ is chosen so large that $M/\Lambda \leq \varepsilon$, then the compactness of $\text{cl } Z_g(\Lambda)$ yields the tightness of $Z_G(M)$. 


(b) Part (a) and Prohorov’s Theorem imply that \( Z_G(M) \) is relatively compact. Since \( g \) is bounded below on \( \mathcal{X} \), \( G \) is bounded below on \( \mathcal{P}(\mathcal{X}) \). It follows that \( G \) is a tightness function on \( \mathcal{P}(\mathcal{X}) \).

The next result in this section is a continuity property needed in Chapter 8.

**Theorem A.3.18.** Let \( \mathcal{X} \) and \( \mathcal{Y} \) be Polish spaces and \( f \in \mathcal{C}_b(\mathcal{X} \times \mathcal{Y}) \). Then the function mapping

\[
(\xi, \gamma) \in \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \mapsto \int_{\mathcal{Y}} f(\xi, b) \gamma(db)
\]

is bounded and continuous.

**Proof.** Let \( \{ (\xi_n, \gamma_n), n \in \mathbb{N} \} \) be a sequence in \( \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \) converging to \( (\xi, \gamma) \in \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \). For any \( \varepsilon > 0 \) there exists a compact subset \( K \) of \( \mathcal{Y} \) such that \( \text{sup}_{n \in \mathbb{N}} \gamma_n(K^c) \leq \varepsilon \) [Theorem A.3.15]. By the continuity of \( f \) and the weak convergence \( \gamma_n \Rightarrow \gamma \), there exists \( N \in \mathbb{N} \) such that for all \( n \geq N \)

\[
\left| \int_{\mathcal{Y}} f(\xi, b) \gamma(db) - \int_{\mathcal{Y}} f(\xi, b) \gamma_n(db) \right| \leq \varepsilon \quad \text{and} \quad \sup_{b \in K} |f(\xi, b) - f(\xi, b_n)| \leq \varepsilon
\]

Hence for all \( n \geq N \)

\[
\begin{align*}
\left| \int_{\mathcal{Y}} f(\xi, b) \gamma(db) - \int_{\mathcal{Y}} f(\xi, b) \gamma_n(db) \right| &\leq \left| \int_{\mathcal{Y}} f(\xi, b) \gamma(db) - \int_{\mathcal{Y}} f(\xi, b) \gamma_n(db) \right| + \left| \int_{\mathcal{Y}} f(\xi, b) \gamma_n(db) - \int_{\mathcal{Y}} f(\xi, b) \gamma_n(db) \right| \\
&\leq \varepsilon + 2\|f\|_{\infty} \gamma_n(K^c) + \sup_{b \in K} |f(\xi, b) - f(\xi_n, b)| \\
&\leq 2\|f\|_{\infty} \varepsilon + 2\varepsilon.
\end{align*}
\]

This proves the continuity of the function defined in (A.2). Since its boundedness is obvious, the proof is complete.

The final three theorems in this section are specialized to \( \mathbb{R}^d \). A sequence \( \{\theta_n, n \in \mathbb{N}\} \) in \( \mathcal{P}(\mathbb{R}^d) \) is said to be **uniformly integrable** if

\[
\lim_{C \to \infty} \lim_{n \to \infty} \int_{\{x \in \mathbb{R}^d : \|x\| > C\}} \|x\| \theta_n(dx) = 0.
\]

According to part (a) of the next result, the sequence of means of a uniformly integrable, weakly convergent sequence \( \{\theta_n\} \) converges. Part (b) gives a criterion for uniform integrability. A measurable function \( f \) mapping \( \mathbb{R}^d \) into \( \mathbb{R} \cup \{\infty\} \) is said to be **superlinear** if

\[
\lim_{N \to \infty} \inf_{\{x \in \mathbb{R}^d : \|x\| = N\}} \frac{1}{\|x\|} f(x) = \infty.
\]
Theorem A.3.19. Let \( \{\theta_n\} \) be a sequence in \( \mathcal{P}(\mathbb{R}^d) \) converging weakly to \( \theta \). The following implications hold.

(a)\citepage[Page 32]{9} If \( \{\theta_n\} \) is uniformly integrable, then

\[
\int_{\mathbb{R}^d} \|x\| \theta(dx) < \infty \quad \text{and} \quad \lim_{n \to \infty} \int_{\mathbb{R}^d} x \theta_n(dx) = \int_{\mathbb{R}^d} x \theta(dx).
\]

(b) If there exists a superlinear function \( f \) mapping \( \mathbb{R}^d \) into \([0, \infty)\) such that

\[
\sup_{n \in \mathbb{N}} \int_{\mathbb{R}^d} f d\theta_n < \infty,
\]

then \( \{\theta_n\} \) is uniformly integrable.

Several times in the text we have appealed to the **continuity theorem for moment generating functions**, which gives a criterion for the weak convergence of a sequence of probability measures. This theorem is stated next. The proof of part (b) relies on the uniqueness theorem for moment generating functions, which is stated afterwards.

Theorem A.3.20. Let \( \{\theta_n, n \in \mathbb{N}\} \) be a sequence in \( \mathcal{P}(\mathbb{R}^d) \). Assume that for all \( n \in \mathbb{N} \) and \( \alpha \in \mathbb{R}^d \) the moment generating functions

\[
g_n(\alpha) = \int_{\mathbb{R}^d} \exp\langle \alpha, y \rangle \theta_n(dy)
\]

are finite. The following conclusions hold.

(a) If for all \( \alpha \in \mathbb{R}^d \)

\[
g(\alpha) = \lim_{n \to \infty} g_n(\alpha)
\]

exists and \( g \) is continuous at the origin, then there exists \( \theta \in \mathcal{P}(\mathbb{R}^d) \) such that \( \theta_n \to \theta \) and for all \( \alpha \in \mathbb{R}^d \)

\[
g(\alpha) = \int_{\mathbb{R}^d} \exp\langle \alpha, y \rangle \theta(dy).
\]

(b) Let \( \mu \in \mathcal{P}(\mathbb{R}^d) \) have the property that for all \( \alpha \in \mathbb{R}^d \)

\[
h(\alpha) = \int_{\mathbb{R}^d} \exp\langle \alpha, y \rangle \mu(dy)
\]

is finite. If for all \( \alpha \in \mathbb{R}^d \lim_{n \to \infty} g_n(\alpha) = h(\alpha) \), then \( \theta_n \to \mu \).

**Proof.** (a) This is proved in Appendix C of \cite{70}.

(b) By the Lebesgue Dominated Convergence Theorem \( h \) is continuous at the origin. Hence by part (a) there exists \( \theta \in \mathcal{P}(\mathbb{R}^d) \) such that \( \theta_n \to \theta \) and for all \( \alpha \in \mathbb{R}^d \)

\[
h(\alpha) = \int_{\mathbb{R}^d} \exp\langle \alpha, y \rangle \theta(dy).
\]

Since \( \mu \) and \( \theta \) have the same moment generating function \( h \), Theorem A.3.21 implies that \( \theta = \mu \). Thus \( \theta_n \to \mu \). \( \blacksquare \)

We now state the **uniqueness theorem for moment generating functions**.
**Theorem A.3.21.** Let $\mu$ and $\theta$ in $\mathcal{P}(\mathbb{R}^d)$ have the property that for all $\alpha \in \mathbb{R}^d$

$$\int_{\mathbb{R}^d} \exp\langle \alpha, y \rangle \mu(dy) = \int_{\mathbb{R}^d} \exp\langle \alpha, y \rangle \theta(dy) < \infty.$$ 

Then $\mu = \theta$.

**Proof.** Let $X$ and $Y$ be random variables having the respective distributions $\mu$ and $\theta$. For any $\alpha \in \mathbb{R}^d$ the functions mapping

$$z \in \mathcal{C} \mapsto E\{e^{z\langle \alpha, X \rangle}\} \text{ and } z \in \mathcal{C} \mapsto E\{e^{z\langle \alpha, Y \rangle}\}$$

are analytic and they agree when $z$ is real. Hence for all $z \in \mathcal{C}$ $E\{e^{z\langle \alpha, X \rangle}\} = E\{e^{z\langle \alpha, Y \rangle}\}$. It follows that $X$ and $Y$ have the same characteristic function. Since random variables having the same characteristic function have the same distribution [9, Thm. 7.5], the proof is done. 

In Lemma 5.3.6 we consider a sequence of stochastic processes $\{S^n, n \in \mathbb{N}\}$ taking values in $C([0, 1] : \mathbb{R}^d)$. The verification that the sequence of distributions of these processes is tight relies on the following theorem. The proof of the theorem uses the Arzéla–Ascoli Theorem [Theorem A.6.2], which characterizes relatively compact subsets of $C([0, 1] : \mathbb{R}^d)$. For $\delta \in (0, 1]$ the modulus of continuity of $\varphi \in C([0, 1] : \mathbb{R}^d)$ is defined by

$$w_\varphi(\delta) = \sup_{\{s, t \in [0, 1] | |s - t| \leq \delta\}} \|\varphi(s) - \varphi(t)\|.$$ 

**Theorem A.3.22 [9, page 55].** A sequence $\{\theta_n\}$ in $\mathcal{P}(C([0, 1] : \mathbb{R}^d))$ is tight if and only if the following two conditions hold:

(a) For each $\eta > 0$ there exists $b < \infty$ such that for all $n \in \mathbb{N}$

$$\theta_n\{\varphi : \|\varphi(0)\| > b\} \leq \eta.$$ 

(b) For each $\varepsilon > 0$ and $\eta > 0$ there exists $\delta \in (0, 1)$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$

$$\theta_n\{\varphi : w_\varphi(\delta) \geq \varepsilon\} \leq \eta.$$ 

### A.4 Probability Theory

Throughout this section $(\Omega, \mathcal{F}, P)$ is a probability space and $\mathcal{Y}$ is a Polish space. Let $Y$ be a random variable mapping $\Omega$ into $\mathcal{Y}$ and $\mathcal{G}$ a sub-$\sigma$-field of $\mathcal{F}$. A **regular conditional distribution for $Y$ given $\mathcal{G}$** is defined to be a quantity $\hat{P}(dy|\mathcal{G})(\omega)$ taking values in $[0, 1]$ and having the following properties:

(a) For each Borel subset $B$ of $\mathcal{Y}$ the function mapping $\omega \in \Omega \mapsto \hat{P}(B|\mathcal{G})(\omega)$ is measurable with respect to $\mathcal{G}$. 
(b) For each \( \Gamma \in \mathcal{G} \) and Borel subset \( B \) of \( \mathcal{Y} \)

\[
P\{\Gamma \cap \{Y \in B\}\} = \int_{\mathcal{Y}} \hat{P}(B|\mathcal{G})(\omega) \, P(d\omega).
\]

(c) For each \( \omega \in \Omega \), \( \hat{P}(dy|\mathcal{G})(\omega) \) is a probability measure on \( \mathcal{Y} \).

The first two properties state that \( \hat{P}(B|\mathcal{G})(\omega) \) is a version of the conditional probability \( P\{Y \in B|\mathcal{G}\}(\omega) \). The issue in proving the existence of a regular conditional distribution is to show that a version of the conditional probability can be chosen to be a probability measure on \( \mathcal{Y} \) for each fixed value of \( \omega \). According to part (a) of the following standard result, this is possible when \( \mathcal{Y} \) is a Polish space. Part (b) states the useful property that \( P \)-a.s. for \( \omega \in \Omega \) conditional expectations can be obtained by integrating with respect to regular conditional distributions. The theorem is proved in Theorems 10.2.2 and 10.2.5 in [36]. If \( X \) is a random variable mapping \( (\Omega, \mathcal{F}) \) into a measurable space \( (\mathcal{Y}, \mathcal{A}) \) and \( \mathcal{G} \) denotes the sub-\( \sigma \)-field of \( \mathcal{F} \) generated by \( X \), then we will write \( \hat{P}(dy|X)(\omega) \) and call it a regular conditional distribution for \( Y \) given \( X \).

**Theorem A.4.1.** Let \( Y \) be a random variable mapping \( \Omega \) into \( \mathcal{Y} \), \( \mathcal{G} \) a sub-\( \sigma \)-field of \( \mathcal{F} \), and \( f \) a measurable function mapping \( \mathcal{Y} \) into \( \mathbb{R} \) such that \( E\{|f|\} < \infty \). The following conclusions hold.

(a) A regular conditional distribution \( \hat{P}(dy|\mathcal{G})(\omega) \) for \( Y \) given \( \mathcal{G} \) exists. It is unique in the sense that if \( \tilde{Q}(dy|\mathcal{G})(\omega) \) also satisfies the definition, then \( P \)-a.s. for \( \omega \in \Omega \) the two distributions \( \hat{P}(dy|\mathcal{G})(\omega) \) and \( \tilde{Q}(dy|\mathcal{G})(\omega) \) agree.

(b) \( P \)-a.s. for \( \omega \in \Omega \) \( f \) is integrable with respect to \( \hat{P}(dy|\mathcal{G})(\omega) \) and

\[
E\{f(Y)|\mathcal{G}\}(\omega) = \int_{\mathcal{Y}} f(y) \, \hat{P}(dy|\mathcal{G})(\omega).
\]

Now let \( (\mathcal{Y}, \mathcal{A}) \) be a measurable space and \( X \) a random variable mapping \( \Omega \) into \( \mathcal{Y} \). A regular conditional distribution for \( Y \) given \( X = x \) is defined to be a quantity \( \hat{P}(dy|X = x) \) taking values in \([0, 1]\) and having the following properties:

(a) For each Borel subset \( B \) of \( \mathcal{Y} \) the function mapping \( x \in \mathcal{X} \mapsto \hat{P}(B|X = x) \) is measurable.

(b) For each measurable subset \( A \) of \( (\mathcal{Y}, \mathcal{A}) \) and Borel subset \( B \) of \( \mathcal{Y} \)

\[
P\{\{X \in A\} \cap \{Y \in B\}\} = \int_A \hat{P}(B|X = x) \, P\{X \in dx\}.
\]

(c) For each \( x \in \mathcal{X} \), \( \hat{P}(dy|X = x) \) is a probability measure on \( \mathcal{Y} \).

The first two properties state that \( \hat{P}(B|X = x) \) is a version of the conditional probability \( P\{Y \in B|X = x\} \). The next result is an immediate consequence of Theorem A.4.1.

**Theorem A.4.2.** Let \( (\mathcal{Y}, \mathcal{A}) \) be a measurable space, \( X \) a random variable mapping \( \Omega \) into \( \mathcal{Y} \), and \( Y \) a random variable mapping \( \Omega \) into \( \mathcal{Y} \). Then a regular conditional distribution \( \hat{P}(dy|X = x) \) for \( Y \) given \( X = x \) exists. It is unique in the sense that if \( \tilde{Q}(dy|X = x) \) also satisfies the definition, then for almost every \( x \) with respect to the distribution of \( X \) the two measures \( \hat{P}(dy|X = x) \) and \( \tilde{Q}(dy|X = x) \) agree.
Proof. A regular conditional distribution \( \hat{P}(dy|X) \) is measurable with respect to the sub-\( \sigma \)-field generated by \( X \) and so is a measurable function of \( X \), say \( \varphi(X) \). The quantity \( \varphi(x) \) is a regular conditional distribution for \( Y \) given \( X = x \). The uniqueness follows from the uniqueness asserted in part (a) of Theorem A.4.1. \( \blacksquare \)

The next theorem generalizes to the setting of a Polish space a standard fact concerning limits of random variables taking values in \( \mathbb{R} \). It will be used in the following section.

**Theorem A.4.3.** Let \( \{Y_n, n \in \mathbb{N}\} \) be a sequence of random variables mapping \( \Omega \) into \( \mathcal{Y} \). Assume that \( Y(\omega) \) \( = \) \( \lim_{n \to \infty} Y_n(\omega) \) exists for every \( \omega \in \Omega \). Then \( Y \) is a random variable mapping \( \Omega \) into \( \mathcal{Y} \).

**Proof.** It suffices to show that \( Y^{-1}(A) \) lies in \( \mathcal{F} \) for every open set \( A \) in \( \mathcal{Y} \). By Lemma 1.3.5 applied to the lower semicontinuous function mapping \( x \in \mathcal{Y} \mapsto 1_A(x) \), there exists a sequence \( \{g_j, j \in \mathbb{N}\} \subset C_0(\mathcal{Y}) \) satisfying \( g_j \uparrow 1_A \). Since for each \( j \in \mathbb{N} \) \( \lim_{n \to \infty} g_j \circ Y_n = g_j \circ Y \), the \( \mathbb{R} \)-valued mappings \( g_j \circ Y \) are random variables on \( \mathcal{Y} \). We can now conclude that \( Y^{-1}(A) \) lies in \( \mathcal{F} \) because

\[
Y^{-1}(A) = (1_A \circ Y)^{-1}((1/2, \infty)) = \bigcup_{j \in \mathbb{N}} (g_j \circ Y)^{-1}((1/2, \infty)).
\]

This completes the proof. \( \blacksquare \)

The ergodic theorem is applied in Sections 8.6 and 9.3 in the proofs of Laplace principle lower bounds. Let \( \{Y_j, j \in \mathbb{N}_0\} \) be a sequence of real-valued random variables defined on \( (\Omega, \mathcal{F}, P) \). The sequence is called **stationary** if for every Borel subset \( B \) of \( \mathbb{R}^{\mathbb{N}_0} \) and every \( k \in \mathbb{N}_0 \)

\[
P\{(Y_k, Y_{k+1}, \ldots) \in B\} = P\{(Y_0, Y_1, \ldots) \in B\}.
\]

An event \( A \in \mathcal{F} \) is called **invariant** if there exists a Borel subset \( B \in \mathbb{R}^{\mathbb{N}_0} \) such that for every \( n \in \mathbb{N}_0 \)

\[
A = \{(Y_n, Y_{n+1}, \ldots) \in B\}.
\]

A stationary sequence \( \{Y_j\} \) is called **ergodic** if every invariant event has probability 0 or 1.

We now assume that the sequence \( \{Y_j\} \) is stationary and ergodic. The next theorem states the **pointwise ergodic theorem** and the **\( L^1 \)-ergodic theorem** for such a sequence. These results are proved in Sections 5–7 of [14].

**Theorem A.4.4.** Let \( \{Y_j, j \in \mathbb{N}_0\} \) be a stationary, ergodic sequence taking values in \( \mathbb{R} \) and assume that \( E|Y_1| < \infty \). Then we have the probability–1 limit

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} Y_j = E\{Y_1\}
\]

and the \( L^1 \) limit

\[
\lim_{n \to \infty} E\left\{ \left| \frac{1}{n} \sum_{j=0}^{n-1} Y_j - E\{Y_1\} \right| \right\} = 0.
\]
A.5. STOCHASTIC KERNELS

For application in Section 8.6 a criterion is needed for the ergodicity of a Markov chain. Let \( \{Y_j, j \in \mathbb{N}_0\} \) be a sequence of random variables mapping \((\Omega, \mathcal{F}, P)\) into \(\mathcal{Y}\). \(\{Y_j\}\) is said to be a Markov chain if for each Borel subset \(A\) of \(\mathcal{Y}\) and \(j \in \mathbb{N}_0\)

\[
P\{Y_{j+1} \in A|Y_0, Y_1, \ldots, Y_j\} = P\{Y_{j+1} \in A|Y_j\} \text{ w.p.1.}
\]

The Markov chain is said to have stationary transition probabilities if for each Borel set \(A\) and point \(x \in \mathcal{Y}\) the conditional probabilities \(P\{Y_{j+1} \in A|Y_j = x\}\) are independent of \(j \in \mathbb{N}_0\). The transition probability function of such a Markov chain is defined as a quantity \(\pi(x, dy)\) which for each \(j \in \mathbb{N}_0\) is a regular conditional distribution for \(Y_{j+1}\) given \(Y_j = x\). In particular, \(\pi(x, dy)\) is a stochastic kernel on \(\mathcal{Y}\) given \(\mathcal{Y}\). The initial distribution is defined as the probability measure \(P\{Y_0 \in dy\}\). Similar concepts are defined if the index set \(\mathbb{N}_0\) is replaced by a finite index set \(\{0, 1, \ldots, n\}\).

Let \(\pi(x, dy)\) be a stochastic kernel on \(\mathcal{Y}\) given \(\mathcal{Y}\). An invariant measure for \(\pi(x, dy)\) is a probability measure \(\theta\) on \(\mathcal{Y}\) such that for all Borel sets \(A\)

\[
\theta(A) = \int_{\mathcal{Y}} \pi(x, A) \theta(dx).
\]

If \(\theta\) is an invariant measure for \(\pi(x, dy)\), then there exists a Markov chain \(\{Y_j, j \in \mathbb{N}_0\}\) that takes values in \(\mathcal{Y}\), has the transition probability function \(\pi(x, dy)\), and has the initial distribution \(\theta\). Furthermore, \(\{Y_j\}\) is stationary.

Let \(\{Y_j, j \in \mathbb{N}_0\}\) be a Markov chain that takes values in \(\mathcal{Y}\) and has the transition probability function \(\pi(x, dy)\). A Borel subset \(A\) of \(\mathcal{Y}\) is said to be closed for \(\pi(x, dy)\) if \(\pi(x, A) = 1\) for all \(x \in A\). The transition probability function \(\pi(x, dy)\) is said to be indecomposable if there are no two disjoint Borel sets \(A_1\) and \(A_2\) which are closed for \(\pi(x, dy)\). The following theorem, proved in Theorem 7.16 in [14] in the case \(\mathcal{Y} = \mathbb{R}\), is valid for an arbitrary Polish space with the same proof.

**Theorem A.4.5.** Let \(\pi(x, dy)\) be an indecomposable transition probability function on \(\mathcal{Y}\). Then if \(\pi(x, dy)\) has an invariant measure \(\theta\), \(\theta\) is unique and the Markov chain associated with \(\theta\) and \(\pi(x, dy)\) is stationary and ergodic.

A.5 Stochastic Kernels

Throughout this section \(\mathcal{X}\) and \(\mathcal{Y}\) are Polish spaces and \((\mathcal{Y}, \mathcal{A})\) is a measurable space. Let us recall the definition of a stochastic kernel, which was introduced in Section 1.4. Let \(\tau(dy|x)\) be a family of probability measures on \(\mathcal{Y}\) parametrized by \(x \in \mathcal{Y}\). We call \(\tau(dy|x)\) a stochastic kernel on \(\mathcal{Y}\) given \(\mathcal{Y}\) if for every Borel subset \(E\) of \(\mathcal{Y}\) the function mapping \(x \in \mathcal{Y} \mapsto \tau(E|x) \in [0, 1]\) is measurable.

In order to establish a useful equivalent condition for a stochastic kernel, we need a preliminary fact given in the next lemma. It is a special case of Proposition 7.25 in [8]. For \(\mathcal{Z}\) a Polish space, \(\mathcal{B}_\mathcal{Z}\) denotes the Borel \(\sigma\)-field.
Lemma A.5.1. For $E \in \mathcal{B}_Y$ define $f_E : \mathcal{P}(Y) \rightarrow [0, 1]$ by $f_E(\theta) = \theta(E)$. Then

$$\mathcal{B}_{\mathcal{P}(Y)} = \sigma\left[ \bigcup_{E \in \mathcal{B}_Y} f_E^{-1}(\mathcal{B}_R) \right].$$

In other words, $\mathcal{B}_{\mathcal{P}(Y)}$ is the smallest $\sigma$-field with respect to which $f_E$ is measurable for every $E \in \mathcal{B}_Y$.

Proof. We write $\Sigma = \sigma[\bigcup_{E \in \mathcal{B}_Y} f_E^{-1}(\mathcal{B}_R)]$. In order to prove that $\Sigma \subset \mathcal{B}_{\mathcal{P}(Y)}$, we show that $f_E$ is $\mathcal{B}_{\mathcal{P}(Y)}$-measurable for every $E \in \mathcal{B}_Y$; i.e., that $f_E^{-1}(\mathcal{B}_R) \subset \mathcal{B}_{\mathcal{P}(Y)}$. Let

$$\mathcal{D} = \{ E \in \mathcal{B}_Y : f_E \text{ is } \mathcal{B}_{\mathcal{P}(Y)} \text{-measurable} \}.$$

For any closed set $F \in \mathcal{B}_Y$ and real number $\alpha$, the Portmanteau Theorem implies that the set $\{ \theta \in \mathcal{P}(Y) : \theta(F) \geq \alpha \}$ is closed. Hence $F \in \mathcal{D}$. It is straightforward to verify that $\mathcal{D}$ is Dynkin class. Since $\mathcal{D}$ contains every closed set, the Dynkin Class Theorem [Theorem A.2.1] implies that $\mathcal{D}$ equals $\mathcal{B}_Y$ and thus that $\Sigma \subset \mathcal{B}_{\mathcal{P}(Y)}$.

The proof that $\mathcal{B}_{\mathcal{P}(Y)} \subset \Sigma$ is based on a standard approximation argument. By definition of $\Sigma$ the function

$$\alpha_\varphi : \theta \in \mathcal{P}(Y) \mapsto \int_Y \varphi \, d\theta \in \mathbb{R}$$

is $\Sigma$-measurable when $\varphi = 1_E$ for any $E \in \mathcal{B}_Y$; indeed, in this case $\alpha_\varphi(\theta) = f_E(\theta)$.

Thus $\alpha_\varphi$ is $\Sigma$-measurable when $\varphi$ is a $\mathcal{B}_Y$-simple function. Since when $\varphi \in \mathcal{C}_b(Y)$ there exists a sequence of $\mathcal{B}_Y$-simple functions $\{ \varphi_n \}$ that are uniformly bounded below and satisfy $\varphi_n \uparrow \varphi$, the Monotone Convergence Theorem implies that $\alpha_{\varphi_n} \uparrow \alpha_\varphi$. Thus $\alpha_\varphi$ is $\Sigma$-measurable for every $\varphi \in \mathcal{C}_b(Y)$. For $\gamma \in \mathcal{P}(Y)$, $\varphi \in \mathcal{C}_b(Y)$, and $\varepsilon > 0$ we define

$$N(\gamma, \varphi, \varepsilon) \doteq \left\{ \theta \in \mathcal{P}(Y) : \left| \int_Y \varphi \, d\theta - \int_Y \varphi \, d\gamma \right| < \varepsilon \right\}.$$

Since $N(\gamma, \varphi, \varepsilon) = \alpha_\varphi^{-1}(\int_Y \varphi \, d\gamma - \varepsilon, \int_Y \varphi \, d\gamma + \varepsilon)$, it follows that $N(\gamma, \varphi, \varepsilon)$ is an element of $\Sigma$, and since the class of sets $\{ N(\gamma, \varphi, \varepsilon) \}$ forms an open subbase for $\mathcal{B}_{\mathcal{P}(Y)}$, we conclude that $\mathcal{B}_{\mathcal{P}(Y)} \subset \Sigma$. This completes the proof of the lemma. $\blacksquare$

The following result, taken from Proposition 7.26 in [8], gives a useful equivalent condition for a stochastic kernel. In the latter reference it is assumed that $(Y, \mathcal{A})$ is a Borel space. However, the proof applies without change when $(V, \mathcal{A})$ is a measurable space.

Theorem A.5.2. Let $\tau(dy|x)$ be a family of probability measures on $Y$ parametrized by $x \in Y$. Then $\tau(dy|x)$ is a stochastic kernel if and only the function mapping $x \in V \mapsto \tau(\cdot|x) \in \mathcal{P}(Y)$ is measurable; i.e., if and only if $\tau(\cdot|x)$ is a random variable mapping $V$ into $\mathcal{P}(Y)$. 


A.5. STOCHASTIC KERNELS

Proof. We define $g : \mathcal{V} \mapsto \mathcal{P}(\mathcal{Y})$ by $g(x) = \tau(\cdot|x)$ and for $E \in \mathcal{B}_Y$ we define $h_E : \mathcal{V} \mapsto [0,1]$ by $h_E(x) = \tau(E|x)$. For $E \in \mathcal{B}_Y$ we also recall $f_E : \mathcal{P}(\mathcal{Y}) \mapsto [0,1]$ defined in the previous lemma by $f_E(\theta) = \theta(E)$. These mappings are related by $h_E = f_E \circ g$. The assertion of the theorem is that $g$ is $\mathcal{A}$-measurable if and only if $h_E$ is $\mathcal{A}$-measurable for every $E \in \mathcal{B}_Y$.

Lemma A.5.1 implies that $f_E$ is $\mathcal{B}_{\mathcal{P}(\mathcal{Y})}$-measurable for every $E \in \mathcal{B}_Y$. Since $h_E = f_E \circ g$, it follows that if $g$ is $\mathcal{A}$-measurable, then $h_E$ is $\mathcal{A}$-measurable for every $E \in \mathcal{B}_Y$. Conversely, if $h_E$ is $\mathcal{A}$-measurable for every $E \in \mathcal{B}_Y$, then again by Lemma A.5.1

$$g^{-1}(\mathcal{B}_{\mathcal{P}(\mathcal{Y})}) = g^{-1}\left(\sigma\left[\bigcup_{E \in \mathcal{B}_Y} f_E^{-1}(\mathcal{B}_R)\right]\right)$$

$$= \sigma\left[\bigcup_{E \in \mathcal{B}_Y} g^{-1}\left(f_E^{-1}(\mathcal{B}_R)\right)\right] = \sigma\left[\bigcup_{E \in \mathcal{B}_Y} h_E^{-1}(\mathcal{B}_R)\right] \subset \mathcal{A}.$$

We conclude that $g$ is $\mathcal{A}$-measurable. This completes the proof. ■

A stochastic kernel on $\mathcal{Y}$ given $\mathcal{V}$ is a random variable mapping $\mathcal{V}$ into $\mathcal{P}(\mathcal{Y})$ [Theorem A.5.2]. Since the limit of such a sequence of random variables is a random variable [Theorem A.4.3], it follows that the class of stochastic kernels is closed under weak limits. We record this observation in the next theorem.

Theorem A.5.3. Let $\{\tau_n(dy|x), n \in \mathbb{N}\}$ be a sequence of stochastic kernels on $\mathcal{Y}$ given $\mathcal{V}$ and $\tau(dy|x)$ a family of probability measures on $\mathcal{Y}$ parametrized by $x \in \mathcal{V}$. Assume that for each $x \in \mathcal{V}$, $\tau_n(\cdot|x) \rightharpoonup \tau(\cdot|x)$. Then $\tau(dy|x)$ is a stochastic kernel on $\mathcal{Y}$ given $\mathcal{V}$.

The next theorem shows that a probability measure on a product of spaces can be decomposed into its first marginal and a stochastic kernel. The proof relies on the existence of a regular conditional distribution. There is an analogous decomposition that expresses a probability measure on a product of spaces in terms of its second marginal. For simplicity, we refer in the text to the following theorem both for the decomposition in terms of the first marginal and for that in terms of the second marginal.

Theorem A.5.4. Let $\tau = \tau(dx \times dy)$ be a probability measure on $\mathcal{V} \times \mathcal{Y}$ with the product $\sigma$-field $\mathcal{A} \times \mathcal{B}_Y$. We denote by $\tau_1$ the first marginal of $\tau$, which is the probability measure on $\mathcal{V}$ defined by $\tau_1(A) = \tau(A \times \mathcal{Y})$ for $A \in \mathcal{A}$. Then there exists a stochastic kernel $\tau(dy|x)$ on $\mathcal{Y}$ given $\mathcal{V}$ such that

$$\tau(A \times B) = \int_{A \times B} \tau_1(dx) \tau(dy|x) = \int_A \tau(B|x) \tau_1(dx)$$

for all $A \in \mathcal{A}$ and $B \in \mathcal{B}_Y$. This decomposition is summarized as $\tau(dx \times dy) = \tau_1(dx) \otimes \tau(dy|x)$.

Proof. On the probability space $(\mathcal{V} \times \mathcal{Y}, \mathcal{A} \times \mathcal{B}_Y, \tau)$ we define the coordinate functions $\hat{X}(x,y) = x$ and $\hat{Y}(x,y) = y$ for $(x,y) \in \mathcal{V} \times \mathcal{Y}$. The required stochastic kernel $\tau(dy|x)$ is then taken to be a regular conditional distribution for $\hat{Y}$ given $\hat{X} = x$. ■
Theorem A.5.4 was just proved using regular conditional distributions. Conversely, given the decomposition of \( \tau \) in this theorem, one can deduce the existence of regular conditional distributions. To see this, let \( X \) and \( Y \) be random variables taking values in \( \mathcal{V} \) and \( \mathcal{Y} \), respectively, and define \( \tau \) to be the joint distribution of \((X, Y)\). Then the stochastic kernel \( \tau(dy|x) \) appearing in the decomposition \( \tau(dx \times dy) = \tau_1(dx) \otimes \tau(dy|x) \) is a regular conditional distribution for \( Y \) given \( X = x \).

It is tempting to try to prove the theorem by a purely measure theoretic argument. Denote by \( \tau_1 \) and \( \tau_2 \) the first and second marginals of \( \tau \) and by \( \tau_1 \times \tau_2 \) the corresponding product measure on \( \mathcal{V} \times \mathcal{Y} \). Although for any measurable rectangle \( R \) \( (\tau_1 \times \tau_2)(R) = 0 \) implies that \( \tau(R) = 0 \), in general \( \tau \) is not absolutely continuous with respect to \( \tau_1 \times \tau_2 \). However, if in a specific case this absolute continuity does hold, then the stochastic kernel \( \tau(dy|x) \) in the statement of the theorem is given by

\[
\tau(B|x) = \int_B \frac{d\tau(x, y)}{d(\tau_1 \times \tau_2)}(x, y) \tau_2(dy).
\]

If in the previous theorem \( \mathcal{V} \) and \( \mathcal{Y} \) coincide, then \( \tau(dy|x) \) is a transition probability function on \( \mathcal{Y} \), and the theorem takes the following form.

**Corollary A.5.5** Let \( \tau = \tau(dx \times dy) \) be a probability measure on \( \mathcal{V} \times \mathcal{Y} \). Then there exists a transition probability function \( q_\tau(x, dy) \) on \( \mathcal{Y} \) such that

\[
\tau(A \times B) = \int_{A \times B} \tau_1(dx) q_\tau(x, dy) = \int_A q_\tau(x, B) \tau_1(dx)
\]

for all Borel subsets \( A \) and \( B \) of \( \mathcal{Y} \). This formula is summarized as \( \tau(dx \times dy) = \tau_1(dx) \otimes q_\tau(x, dy) \).

Let \( \mathcal{X} \) and \( \mathcal{Y} \) be Polish spaces and \((\mathcal{V}, \mathcal{A})\) a measurable space. The decomposition, given in Theorem A.5.4, of a probability measure on \( \mathcal{X} \times \mathcal{Y} \) generalizes to the context of stochastic kernels for which a measurable dependence on a parameter \( z \in \mathcal{V} \) is present. The resulting decomposition is stated in the next theorem. Almost surely with respect to a given measure on \((\mathcal{V}, \mathcal{A})\), the theorem expresses a stochastic kernel on \( \mathcal{X} \times \mathcal{Y} \) given \( \mathcal{V} \) in terms of its first marginal. An analogous decomposition expresses a stochastic kernel on \( \mathcal{X} \times \mathcal{Y} \) given \( \mathcal{V} \) in terms of its second marginal. For simplicity, in the text we refer to the following theorem for both decompositions.

**Theorem A.5.6.** Let \( \sigma(dx \times dy|z) \) be a stochastic kernel on \( \mathcal{X} \times \mathcal{Y} \) given \( \mathcal{V} \) and \( \theta \) a probability measure on \((\mathcal{V}, \mathcal{A})\). We denote by \( \sigma_1(dx|z) \) the first marginal of \( \sigma(dx \times dy|z) \), which is the stochastic kernel on \( \mathcal{X} \) given \( \mathcal{V} \) defined by \( \sigma_1(A|z) = \sigma(A \times \mathcal{Y}|z) \) for Borel sets \( A \) and \( z \in \mathcal{V} \). Then there exists a stochastic kernel \( \sigma(dy|x, z) \) on \( \mathcal{Y} \) given \( \mathcal{X} \times \mathcal{V} \) such that \( \theta \)-a.s. for \( z \in \mathcal{V} \)

\[
\sigma(A \times B|z) = \int_{A \times B} \sigma_1(dx|z) \sigma(dy|x, z) = \int_A \sigma(B|x, z) \sigma_1(dx|z)
\]

for all Borel subsets \( A \) of \( \mathcal{X} \) and \( B \) of \( \mathcal{Y} \). This decomposition is summarized as \( \sigma(dx \times dy|z) = \sigma_1(dx|z) \otimes \sigma(dy|x, z) \).
A.5. STOCHASTIC KERNELS

Proof. Let \( \sigma = \sigma(dx \times dy \times dz) \) be the probability measure on \((X \times Y \times V, B_X \times B_Y \times A)\) whose value on measurable rectangles \( A \times B \times C \) is given by

\[
\sigma(A \times B \times C) = \int_{X \times B \times C} \sigma(dx \times dy | z) \theta(dz).
\]

On this product space we also define the coordinate functions \( \hat{X}(x, y, z) \equiv x, \hat{Y}(x, y, z) \equiv y, \) and \( \hat{Z}(x, y, z) \equiv z \) together with the marginal \( \sigma_{1,3}(dx \times dz) \) of \( \sigma \) obtained by projecting onto the first and third coordinates. Since the distribution of \( \hat{Z} \) is \( \theta \), \( \sigma(dx \times dy | z) \) is a regular conditional distribution for \((\hat{X}, \hat{Y})\) given \( \hat{Z} = z \). We prove the decomposition stated in the theorem by taking \( \sigma(dy|x, z) \) to be a regular conditional distribution for \( \hat{Y} \) given \((\hat{X}, \hat{Z}) = (x, z)\) and showing that \( \theta \)-a.s. for \( z \in V \), \( \sigma_1(dx | z) \otimes \sigma(dy | x, z) \) is a version of \( \sigma(dx \times dy | z) \). This entails showing that for all \( A \in B_X, B \in B_Y, \) and \( C \in A \)

\[
\sigma(A \times B \times C) = \int_C \left( \int_{X \times B} \sigma_1(dx | z) \sigma(dy | x, z) \right) \theta(dz).
\]

Since \( \sigma_1(dx | z) \) is a regular conditional distribution for \( \hat{X} \) given \( \hat{Z} = z \), \( \sigma_1(dx | z) \otimes \theta(dz) = \sigma_{1,3}(dx \times dz) \). Hence

\[
\int_C \left( \int_{X \times B} \sigma_1(dx | z) \sigma(dy | x, z) \right) \theta(dz) = \int_{X \times C} \sigma(B|x, z) \sigma_1(dx | z) \otimes \theta(dz)
\]

\[
= \int_{X \times C} \sigma(B|x, z) \sigma_{1,3}(dx \times dz) = \sigma(A \times B \times C).
\]

This completes the proof. ■

The almost sure decomposition given in Theorem A.5.6 suffices for all our applications. Proposition 7.27 in [8] proves a stronger result; namely, that there exists a stochastic kernel \( \sigma(dy|x, z) \) on \( Y \) given \( X \times V \) such that \( \sigma(dx \times dy | z) = \sigma_1(dx | z) \otimes \sigma(dy | x, z) \) is valid for all \( z \in V \).

Let \( \sigma(dy|x) \) and \( \tau(dy|x) \) be stochastic kernels on \( Y \) given \( X \) with the property that for each \( x \) in some measurable subset \( A \) of \((V, A)\) the probability measure \( \sigma(dy|x) \) is absolutely continuous with respect to \( \tau(dy|x) \). A basic question is the existence of a Radon–Nikodym derivative which is a jointly measurable function of \((x, y) \in A \times Y \). This is settled by the following result. It is an immediate consequence of Theorem V.58 in [27], which is proved by applying the martingale convergence theorem in the context of a separable measurable space \((Z, G)\). By definition a measurable space \((Z, G)\) is separable if there exists a sequence of sets \( \{A_n, n \in N\} \) so that \( G = \sigma(\{A_n\}) \). A Polish space with the Borel \( \sigma \)-field is an example of a separable measurable space.

Theorem A.5.7. Let \( A \) be a measurable subset of \((V, A)\). Let \( \sigma(dy|x) \) and \( \tau(dy|x) \) be stochastic kernels on \( Y \) given \( V \) with the property that for each \( x \in A \) the probability measure \( \sigma(dy|x) \) is absolutely continuous with respect to \( \tau(dy|x) \). Then there exists a version of the Radon–Nikodym derivative

\[
f(x, y) \doteq \frac{d\sigma(\cdot|x)}{d\tau(\cdot|x)}(y).
\]

which is a nonnegative measurable function of \((x, y) \in A \times Y \).
The next theorem concerns the weak convergence of sequences of measures constructed from sequences of stochastic kernels. It is needed in order to prove the Laplace principle upper bound in Chapter 6.

**Theorem A.5.8.** Let $\Delta$ and $\mathcal{Y}$ be Polish spaces, $\theta$ a probability measure on $\Delta$, and $\{\mu_n(dy|t), n \in \mathbb{N}\}$ and $\mu(dy|t)$ stochastic kernels on $\mathcal{Y}$ given $\Delta$. Assume that $\mu_n(\cdot|t) \Rightarrow \mu(\cdot|t)$ for each $t \in \Delta$. Then

$$\mu_n(dy|t) \otimes \theta(dt) \Rightarrow \mu(dy|t) \otimes \theta(dt).$$

**Proof.** By Theorem A.3.14 one can prove the weak convergence by considering integrals with respect to $g(t)h(y)$ for $g \in C_b(\Delta)$ and $h \in C_b(\mathcal{Y})$. The Lebesgue Dominated Convergence Theorem yields

$$\lim_{n \to \infty} \int_{\mathcal{Y} \times \Delta} h(y) g(t) \mu_n(dy|t) \otimes \theta(dt) = \lim_{n \to \infty} \int_{\Delta} \left( \int_{\mathcal{Y}} h(y) \mu_n(dy|t) \right) g(t) \theta(dt)$$

$$= \int_{\Delta} \left( \int_{\mathcal{Y}} h(y) \mu(dy|t) \right) g(t) \theta(dt)$$

$$= \int_{\mathcal{Y} \times \Delta} h(y) g(t) \mu(dy|t) \otimes \theta(dt).$$

This completes the proof. $\blacksquare$

In order to prove the Laplace principle upper bound in Chapter 7, we need a variation on the previous theorem which allows us to deal with the more complicated limit arising there. We first derive a general result.

**Theorem A.5.9.** Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, $\{\theta_n, n \in \mathbb{N}\}$ and $\theta$ probability measures on $\mathcal{X}$, and $\{\mu_n(dy|x), n \in \mathbb{N}\}$ and $\mu(dy|x)$ stochastic kernels on $\mathcal{Y}$ given $\mathcal{X}$. Assume that $\theta_n \Rightarrow \theta$, that the function mapping $x \in \mathcal{X} \mapsto \mu(dy|x) \in \mathcal{P}(\mathcal{Y})$ is continuous with respect to the weak topology, and that for each compact subset $K$ of $\mathcal{X}$ and each $h \in C_b(\mathcal{Y})$

$$\lim_{n \to \infty} \sup_{x \in K} \left| \int_{\mathcal{Y}} h(y) \mu(dy|x) - \int_{\mathcal{Y}} h(y) \mu_n(dy|x) \right| = 0.$$

Then

$$\mu_n(dy|x) \otimes \theta_n(dx) \Rightarrow \mu(dy|x) \otimes \theta(dx).$$

**Proof.** Again by Theorem A.3.14 one can prove the weak convergence by considering integrals with respect to $g(x)h(y)$ for $g \in C_b(\mathcal{X})$ and $h \in C_b(\mathcal{Y})$. Let $\varepsilon > 0$ be given. According to Prohorov’s Theorem, there exists a compact subset $K$ of $\mathcal{X}$ such that $\sup_{n \in \mathbb{N}} \theta_n(K^c) \leq \varepsilon$. The continuity of the function mapping $x \in \mathcal{X} \mapsto \mu(dy|x) \in \mathcal{P}(\mathcal{Y})$ implies that the function mapping

$$x \in \mathcal{X} \mapsto \int_{\mathcal{Y}} h(y) \mu(dy|x) \in \mathbb{R}$$
is continuous. Hence by the weak convergence \( \theta_n \Rightarrow \theta \) and the hypothesis on the stochastic kernels \( \mu_n(dy|x) \), there exists \( N \in \mathbb{N} \) such that for all \( n \geq N \)

\[
\left| \int_X \left( \int_Y h(y) \mu(dy|x) \right) g(x) \theta(dx) - \int_X \left( \int_Y h(y) \mu(dy|x) \right) g(x) \theta_n(dx) \right| \leq \varepsilon
\]

and

\[
\sup_{x \in K} \left| \int_Y h(y) \mu_n(dy|x) - \int_Y h(y) \mu_n(dy|x) \right| \leq \varepsilon.
\]

Hence for all \( n \geq N \)

\[
\left| \int_{Y \times X} h(y) g(x) \mu(dy|x) \otimes \theta(dx) - \int_{Y \times X} h(y) g(x) \mu_n(dy|x) \otimes \theta_n(dx) \right|
\]

\[
\leq \left| \int_X \left( \int_Y h(y) \mu(dy|x) \right) g(x) \theta(dx) - \int_X \left( \int_Y h(y) \mu(dy|x) \right) g(x) \theta_n(dx) \right|
\]

\[
+ \left| \int_X \left( \int_Y h(y) \mu(dy|x) \right) g(x) \theta_n(dx) - \int_X \left( \int_Y h(y) \mu_n(dy|x) \right) g(x) \theta_n(dx) \right|
\]

\[
\leq \varepsilon + 2 \| g \|_{\infty} \| h \|_{\infty} \varepsilon \leq \varepsilon + 2 \| g \|_{\infty} \| h \|_{\infty} \varepsilon + \| g \|_{\infty} \varepsilon.
\]

Since \( \varepsilon > 0 \) is arbitrary, the proof is complete. □

We now derive what is needed for the proof of the Laplace principle upper bound in Chapter 7.

**Corollary A.5.10.** Let \( \Delta \) be a compact subset of \( \mathbb{R}^d \), \( \{ \theta_n, n \in \mathbb{N} \} \) and \( \theta \) probability measures on \( \Delta \), \( \mu(dy|x) \) a stochastic kernel on \( \mathbb{R}^d \) given \( \mathbb{R}^d \), \( \{ \varphi_n, n \in \mathbb{N} \} \) a sequence of bounded measurable functions mapping \( \Delta \) into \( \mathbb{R}^d \), and \( \varphi \) a bounded continuous function mapping \( \Delta \) into \( \mathbb{R}^d \). Assume that \( \theta_n \Rightarrow \theta \), that the function mapping \( x \in \mathbb{R}^d \mapsto \mu(dy|x) \in \mathcal{P}(\mathbb{R}^d) \) is continuous with respect to the weak topology, and that the sequence \( \{ \varphi_n, n \in \mathbb{N} \} \) converges to \( \varphi \) uniformly on \( \Delta \). Then

\[
\mu(dy|\varphi_n(t)) \otimes \theta_n(dt) \Rightarrow \mu(dy|\varphi(t)) \otimes \theta(dt).
\]

**Proof.** This result follows from the previous theorem once we verify that for any \( h \in C_b(\Delta) \)

\[
\lim_{n \to \infty} \sup_{t \in \Delta} \left| \int_{\mathbb{R}^d} h(y) \mu(dy|\varphi(t)) - \int_{\mathbb{R}^d} h(y) \mu(dy|\varphi_n(t)) \right| = 0.
\]

By hypothesis on \( \mu(dy|x) \), given \( \varepsilon > 0 \) and any compact subset \( \Lambda \) of \( \mathbb{R}^d \) there exists \( \delta > 0 \) such that

\[
\sup \left\{ \left| \int_{\mathbb{R}^d} h(y) \mu(dy|z) - \int_{\mathbb{R}^d} h(y) \mu(dy|z) \right| : \zeta \in \Lambda, z \in \Lambda, \| \zeta - z \| \leq \delta \right\} \leq \varepsilon.
\]

We now choose \( N \in \mathbb{N} \) such that \( \sup_{t \in \Delta} \| \varphi(t) - \varphi_n(t) \| \leq \delta \) for all \( n \geq N \) and define \( \Lambda \) to be the compact set \( \{ \alpha \in \mathbb{R}^d : \| \alpha - \beta \| \leq \delta \text{ for some } \beta \in \varphi(\Delta) \} \). Then for all \( n \geq N \)

\[
\sup_{t \in \Delta} \left| \int_{\mathbb{R}^d} h(y) \mu(dy|\varphi(t)) - \int_{\mathbb{R}^d} h(y) \mu(dy|\varphi_n(t)) \right| \leq \varepsilon.
\]

Since \( \varepsilon > 0 \) is arbitrary, the proof is complete. □
A.6 Analysis

Let $\mathcal{X}$ be a Polish space and let $d_1$ and $d_2$ be two metrics on $\mathcal{X}$. For $i = 1, 2$, $x \in \mathcal{X}$, and $\varepsilon > 0$, we define the open balls $B_i(x, \varepsilon) = \{ y \in \mathcal{X} : d_i(y, x) < \varepsilon \}$. The metrics $d_1$ and $d_2$ are said to be equivalent if for each $x \in \mathcal{X}$ and $\varepsilon > 0$ there exists $\delta > 0$ so that $B_1(x, \delta) \subseteq B_2(x, \varepsilon)$ and $B_2(x, \delta) \subseteq B_1(x, \varepsilon)$. Thus, if the metrics $d_1$ and $d_2$ are equivalent, then $(\mathcal{X}, d_1)$ is homeomorphic to $(\mathcal{X}, d_2)$.

The following separability result is applied numerous times in the text. It is proved in Lemma 3.1.4 in [85]. The proof depends on the fact that a separable metric space admits an equivalent metric with respect to which it is totally bounded.

**Theorem A.6.1.** Let $\mathcal{X}$ be a Polish space. Then $\mathcal{X}$ admits an equivalent metric $m(x, y)$ with the property that the space $\mathcal{U}_b(\mathcal{X}, m)$ of bounded, uniformly continuous functions mapping $\mathcal{X}$ into $\mathbb{R}$ is separable with respect to the uniform metric.

The Arzelà–Ascoli Theorem, stated next, characterizes relatively compact subsets of $C([0, 1] : \mathbb{R}^d)$. A subset $\Gamma$ of $C([0, 1] : \mathbb{R}^d)$ is said to be uniformly equicontinuous if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $|s - t| \leq \delta$ implies $||\varphi(s) - \varphi(t)|| \leq \varepsilon$ for all $s$ and $t$ in $[0, 1]$ and all $\varphi \in \Gamma$.

**Theorem A.6.2** [9, page 221]. A subset $\Gamma$ of $C([0, 1] : \mathbb{R}^d)$ is relatively compact if and only if $\sup \{ ||\varphi(0)|| : \varphi \in \Gamma \} < \infty$ and $\Gamma$ is uniformly equicontinuous.

The following fact about absolutely continuous functions can be proved as outlined in Problem 11 on pages 334–335 of [50]. We present a more elementary proof.

**Theorem A.6.3.** Let $g$ be an absolutely continuous function mapping $[0, 1]$ into $\mathbb{R}$. Then for each real number $\alpha$ the set $\{ s \in [0, 1] : g(s) = \alpha, g(s) \neq 0 \}$ has Lebesgue measure 0.

**Proof.** We define $B = \{ s \in [0, 1] : g(s) = \alpha, g(s) \neq 0 \}$ and note that $B \subseteq \bigcup_{n \in \mathbb{N}} C_n$, where $C_n$ denotes the set of $s \in [0, 1]$ such that $g(s) = \alpha$ and $|g(t) - g(s)| \geq |t - s|/n$ for all $t \in [0, 1]$ satisfying $|t - s| < 1/n$. If $s$ lies in $C_n$, then no point $t \in [0, 1]$ satisfying $|t - s| < 1/n$ can lie in $C_n$ since $|g(t) - g(s)| = |g(t) - g(s)| \geq |t - s|/n > 0$. Thus each $C_n$ contains only finitely many points. Since $B$ is a subset of a countable set, its Lebesgue measure equals 0. □

We next prove Gronwall's Inequality.

**Theorem A.6.4.** Take $T \in [0, \infty)$. Let $A$ and $B$ be positive constants and let $f$ be a bounded measurable function on $[0, T]$ that satisfies for all $t \in [0, T]$

$$0 \leq f(t) \leq A + B \int_0^t f(s) \, ds. \quad (A.3)$$

Then for all $t \in [0, T]$

$$0 \leq f(t) \leq Ae^{BR} \quad \text{and} \quad 0 \leq \sup_{t \in [0,T]} f(t) \leq Ae^{BT}.$$
A.6. ANALYSIS

Proof. For $t \in [0, T]$ iterating equation (A.3) yields

$$
0 \leq f(t) \leq A + A \sum_{j=1}^{\infty} \int_{0}^{s_{1}} \int_{0}^{s_{2}} \cdots \int_{0}^{s_{j-1}} B^{j} ds_{j} \cdots ds_{2} ds_{1}
= A + A \sum_{j=1}^{\infty} \frac{1}{j!} B^{j} = Ae^{Bt}.
$$

This implies that $0 \leq \sup_{t \in [0, T]} f(t) \leq Ae^{Bt}$. □

$\mathcal{D}([0, 1] : \mathbb{R}^{d})$ is defined to be the space of functions that map $[0, 1]$ into $\mathbb{R}^{d}$, are right continuous, and have left-hand limits. Any function in $\mathcal{D}([0, 1] : \mathbb{R}^{d})$ has at most countably many discontinuities and is bounded [9, page 110]. In order to metrize $\mathcal{D}([0, 1] : \mathbb{R}^{d})$, we let $\Lambda$ denote the class of strictly increasing, continuous mappings of $[0, 1]$ into itself. For $f$ and $g$ in $\mathcal{D}([0, 1] : \mathbb{R}^{d})$, we define $\rho(f, g)$ to be the infimum of those positive numbers $b$ for which there exists $\lambda \in \Lambda$ such that

$$
\sup_{0 \leq s \neq t \leq 1} \left| \log \left( \frac{\lambda(t) - \lambda(s)}{t - s} \right) \right| \leq b \quad \text{and} \quad \sup_{t \in [0, 1]} \| f(t) - g(\lambda(t)) \| \leq b.
$$

The next theorem states that $\rho(f, g)$ defines a metric on $\mathcal{D}([0, 1] : \mathbb{R}^{d})$ and gives some of its other properties. We call $\rho(f, g)$ the Skorohod metric and the topology defined by this metric the Skorohod topology.

Theorem A.6.5.
(a) $\rho(f, g)$ defines a metric on $\mathcal{D}([0, 1] : \mathbb{R}^{d})$ with respect to which $\mathcal{D}([0, 1] : \mathbb{R}^{d})$ is complete and separable.
(b) For any $f$ and $g$ in $\mathcal{D}([0, 1] : \mathbb{R}^{d})$, $\rho(f, g) \leq \| f - g \|_{\infty}$.
(c) Let $\{ \varphi_{n}, n \in \mathbb{N} \}$ be a sequence in $\mathcal{D}([0, 1] : \mathbb{R}^{d})$ and $\varphi \in C([0, 1] : \mathbb{R}^{d})$. If $\rho(\varphi_{n}, \varphi) \to 0$, then $\| \varphi_{n} - \varphi \|_{\infty} \to 0$.
(d) $\mathcal{C}([0, 1] : \mathbb{R}^{d})$ is a closed subset of $\mathcal{D}([0, 1] : \mathbb{R}^{d})$, and when relativized to $\mathcal{C}([0, 1] : \mathbb{R}^{d})$ the Skorohod topology coincides with the uniform topology.

Proof. Parts (a) and (c) and the second assertion in part (d) are proved in Section 14 of [9]. Part (b) is an immediate consequence of the definition of $\rho(f, g)$. Let $\{ \varphi_{n}, n \in \mathbb{N} \}$ be a sequence in $\mathcal{C}([0, 1] : \mathbb{R}^{d})$ such that $\rho(\varphi_{n}, \varphi) \to 0$ for some $\varphi \in \mathcal{D}([0, 1] : \mathbb{R}^{d})$. Then for each $n \in \mathbb{N}$ there exists a strictly increasing, continuous mapping $\lambda_{n}$ of $[0, 1]$ into itself such that

$$
\sup_{t \in [0, 1]} \| \varphi(t) - \varphi_{n}(\lambda_{n}(t)) \| = 0.
$$

Since this exhibits $\varphi$ as the uniform limit of the continuous functions $\varphi_{n} \circ \lambda_{n}$, it follows that $\varphi$ is continuous. This proves the first assertion in part (d). The proof of the theorem is complete. □
Appendix B

Deriving the Representation Formulas via Measure Theory

B.1 Introduction

In Section B.2 we give a measure-theoretic proof of a representation formula for Sanov’s
Theorem, following an approach suggested by Zeitouni [95]. Although the representation
formula is similar to the one used in Chapter 2, its proof is completely different from
the proof based on dynamic programming given in Section 2.3. In the special case of
Sanov’s Theorem the approach of Section B.2 gives an elegant derivation. However, as
we discuss in Section B.3, in more complicated examples we prefer the derivation of the
representation formula which is based on dynamic programming.

B.2 A Measure Theoretic Proof of the Representation Formula for Sanov’s Theorem

We use the notation of Chapter 2. Let $\mathcal{S}$ be a Polish space, $\rho$ a probability measure on
$\mathcal{X}$, and $\{X_j, j \in \mathbb{N}_0\}$ a sequence of i.i.d. random variables taking values in $\mathcal{S}$ and having
the common distribution $\rho$. For $n \in \mathbb{N}$ we consider the empirical measure

$$L^n = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{X_j}.$$ 

Given $h$ a bounded measurable function mapping $\mathcal{P}(\mathcal{S})$ into $\mathbb{R}$, a dynamic programming
argument yielded a representation formula for

$$W^n = \frac{1}{n} \log E\{\exp[-n h(L^n)]\}.$$ 

The representation, given in Theorem 2.3.2, states that

$$W^n = \inf_{\nu_j} \bar{E} \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu_j^n(\cdot||\bar{L}_j^n)) + h(\bar{L}^n) \right\}.$$  

(B.1)
In this formula the infimum is taken over all admissible control sequences \( \{\nu^n_j\} \), each \( \nu^n_j \) being a stochastic kernel on \( \mathcal{S} \) given \( \mathcal{M}_{j/n}(\mathcal{S}) \). The quantities \( \{\bar{L}^n_j, j = 0, 1, \ldots, n\} \) are the random measures defined recursively by \( \bar{L}^n_0 = 0 \) and

\[
\bar{L}^n_{j+1} = \bar{L}^n_j + \frac{1}{n} \delta_{\bar{X}^n_j},
\]

where the conditional distribution of \( \bar{X}^n_j \) is given by

\[
\tilde{P}\{\bar{X}^n_j \in dy|\bar{L}^n_1, \bar{L}^n_2, \ldots, \bar{L}^n_j\} \doteq \nu^n_j(dy|\bar{L}^n_j).
\]

In formula (B.1) \( \bar{L}^n \) denotes \( \bar{L}^n_n \).

We now give a measure-theoretic proof of an analogous representation. The only difference is that in the new formula the infimum is taken over a larger class of sequences of stochastic kernels than the class of admissible control sequences appearing in formula (B.1). The measure-theoretic proof is based on the following theorem, which is known as the **chain rule**. It is proved in Section C.3.

**Theorem B.2.1.** Let \( \mathcal{X} \) and \( \mathcal{Y} \) be Polish spaces and \( \alpha \) and \( \beta \) probability measures on \( \mathcal{X} \times \mathcal{Y} \). We denote by \( \alpha_1 \) and \( \beta_1 \) the first marginals of \( \alpha \) and \( \beta \) and by \( \alpha(dy|x) \) and \( \beta(dy|x) \) the stochastic kernels on \( \mathcal{Y} \) given \( \mathcal{X} \) for which we have the decompositions [Theorem A.5.4]

\[
\alpha(dx \times dy) = \alpha_1(dx) \otimes \alpha(dy|x) \quad \text{and} \quad \beta(dx \times dy) = \beta_1(dx) \otimes \beta(dy|x).
\]

Then the function mapping \( x \in \mathcal{X} \mapsto R(\alpha(\cdot|x)||\beta(\cdot|x)) \) is measurable and

\[
R(\alpha||\beta) = R(\alpha_1||\beta_1) + \int_{\mathcal{X}} R(\alpha(\cdot|x)||\beta(\cdot|x)) \alpha_1(dx).
\]

We now turn to the proof of the representation formula for Sanov’s Theorem. Let \( \rho^n \) denote the product measure on the \( n \)-fold product space \( \mathcal{S}^n \) with identical one-dimensional marginals \( \rho \). For \( (x_0, x_1, \ldots, x_{n-1}) \in \mathcal{S}^n \) we define the probability measure

\[
\ell^n \doteq \frac{1}{n} \sum_{j=0}^{n-1} \delta_{x_j}.
\]

Part (a) of Proposition 1.4.2 gives the representation

\[
W^n \doteq -\frac{1}{n} \log E\left\{\exp[-nh(L^n)]\right\} \quad (B.2)
\]

\[
= -\frac{1}{n} \log \int_{\mathcal{S}^n} \exp[-nh(L^n)] \rho^n(dx)
\]

\[
= \inf_{\nu^n \in \mathcal{P}(\mathcal{S}^n)} \left\{ \frac{1}{n}R(\nu^n||\rho^n) + \int_{\mathcal{S}^n} h(L^n) \nu^n(dx) \right\}.
\]

We now repeatedly apply Theorem A.5.4, factoring \( \nu^n(dx_0 \times dx_1 \times \cdots \times dx_{n-1}) \) in the form

\[
\nu^n(dx_0 \times dx_1 \times \cdots \times dx_{n-1}) = \nu^n_0(dx_0) \otimes \nu^n_1(dx_1|x_0) \otimes \cdots \otimes \nu^n_{n-1}(dx_{n-1}|x_0, x_1, \ldots, x_{n-2}).
\]
Here \( \nu^n_0 \) is the probability measure on \( \mathcal{S} \) obtained by projecting onto the first coordinate \( x_0 \) of \( \mathcal{S} \), and for \( j \in \{1, 2, \ldots, n - 2\} \) \( \nu^n_j(dx_j|x_0, x_1, \ldots, x_{j-1}) \) is a stochastic kernel on \( \mathcal{S} \) given \( \mathcal{S}^j \). Repeated application of Theorem B.2.1 yields

\[
R(\nu^n || \rho^n) = \int_{\mathcal{S}^n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot|x_0, x_1, \ldots, x_{j-1}) || \rho(\cdot)) \nu^n(dx_0 \times dx_1 \times \ldots \times dx_{n-1}).
\]

Substituting this into the last line of (B.2) gives the representation formula

\[
W^n = \inf_{\nu^n \in \mathcal{P}(\mathcal{S}^n)} \left\{ \int_{\mathcal{S}^n} \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot) || \rho(\cdot)) \nu^n(dx) + \int_{\mathcal{S}^n} h(\ell^n) \nu^n(dx) \right\}
\]

where \( \nu^n_j = \nu^n_j(\cdot|x_0, x_1, \ldots, x_{j-1}) \). The last display can also be written as

\[
W^n = \inf_{\{\nu^n_j\}} \left\{ \int_{\mathcal{S}} \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot) || \rho(\cdot)) \nu^n(dx) + \int_{\mathcal{S}^n} h(\ell^n) \nu^n(dx) \right\},
\]

where now the infimum is taken over all sequences of stochastic kernels \( \nu^n_j \) on \( \mathcal{S} \) given \( \mathcal{S}^j \).

Let us rewrite this formula so that it bears a stronger resemblance to equation (B.1). Given a sequence \( \{\nu^n_j, j = 0, 1, \ldots, n - 1\} \) of stochastic kernels \( \nu^n_j \) on \( \mathcal{S} \) given \( \mathcal{S}^j \), we recursively define a sequence of random variables \( \{\bar{X}^n_j, j = 0, 1, \ldots, n - 1\} \) taking values in \( \mathcal{S} \) by specifying the conditional distributions

\[
P\{\bar{X}^n_j \in dy | \bar{X}_0^n, \bar{X}_1^n, \ldots, \bar{X}_{j-1}^n\} = \nu^n_j(dy | \bar{X}_0^n, \bar{X}_1^n, \ldots, \bar{X}_{j-1}^n).
\]

Setting

\[
\bar{L}^n = \frac{1}{n} \sum_{j=1}^{n-1} \delta_{\bar{X}^n_j},
\]

we now have the following result.

**Theorem B.2.2** Let \( h \) be a bounded measurable function mapping \( \mathcal{P}(\mathcal{S}) \) into \( \mathbb{R} \). Then for all \( n \in \mathbb{N} \)

\[
W^n \doteq \frac{1}{n} \log E\{\exp[-n h(\bar{L}^n)]\}
\]

equals

\[
\inf_{\{\nu^n_j\}} \bar{E} \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu^n_j(\cdot|\bar{X}_0^n, \bar{X}_1^n, \ldots, \bar{X}_{j-1}^n) || \rho(\cdot)) + h(\bar{L}^n) \right\}.
\]

The infimum is taken over all sequences of stochastic kernels \( \nu^n_j \) on \( \mathcal{S} \) given \( \mathcal{S}^j \).

The measure–theoretic derivation of a representation formula for Sanov’s Theorem is complete. In the next section we remark on this approach.
B.3 Discussion

As we have just seen, measure theory provides a simple derivation of a representation formula for Sanov’s Theorem. However, when one considers problems with a more complicated structure than Sanov’s Theorem, this approach is less than ideal.

In the representation (B.1) derived via dynamic programming, the infimum is taken over all sequences \( \{ \nu^n_j, j = 0, 1, \ldots, n - 1 \} \), each \( \nu^n_j \) being a stochastic kernel on \( \mathcal{S} \) given \( \mathcal{M}_{j/n}(\mathcal{S}) \). In contrast, the infimum in the measure-theoretically derived representation formula stated in Theorem B.2.2 is taken over a much larger class of sequences; namely, all sequences of stochastic kernels \( \nu^n_j \) on \( \mathcal{S} \) given \( \mathcal{S}^j \). In other words, the control applied at time \( j \) in the first representation formula depends on the past \( (\hat{X}_0^n, \hat{X}_1^n, \ldots, \hat{X}_{j-1}^n) \) of the random variable \( \hat{X}_j^n \) only through the controlled empirical subprobability measure \( \hat{L}_j^n \equiv \frac{1}{n} \sum_{i=0}^{j-1} \delta_{\hat{X}_i^n} \). On the other hand, the control applied at time \( j \) in the second representation formula has no simple dependence on the past of \( \hat{X}_j^n \), and in general it is a function of all the random variables \( \hat{X}_0^n, \hat{X}_1^n, \ldots, \hat{X}_{j-1}^n \). While this more complicated dependence does not affect the proof of the Laplace principle for the empirical measures, it has a direct bearing on the numerical calculation of \( W^n \) through the representation, making it far more tedious. This aspect of the theory is not treated in the book, but is currently under investigation in the context of models arising in queueing theory [42].

A second and much more serious disadvantage is that the measure-theoretic derivation can easily give an awkward form of the representation if one happens to pick an inappropriate description of the underlying process. This can be seen if one tries to apply it even in the relatively straightforward context of Mogulskii’s Theorem, for which one has the choice of describing the discrete-time process \( \{ X^n_i, i = 0, 1, \ldots, n \} \) either in terms of the i.i.d. increments \( v_i \) or as a Markov chain with transition probability function given in equation (3.4). If one uses the first description of the process, then the measure-theoretic derivation yields a representation formula that is analogous to that derived via dynamic programming. However, if one happens to choose the second description of the process, then the resulting representation does not have a form that is useful for asymptotic analysis.

Mogulskii’s Theorem treats the simplest case of a class of random walk models. In any more general model, it is awkward even to write down the replacement for \( \rho^n \) which gives a useful representation formula. In contrast, when one uses the standard description of the process—i.e., that used for a law of large numbers analysis—dynamic programming hands one the most convenient representation. For these reasons we have focused on the dynamic programming derivation in the main part of the book.
Appendix C

Proofs of a Number of Results

C.1 Introduction

In this appendix we start by proving three parts of Lemma 1.4.3. They are the Donsker–Varadhan variational formula for the relative entropy, stated in part (a); a factorization property of the relative entropy, stated in part (f) and derived below from the chain rule; and an approximation property of the relative entropy, stated in part (g). We then prove parts (f) and (g) of Lemma 6.2.3, which concern the relationship between the relative entropy and Cramér functions. We end by proving a result on stochastic kernels and a continuity property of Cramér functions.

C.2 Proof of the Donsker–Varadhan Variational Formula for the Relative Entropy

The Donsker–Varadhan variational formula for the relative entropy was stated in part (a) of Lemma 1.4.3. Let $\mathcal{X}$ be a Polish space and denote by $\mathcal{C}_b(\mathcal{X})$ the space of bounded continuous functions mapping $\mathcal{X}$ into $\mathbb{R}$ and by $\Psi_b(\mathcal{X})$ the space of bounded measurable functions mapping $\mathcal{X}$ into $\mathbb{R}$. The variational formula states that for each $\gamma$ and $\theta$ in $\mathcal{P}(\mathcal{X})$

$$R(\gamma \| \theta) = \sup_{g \in \mathcal{C}_b(\mathcal{X})} \left\{ \int_{\mathcal{X}} g \, d\gamma - \log \int_{\mathcal{X}} e^g \, d\theta \right\} = \sup_{\psi \in \Psi_b(\mathcal{X})} \left\{ \int_{\mathcal{X}} \psi \, d\gamma - \log \int_{\mathcal{X}} e^\psi \, d\theta \right\}.$$ 

We first prove that

$$\sup_{g \in \mathcal{C}_b(\mathcal{X})} \left\{ \int_{\mathcal{X}} g \, d\gamma - \log \int_{\mathcal{X}} e^g \, d\theta \right\} = \sup_{\psi \in \Psi_b(\mathcal{X})} \left\{ \int_{\mathcal{X}} \psi \, d\gamma - \log \int_{\mathcal{X}} e^\psi \, d\theta \right\}. \quad (C.1)$$

Given $\varepsilon > 0$, the tightness of $\gamma$ and $\theta$ assures the existence of a compact subset $K$ of $\mathcal{X}$ such that $\gamma(K^c) \leq \varepsilon$ and $\theta(K^c) \leq \varepsilon$ [Theorem A.2.3]. By Lusin’s Theorem and the Tietze–Urysohn Extension Theorem, for any $\psi \in \Psi_b(\mathcal{X})$ there exists a closed subset $F$ of
$K$ and $g \in C_b(\mathcal{X})$ such that $(\gamma + \theta)(K \setminus F) \leq \varepsilon$, the restriction of $g$ to $F$ agrees with the restriction of $\psi$ to $F$, and $\|g\|_{\infty} \leq \|\psi\|_{\infty}$. It follows that

$$\gamma(F^c) \leq \gamma(K^c) + \gamma(K \setminus F) \leq 2\varepsilon \quad \text{and} \quad \theta(F^c) \leq \theta(K^c) + \theta(K \setminus F) \leq 2\varepsilon$$

and that there exists a constant $C < \infty$ independent of $\varepsilon > 0$ such that

$$\int_X \psi d\gamma - \log \int_X e^{\psi} d\theta \leq \int_X g d\gamma - \log \int_X e^{g} d\theta + C\varepsilon.$$ 

Taking the supremum over $g \in C_b(\mathcal{X})$ and sending $\varepsilon \to 0$ yields

$$\int_X \psi d\gamma - \log \int_X e^{\psi} d\theta \leq \sup_{g \in C_b(\mathcal{X})} \left\{ \int_X g d\gamma - \log \int_X e^{g} d\theta \right\},$$

and since $\psi \in \Psi_b(\mathcal{X})$ is arbitrary, we conclude that

$$\sup_{\psi \in \Psi_b(\mathcal{X})} \left\{ \int_X \psi d\gamma - \log \int_X e^{\psi} d\theta \right\} \leq \sup_{g \in C_b(\mathcal{X})} \left\{ \int_X g d\gamma - \log \int_X e^{g} d\theta \right\}.$$ 

Because $C_b(\mathcal{X}) \subset \Psi_b(\mathcal{X})$, the opposite inequality is obvious. This completes the proof of formula (C.1).

We define $H(\gamma, \theta)$ to be the common value of the suprema in formula (C.1). $H(\gamma, \theta)$ is nonnegative since if $g_0$ denotes the zero function on $\mathcal{X}$, then

$$H(\gamma, \theta) \geq \int_X g_0 d\gamma - \log \int_X e^{g_0} d\theta = 0.$$ 

By part (a) of Proposition 1.4.2, for any $k \in \Psi_b(\mathcal{X})$

$$R(\gamma\|\theta) \geq - \int_X k d\gamma - \log \int_X e^{-k} d\theta.$$ 

Replacing $k$ by $\psi \equiv -k$ and taking the supremum over $\psi \in \Psi_b(\mathcal{X})$, we obtain

$$R(\gamma\|\theta) \geq \sup_{\psi \in \Psi_b(\mathcal{X})} \left\{ \int_X \psi d\gamma - \log \int_X e^{\psi} d\theta \right\} = H(\gamma, \theta).$$

In order to show that $H(\gamma, \theta) \geq R(\gamma\|\theta)$, we use the proof given on pages 69–70 of [29]. We may assume that $H(\gamma, \theta) < \infty$ for otherwise there is nothing to prove. The first task is to show that under this condition $\gamma$ is absolutely continuous with respect to $\theta$. Let $A$ be a Borel set for which $\theta(A) = 0$ and take $r > 0$. Since for any $\psi \in \Psi_b(\mathcal{X})$

$$\int_X \psi d\gamma - \log \int_X e^{\psi} d\theta \leq H(\gamma, \theta) < \infty,$$ 

we obtain, upon substituting $\psi = r 1_A$, $r \gamma(A) \leq H(\gamma, \theta)$. Taking $r \to \infty$ gives $\gamma(A) = 0$, as claimed.
Since $\gamma \ll \theta$, we can define the Radon–Nikodym derivative $f = \frac{d\gamma}{d\theta}$. If $f$ is uniformly positive and bounded, then $\psi \doteq \log f$ is bounded and measurable, and substituting this function into (C.2) yields the desired inequality

$$R(\gamma\|\theta) = \int_X (\log f) \, d\gamma \leq H(\gamma, \theta).$$

If $f$ is uniformly positive but not bounded, then for $n \in \mathbb{N}$ we set $f_n = f \wedge n$ and substitute $\psi \doteq \log f_n$ into (C.2). Using the Monotone Convergence Theorem, we again obtain

$$R(\gamma\|\theta) = \int_X (\log f) \, d\gamma = \lim_{n \to \infty} \int_X (\log f_n) \, d\gamma \leq H(\gamma, \theta) + \lim_{n \to \infty} \log \int_X f_n \, d\theta = H(\gamma, \theta).$$

We now treat the general case where $f$ is neither uniformly positive nor bounded. For $t \in [0, 1]$ define

$$\gamma_t \doteq t \theta + (1 - t)\gamma \quad \text{and} \quad f_t = \frac{d\gamma_t}{d\theta} = t + (1 - t)f.$$

For each $t \in (0, 1]$ $f_t$ is uniformly positive, and so by the preceding calculation $R(\gamma_t\|\theta) \leq H(\gamma_t, \theta)$. We now prove that

$$\lim_{t \to 0} R(\gamma_t\|\theta) = R(\gamma\|\theta) \quad \text{and} \quad \lim_{t \to 0} H(\gamma_t, \theta) = H(\gamma, \theta). \quad (C.3)$$

This will complete the proof that $R(\gamma\|\theta) \leq H(\gamma, \theta)$.

Since $s \log s$ is convex on $[0, \infty)$,

$$R(\gamma_t\|\theta) = \int_X f_t (\log f_t) \, d\theta \leq (1 - t) \int_X f (\log f) \, d\theta = (1 - t)R(\gamma\|\theta).$$

On the other hand, since $\log s$ is nondecreasing and concave on $[0, \infty)$, $\log f_t \geq (\log t) \vee [(1 - t) \log f]$, and therefore

$$R(\gamma_t\|\theta) = \int_X f_t (\log f_t) \, d\theta \leq (1 - t) \int_X f (\log f_t) \, d\theta \geq t \log t + (1 - t)^2 R(\gamma\|\theta).$$

Combining the last two displays gives the first limit in equation (C.3).

It is easily checked that $H(\theta, \theta) = 0$ and that the function mapping $t \in [0, 1] \mapsto H(\gamma_t, \theta)$ is convex and lower semicontinuous. Thus for $t \in [0, 1]$

$$0 \leq H(\gamma_t, \theta) \leq tH(\theta, \theta) + (1 - t)H(\gamma_t, \theta) = (1 - t)H(\gamma, \theta) \leq H(\gamma, \theta) < \infty,$$

and so this function is also bounded. It follows that the function mapping $t \in [0, 1] \mapsto H(\gamma_t, \theta)$ is continuous [Theorem D.2.2 (d)] and therefore that

$$\lim_{t \to 0} H(\gamma_t, \theta) = H(\gamma_0, \theta) = H(\gamma, \theta).$$

This gives the second limit in equation (C.3). The proof of the Donsker–Varadhan variational formula is complete. ■
C.3 Proof of the Chain Rule and Part (f) of Lemma 1.4.3

The chain rule was stated in Theorem B.2.1 and applied in Section B.2. After restating it, we will derive from it a factorization property of the relative entropy given in part (f) of Lemma 1.4.3 as well as a related result. The chain rule will then be proved. It is a well known result in information theory [19].

**Theorem C.3.1 (chain rule).** Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and $\alpha$ and $\beta$ probability measures on $\mathcal{X} \times \mathcal{Y}$. We denote by $\alpha_1$ and $\beta_1$ the first marginals of $\alpha$ and $\beta$ and by $\alpha(dy|x)$ and $\beta(dy|x)$ the stochastic kernels on $\mathcal{Y}$ given $\mathcal{X}$ for which we have the decompositions [Theorem A.5.4]

$$\alpha(dx \times dy) = \alpha_1(dx) \otimes \alpha(dy|x) \quad \text{and} \quad \beta(dx \times dy) = \beta_1(dx) \otimes \beta(dy|x).$$

Then the function mapping $x \in \mathcal{X} \mapsto R(\alpha(\cdot|x)||\beta(\cdot|x))$ is measurable and

$$R(\alpha||\beta) = R(\alpha_1||\beta_1) + \int_{\mathcal{X}} R(\alpha(\cdot|x)||\beta(\cdot|x)) \alpha_1(dx).$$

We next derive from this the following fact, which was stated in part (f) of Lemma 1.4.3.

**Corollary C.3.2.** Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, $\sigma(dy|x)$ and $\tau(dy|x)$ stochastic kernels on $\mathcal{Y}$ given $\mathcal{X}$, and $\theta$ a probability measure on $\mathcal{X}$. Then the function mapping $x \in \mathcal{X} \mapsto R(\sigma(\cdot|x)||\tau(\cdot|x))$ is measurable and

$$\int_{\mathcal{X}} R(\sigma(\cdot|x)||\tau(\cdot|x)) \theta(dx) = R(\theta \otimes \sigma || \theta \otimes \tau). \quad (C.4)$$

**Proof.** The measurability of the function mapping $x \in \mathcal{X} \mapsto R(\sigma(\cdot|x)||\tau(\cdot|x))$ is shown in the proof of Theorem C.3.1. In order to prove formula (C.4), we make the following identifications in Theorem C.3.1: $\alpha(dx \times dy) \asymp \theta(dx) \otimes \sigma(dy|x)$ and $\beta(dx \times dy) \asymp \theta(dx) \otimes \tau(dy|x)$. Then the first marginals $\alpha_1(dx)$ and $\beta_1(dx)$ both equal $\theta(dx)$, and Theorem C.3.1 implies that

$$R(\theta \otimes \sigma || \theta \otimes \tau) = R(\theta||\theta) + \int_{\mathcal{X}} R(\sigma(\cdot|x)||\tau(\cdot|x)) \theta(dx) = \int_{\mathcal{X}} R(\sigma(\cdot|x)||\tau(\cdot|x)) \theta(dx).$$

This is formula (C.4). $\blacksquare$

For use elsewhere in the text, we need another corollary of Theorem C.3.1 that applies to product measures.

**Corollary C.3.3.** Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, $\gamma$ and $\theta$ probability measures on $\mathcal{X}$, and $\lambda$ and $\mu$ probability measures on $\mathcal{Y}$. Then

$$R(\gamma \times \lambda||\theta \times \mu) = R(\gamma||\theta) + R(\lambda||\mu).$$
C.3. PROOF OF THE CHAIN RULE AND PART (F) OF LEMMA 1.4.3

Proof of Theorem C.3.1. The stochastic kernels $\alpha(dy|x)$ and $\beta(dy|x)$ are measurable functions mapping $\mathcal{X}$ into $\mathcal{P}(\mathcal{Y})$ [Theorem A.5.2]. Since $R(\cdot,\cdot)$ is lower semicontinuous on $\mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{Y})$ [Lemma 1.4.3 (b)], the measurability of the function mapping $x \in \mathcal{X} \mapsto R(\alpha(\cdot|x),\beta(\cdot|x))$ follows. We now prove that

$$R(\alpha\|\beta) = R(\alpha_1\|\beta_1) + \int_\mathcal{X} R(\alpha(\cdot|x),\beta(\cdot|x)) \alpha_1(dx), \tag{C.5}$$

assuming first that the right hand side is finite. Under this assumption $\alpha_1 \ll \beta_1$ and there exists an $\alpha_1$-null set $\Gamma$ in $\mathcal{X}$ such that $R(\alpha(\cdot|x),\beta(\cdot|x))$ is finite for $x \in \Gamma$. Hence for $x \in \Gamma \alpha(\cdot|x) \ll \beta(\cdot|x)$ as measures on $\mathcal{Y}$. By redefining $\alpha(\cdot|x)$ for $x$ in the null set $\Gamma$, we can assure that $\alpha(\cdot|x) \ll \beta(\cdot|x)$ for all $x \in \mathcal{X}$. Let

$$\psi(x) = \frac{d\alpha_1}{d\beta_1}(x).$$

Theorem A.5.7 guarantees that there exists a version of the Radon–Nikodym derivative

$$\zeta(x,y) = \frac{d\alpha(\cdot|x)}{d\beta(\cdot|x)}(y)$$

which is a nonnegative measurable function on $\mathcal{X} \times \mathcal{Y}$. For any Borel subsets $A$ of $\mathcal{X}$ and $B$ of $\mathcal{Y}$

$$\alpha(A \times B) = \int_A \alpha(B|x) \alpha_1(dx)$$

$$= \int_A \left( \int_B \zeta(x,y) \beta(dy|x) \right) \psi(x) \beta_1(dx) = \int_{A \times B} \psi(x) \zeta(x,y) \beta(dx \times dy).$$

This implies that $\alpha \ll \beta$ and that for $(x,y) \in \mathcal{X} \times \mathcal{Y}$

$$\frac{d\alpha}{d\beta}(x,y) = \psi(x) \zeta(x,y).$$

As a consequence

$$R(\alpha_1\|\beta_1) + \int_\mathcal{X} R(\alpha(\cdot|x),\beta(\cdot|x)) \alpha_1(dx)$$

$$= \int_\mathcal{X} \log \psi(x) \alpha_1(dx) + \int_\mathcal{X} \left( \int_\mathcal{Y} \log \zeta(x,y) \alpha(dy|x) \right) \alpha_1(dx)$$

$$= \int_{\mathcal{X} \times \mathcal{Y}} \log \psi(x) \alpha(dx \times dy) + \int_{\mathcal{X} \times \mathcal{Y}} \log \zeta(x,y) \alpha_1(dx) \otimes \alpha(dy|x)$$

$$= \int_{\mathcal{X} \times \mathcal{Y}} \log(\psi(x) \zeta(x,y)) \alpha(dx \times dy)$$

$$= R(\alpha\|\beta).$$

This is formula (C.5).

We now prove (C.5) under the assumption that the left hand side is finite. Under this assumption $\alpha \ll \beta$, and so for $(x,y) \in \mathcal{X} \times \mathcal{Y}$ we can define

$$\varphi(x,y) = \frac{d\alpha}{d\beta}(x,y).$$
Since $\alpha \ll \beta$, $\alpha_1 \ll \beta_1$ and

$$\psi(x) \overset{d}{=} \frac{d\alpha_1}{d\beta_1}(x) \text{ equals } \int_Y \varphi(x, y) \beta(dy|x).$$

For any Borel subsets $A$ of $X$ and $B$ of $Y$

$$\int_A \alpha(B|x) \psi(x) \beta_1(dx) = \int_A \alpha(B|x) \alpha_1(dx)$$

$$= \alpha(A \times B)$$

$$= \int_{A \times B} \varphi(x, y) \beta(dx \times dy)$$

$$= \int_A \left( \int_B \varphi(x, y) \beta(dy|x) \right) \beta_1(dx).$$

This implies that there exists a $\beta_1$-null set $\Gamma$ such that for all $x \in \Gamma^c$

$$\psi(x) \alpha(B|x) = \int_B \varphi(x, y) \beta(dy|x).$$

Thus for all $x \in \Gamma^c \cap \{ \psi > 0 \}$ $\alpha(\cdot|x) \ll \beta(\cdot|x)$, and for such $x$ and all $y \in Y$

$$\zeta(x, y) \overset{d}{=} \frac{d\alpha(\cdot|x)}{d\beta(\cdot|x)}(y) \text{ equals } \frac{\varphi(x, y)}{\psi(x)}.$$

In other words, for all $x \in \Gamma^c \cap \{ \psi > 0 \}$ the various Radon–Nikodym derivatives are related by

$$\varphi(x, y) = \psi(x) \zeta(x, y).$$

We have $\alpha_1(\{ \psi > 0 \}) = 1$, and since $\alpha_1 \ll \beta_1$ and $\beta_1(\Gamma^c) = 1$, we also have $\alpha_1(\Gamma^c) = 1$. It now follows that

$$R(\alpha\|\beta) = \int_{X \times Y} \log \varphi(x, y) \alpha(dx \times dy)$$

$$= \int_{X \times Y} \log \varphi(x, y) \alpha_1(dx) \otimes \alpha(dy|x)$$

$$= \int_{(\Gamma^c \cap \{ \psi > 0 \}) \times Y} \log[\psi(x) \zeta(x, y)] \alpha_1(dx) \otimes \alpha(dy|x)$$

$$= \int_{\Gamma^c \cap \{ \psi > 0 \}} \log \varphi(x) \alpha_1(dx) + \int_{\Gamma^c \cap \{ \psi > 0 \}} \left( \int_Y \log \zeta(x, y) \alpha(dy|x) \right) \alpha_1(dx)$$

$$= R(\alpha_1\|\beta_1) + \int_X R(\alpha(\cdot|x)\|\beta(\cdot|x)) \alpha_1(dx),$$

which again is equation (C.5). This finishes the proof of the theorem. ■

### C.4 Proof of Part (g) of Lemma 1.4.3

We denote by $\Pi$ the class of all finite measurable partitions of the Polish space $X$. Part (g) of Lemma 1.4.3 states that for each $\gamma$ and $\theta$ in $P(X)$

$$R(\gamma\|\theta) = \sup_{\pi \in \Pi} \sum_{A \in \pi} \gamma(A) \log \frac{\gamma(A)}{\theta(A)}. \quad (C.6)$$
where the summand equals 0 if \( \gamma(A) = 0 \) and equals \( \infty \) if \( \gamma(A) > 0 \) and \( \theta(A) = 0 \). In addition, if \( A \) is any Borel subset of \( \mathcal{X} \), then

\[
R(\gamma \| \theta) \geq \gamma(A) \log \frac{\gamma(A)}{\theta(A)} - 1.
\]

(C.7)

We first prove that for any finite measurable partition \( \pi \) of \( \mathcal{X} \)

\[
R(\gamma \| \theta) \geq \sum_{A \in \pi} \gamma(A) \log \frac{\gamma(A)}{\theta(A)}.
\]

If \( R(\gamma \| \theta) = \infty \), there is nothing to prove, so we assume that \( R(\gamma \| \theta) < \infty \). In this case \( \gamma \ll \theta \), and setting \( B = \cup \{ A \in \pi : \gamma(A) > 0 \} \), we define for \( m \in \mathbb{N} \) the bounded measurable function

\[
\psi_m(x) \doteq \sum_{\{ A \in \pi : \gamma(A) > 0 \}} \left( \log \frac{\gamma(A)}{\theta(A)} \right) 1_A(x) - m 1_B(x).
\]

The Donsker–Varadhan variational formula stated in part (a) of Lemma 1.4.3 implies that

\[
R(\gamma \| \theta) \geq \int_{\mathcal{X}} \psi_m d\gamma - \log \int_{\mathcal{X}} e^{\psi_m} d\theta = \sum_{\{ A \in \pi : \gamma(A) > 0 \}} \gamma(A) \log \frac{\gamma(A)}{\theta(A)} - \log(1 + e^{-m\theta(B)}).
\]

This yields the desired formula since \( \lim_{m \to \infty} \log(1 + e^{-m\theta(B)}) = 0 \).

In order to complete the proof of equation (C.6), we determine a sequence \( \{ \pi_n, n \in \mathbb{N} \} \) of finite measurable partitions of \( \mathcal{X} \) having the property that

\[
R(\gamma \| \theta) = \lim_{n \to \infty} \sum_{A \in \pi_n} \gamma(A) \log \frac{\gamma(A)}{\theta(A)}.
\]

(C.8)

We carry this out via a standard technique, using unpublished notes of Barron [6]. If \( \gamma \) is not absolutely continuous with respect to \( \theta \), then the proof is straightforward. Indeed, in this case there exists a Borel subset \( A \) of \( \mathcal{X} \) having the property that \( \theta(A) = 0 \) and \( \gamma(A) > 0 \). We obtain formula (C.8) by setting \( \pi_n \doteq \{ A, A^c \} \) for each \( n \in \mathbb{N} \).

We now suppose that \( \gamma \) is absolutely continuous with respect to \( \theta \) and let \( f \doteq \frac{\gamma}{\theta} \). For each \( n \in \mathbb{N} \) we then define \( \pi_n \) to be the finite measurable partition of \( \mathcal{X} \) consisting of the disjoint Borel sets

\[
A_{n,k} \doteq \begin{cases} 
\{ x \in \mathcal{X} : \log f(x) \leq -\sqrt{n} \} & \text{if } k = -n \\
\{ x \in \mathcal{X} : \frac{k+1}{\sqrt{n}} < \log f(x) \leq \frac{k}{\sqrt{n}} \} & \text{if } k \in \{-n + 1, -n + 2, \ldots, n - 1, n \} \\
\{ x \in \mathcal{X} : \log f(x) > \sqrt{n} \} & \text{if } k = n + 1.
\end{cases}
\]

For \(-n + 1 \leq k \leq n + 1\)

\[
\gamma(A_{n,k}) = \int_{A_{n,k}} \exp(\log f) \, d\theta \geq \exp \left[ \frac{k - 1}{\sqrt{n}} \right] \theta(A_{n,k}).
\]

(C.9)
The error in the approximation of the relative entropy by the sum over the partition \( \pi_n \) equals

\[
R(\gamma \| \theta) - \sum_{A \in \pi_n} \gamma(A) \log \frac{\gamma(A)}{\theta(A)} = \sum_{k=-n}^{n+1} 1_{\{j: \gamma(A_{n,j}) > 0\}}(k) \int_{A_{n,k}} \log \left( f \frac{\theta(A_{n,k})}{\gamma(A_{n,k})} \right) d\gamma.
\]

We now bound each term in this sum. For \(-n + 1 \leq k \leq n\) and \( x \in A_{n,k}\), if \( \gamma(A_{n,k}) > 0 \), then the integrand satisfies

\[
\log \left( f(x) \frac{\theta(A_{n,k})}{\gamma(A_{n,k})} \right) \leq \frac{k}{\sqrt{n}} + \frac{k - 1}{\sqrt{n}} = \frac{1}{\sqrt{n}}
\]

which implies that

\[
\sum_{k=-n+1}^{n} 1_{\{j: \gamma(A_{n,j}) > 0\}}(k) \int_{A_{n,k}} \log \left( f \frac{\theta(A_{n,k})}{\gamma(A_{n,k})} \right) d\gamma \leq \frac{1}{\sqrt{n}} \sum_{k=-n+1}^{n} \gamma(A_{n,k}) \leq \frac{1}{\sqrt{n}}.
\]

For \( k = -n \) and \( x \in A_{n,-n} \), if \( \gamma(A_{n,-n}) > 0 \), then the integrand satisfies

\[
\log \left( f(x) \frac{\theta(A_{n,-n})}{\gamma(A_{n,-n})} \right) \leq \log \left( e^{-\sqrt{n}} \frac{1}{\gamma(A_{n,-n})} \right),
\]

and so, since \( s \log s \geq -e^{-1} \) for \( s \in [0, \infty) \),

\[
\int_{A_{n,-n}} \log \left( f \frac{\theta(A_{n,-n})}{\gamma(A_{n,-n})} \right) d\gamma \leq -\gamma(A_{n,-n}) \log \left( e^{\sqrt{n}} \gamma(A_{n,-n}) \right) \leq e^{-\sqrt{n}-1}.
\]

Finally for \( k = n+1 \), if \( \gamma(A_{n,n+1}) > 0 \), then (C.9) implies that \( \theta(A_{n,n+1})/\gamma(A_{n,n+1}) \leq 1 \). Thus

\[
\int_{A_{n,n+1}} \log \left( f \frac{\theta(A_{n,n+1})}{\gamma(A_{n,n+1})} \right) d\gamma \leq \int_{A_{n,n+1}} (\log f) d\gamma = \int_{\{\log f > \sqrt{n}\}} (\log f) d\gamma.
\]

Combining these inequalities yields

\[
0 \leq R(\gamma \| \theta) - \sum_{A \in \pi_n} \gamma(A) \log \frac{\gamma(A)}{\theta(A)} \leq \frac{1}{\sqrt{n}} + e^{-\sqrt{n}-1} + \int_{\{\log f > \sqrt{n}\}} (\log f) d\gamma.
\]

If \( R(\gamma \| \theta) < \infty \), then the integral in this inequality converges to 0 as \( n \to \infty \), and thus

\[
\lim_{n \to \infty} \sum_{A \in \pi_n} \gamma(A) \log \frac{\gamma(A)}{\theta(A)} = R(\gamma \| \theta).
\]

Now assume that \( \gamma \) is absolutely continuous with respect to \( \theta \) but that \( R(\gamma \| \theta) = \infty \). For any Borel set \( B \), if \( \theta(B) = 0 \), then \( \gamma(B) = 0 \) and \( \gamma(B) \log [\gamma(B)/\theta(B)] = 0 \), while if \( \theta(B) > 0 \), then since \( s \log s \geq s - 1 \) for \( s \in [0, \infty) \),

\[
\gamma(B) \log \frac{\gamma(B)}{\theta(B)} = \theta(B) \left[ \frac{\gamma(B)}{\theta(B)} \log \frac{\gamma(B)}{\theta(B)} \right] \geq \theta(B) \left[ \frac{\gamma(B)}{\theta(B)} - 1 \right] \geq -1.
\]

(C.10)
Since \( \{A_{n,k}, -n \leq k \leq n\} \) is a finite measurable partition of \( \{\log f \leq \sqrt{n}\} \), similar estimates as in the case where \( R(\gamma\|\theta) < \infty \) yield

\[
\int_{\{\log f \leq \sqrt{n}\}} (\log f) \, d\gamma - \sum_{k=-n}^{n} \gamma(A_{n,k}) \log \frac{\gamma(A_{n,k})}{\theta(A_{n,k})} \leq \frac{1}{\sqrt{n}} + e^{-\sqrt{n}-1}.
\]

Thus

\[
\sum_{A \in \pi_n} \gamma(A) \log \frac{\gamma(A)}{\theta(A)} \geq \int_{\{\log f \leq \sqrt{n}\}} (\log f) \, d\gamma - 1 - \frac{1}{\sqrt{n}} - e^{-\sqrt{n}-1}.
\]

Since the right hand side converges to \( \infty = R(\gamma\|\theta) \) as \( n \to \infty \), we have completed the proof of (C.8) and thus the proof of (C.6).

We now prove formula (C.7). Given \( A \) a Borel subset of \( \mathcal{X} \), (C.6) yields for the finite measurable partition \( \pi \triangleq \{A, A^c\} \)

\[
R(\gamma\|\theta) \geq \gamma(A) \log \frac{\gamma(A)}{\theta(A)} + \gamma(A^c) \log \frac{\gamma(A^c)}{\theta(A^c)}.
\]

If \( \theta(A^c) = 0 \), then the last term in this display equals either 0 or \( \infty \) depending on whether \( \gamma(A^c) \) equals 0 or is positive. In either case formula (C.7) follows. On the other hand, if \( \theta(A^c) > 0 \), then by (C.10)

\[
R(\gamma\|\theta) \geq \gamma(A) \log \frac{\gamma(A)}{\theta(A)} - 1.
\]

This is what we want to prove. The proof of part (g) of Lemma 1.4.3 is complete. ■

### C.5 Proof of Part (f) of Lemma 6.2.3

In this section we prove under Condition 6.2.1 the relationship between the relative entropy and \( L(x, \beta) \) given in part (f) of Lemma 6.2.3. In order to streamline the notation, we will suppress \( x \), writing \( H(\alpha), L(\beta), \) and \( \mu \) instead of \( H(x, \alpha), L(x, \beta), \) and \( \mu(\cdot|x) \).

Our goal is to establish the variational formula

\[
L(\beta) = \inf \left\{ R(\gamma\|\mu) : \gamma \in \mathcal{P}(\mathbb{R}^d), \int_{\mathbb{R}^d} y \gamma(dy) = \beta \right\}
\]

for each \( \beta \in \mathbb{R}^d \) and to show that if \( L(\beta) \) is finite, then the infimum is uniquely attained at some measure \( \gamma \). Obviously, if the infimum equals \( \infty \), then the infimum is attained at any measure \( \gamma \) having mean \( \beta \).

We first carry this out under the assumption that \( \beta \) lies in \( \text{ri(dom } L) \). For \( \alpha \in \mathbb{R}^d \) we denote by \( \mu_\alpha \) the probability measure on \( \mathbb{R}^d \) which is absolutely continuous with respect to \( \mu \) and satisfies

\[
\frac{d\mu_\alpha}{d\mu}(y) = \exp(\langle \alpha, y \rangle) \cdot \frac{1}{\int_{\mathbb{R}^d} \exp(\langle \alpha, y \rangle) \, \mu(dy)}.
\]
We consider this measure $\mu_\alpha$ because a formal use of Lagrange multipliers suggests that if $\int_{\mathbb{R}^d} y \mu_\alpha(dy) = \beta$, then $\mu_\alpha$ is an extremal point of $R(\gamma||\mu)$ on the set of $\gamma \in \mathcal{P}(\mathbb{R}^d)$ satisfying $\int_{\mathbb{R}^d} y \gamma(dy) = \beta$. We now turn to the proof of (C.11). Since

$$\nabla H(\alpha) = \int_{\mathbb{R}^d} y \exp \langle \alpha, y \rangle \mu(dy) \cdot \frac{1}{\int_{\mathbb{R}^d} \exp \langle \alpha, y \rangle \mu(dy)} = \int_{\mathbb{R}^d} y \mu_\alpha(dy),$$

part (d) of Lemma 6.2.3 implies that for $\beta \in \text{ri}(\text{dom } L)$ there exists $\alpha = \alpha(\beta) \in \mathbb{R}^d$ such that $\int_{\mathbb{R}^d} y \mu_{\alpha(\beta)}(dy) = \beta$ and

$$\langle \alpha(\beta), \beta \rangle - H(\alpha(\beta)) = L(\beta).$$

Now let $\gamma$ be any probability measure on $\mathbb{R}^d$ which has mean $\beta$ and for which $R(\gamma||\mu)$ is finite. Then $\gamma$ is absolutely continuous both with respect to $\mu$ and with respect to $\mu_{\alpha(\beta)}$ and

$$R(\gamma||\mu) = \int_{\mathbb{R}^d} \left( \log \frac{d\gamma}{d\mu}(y) \right) \gamma(dy)$$

$$= \int_{\mathbb{R}^d} \left( \log \frac{d\gamma}{d\mu_{\alpha(\beta)}}(y) \right) \gamma(dy) + \int_{\mathbb{R}^d} \left( \log \frac{d\mu_{\alpha(\beta)}}{d\mu}(y) \right) \gamma(dy)$$

$$= R(\gamma||\mu_{\alpha(\beta)}) + \int_{\mathbb{R}^d} \left( \langle \alpha(\beta), y \rangle - \log \int_{\mathbb{R}^d} \exp \langle \alpha(\beta), y \rangle \mu(dy) \right) \gamma(dy)$$

$$= R(\gamma||\mu_{\alpha(\beta)}) + \langle \alpha(\beta), \beta \rangle - H(\alpha(\beta))$$

$$= R(\gamma||\mu_{\alpha(\beta)}) + L(\beta).$$

Since $R(\gamma||\mu_{\alpha(\beta)})$ is nonnegative and equality holds if and only if $\gamma = \mu_{\alpha(\beta)}$ [Lemma 1.4.1], it follows that for $\gamma \in \mathcal{P}(\mathbb{R}^d)$ satisfying $\int_{\mathbb{R}^d} y \gamma(dy) = \beta$

$$R(\gamma||\mu) \geq L(\beta) = R(\mu_{\alpha(\beta)}||\mu)$$

and equality holds if and only if $\gamma = \mu_{\alpha(\beta)}$. This completes the proof of part (f) of Lemma 6.2.3 for $\beta \in \text{ri}(\text{dom } L)$. The measure $\mu_\alpha$ will also play a key role in the proof of part (g) of this lemma, which is given in the next section of the appendix. We point out that if $\alpha_1 \neq \alpha_2$ are two distinct points in $\mathbb{R}^d$ satisfying $\nabla H(\alpha_1) = \nabla H(\alpha_2) = \beta$, then $\mu_{\alpha_1}$ and $\mu_{\alpha_2}$ are equal [45, Thm. VIII.4.4].

We now consider $\beta \not\in \text{ri}(\text{dom } L)$ for which $L(\beta) < \infty$. Such a point $\beta$ must lie in the relative boundary of $\text{dom } L$, and so by part (c) of Theorem D.2.2 there exists a sequence $\{\beta_n, n \in \mathbb{N}\} \subset \text{ri}(\text{dom } L)$ such that

$$\lim_{n \to \infty} \beta_n = \beta \quad \text{and} \quad \lim_{n \to \infty} L(\beta_n) = L(\beta).$$

By our work in the last paragraph, for each $n \in \mathbb{N}$ $\gamma_n \equiv \mu_{\alpha(\beta_n)}$ is the unique probability measure on $\mathbb{R}^d$ with the properties that

$$R(\gamma_n||\mu) = L(\beta_n) \quad \text{and} \quad \int_{\mathbb{R}^d} y \gamma_n(dy) = \beta_n.$$
It follows that $\sup_{n \in \mathbb{N}} R(\gamma_n \| \mu) < \infty$ and thus that the sequence $\{\gamma_n, n \in \mathbb{N}\}$ is tight and uniformly integrable [Lemma 1.4.3 (d)]. Given any subsequence of $\{\gamma_n\}$, there exists a subsequence and $\gamma^* \in \mathcal{P}(\mathbb{R}^d)$ such that [Theorem A.3.19 (a)]

$$\gamma_n \Longrightarrow \gamma^* \quad \text{and} \quad \lim_{n \to \infty} \int_{\mathbb{R}^d} y \gamma_n(dy) = \lim_{n \to \infty} \beta_n = \beta = \int_{\mathbb{R}^d} y \gamma^*(dy).$$

We claim that the infimum in equation (C.11) is uniquely attained at $\gamma^*$. Indeed, since $R(\cdot \| \mu)$ is lower semicontinuous,

$$L(\beta) = \lim_{n \to \infty} L(\beta_n) = \lim_{n \to \infty} R(\gamma_n \| \mu) \geq R(\gamma^* \| \mu) \geq L(\beta).$$

Thus $R(\gamma^* \| \mu) = L(\beta)$, as claimed. In addition, if $\gamma \neq \gamma^*$ is a distinct probability measure on $\mathbb{R}^d$ satisfying $R(\gamma \| \mu) = L(\beta)$ and $\int_{\mathbb{R}^d} y \gamma(dy) = \beta$, then by the strict convexity of $R(\cdot \| \mu)$ [Lemma 1.4.3 (b)] we have for any $s \in (0,1)$

$$L(\beta) \leq R(s \gamma^* + (1-s) \gamma \| \mu) < s R(\gamma^* \| \mu) + (1-s) R(\gamma \| \mu) = L(\beta).$$

This contradiction proves the uniqueness and shows that any subsequence of $\{\gamma_n, n \in \mathbb{N}\}$ has a subsequence converging weakly to $\gamma^*$. The standard argument by contradiction applied to an arbitrary subsequence implies that the entire sequence $\{\gamma_n, n \in \mathbb{N}\}$ converges weakly to $\gamma^*$.

We have proved the following lemma.

**Lemma C.5.1.** We assume Condition 6.2.1. Let $\beta$ be any point in $\mathbb{R}^d$ for which $L(\beta) < \infty$. Then

$$L(\beta) = \inf \left\{ R(\gamma \| \mu) : \gamma \in \mathcal{P}(\mathbb{R}^d), \int_{\mathbb{R}^d} y \gamma(dy) = \beta \right\}$$

and the infimum is uniquely attained at some measure $\gamma^*$. If $\beta \in \text{ri}(\text{dom } L)$, then $\gamma^* = \mu_{\alpha(\beta)}$ [see (C.12)], and if $\beta$ lies in the relative boundary of $\text{dom } L$, then $\mu_{\alpha(\beta_n)} \rightharpoonup \gamma^*$, where $\{\beta_n\}$ is some sequence in $\text{ri}(\text{dom } L)$ satisfying $\beta_n \to \beta$.

We now prove that if $\beta$ is any point in $\mathbb{R}^d$ for which

$$\inf \left\{ R(\gamma \| \mu) : \gamma \in \mathcal{P}(\mathbb{R}^d), \int_{\mathbb{R}^d} y \gamma(dy) = \beta \right\} < \infty,$$  \hspace{1cm} (C.13)

then

$$L(\beta) = \inf \left\{ R(\gamma \| \mu) : \gamma \in \mathcal{P}(\mathbb{R}^d), \int_{\mathbb{R}^d} y \gamma(dy) = \beta \right\}$$ \hspace{1cm} (C.14)

and the infimum is uniquely attained at some measure $\gamma^*$. Indeed, if (C.13) holds, then there exists $\tilde{\gamma} \in \mathcal{P}(\mathbb{R}^d)$ for which $R(\tilde{\gamma} \| \mu) < \infty$ and $\int_{\mathbb{R}^d} y \tilde{\gamma}(dy) = \beta$. We claim that

$$L(\beta) \leq R(\tilde{\gamma} \| \mu) < \infty.$$  \hspace{1cm} (C.15)

Once this is shown, we are done because we can then appeal to the earlier part of the proof and assert that since $L(\beta) < \infty$ (C.14) is valid and the infimum is uniquely attained at some measure $\gamma^*$. 

According to the Donsker–Varadhan variational formula for the relative entropy [Lemma 1.4.3 (a)], for any $g \in C_b(\mathbb{R}^d)$

$$\infty > R(\gamma \| \mu) \geq \int_{\mathbb{R}^d} g(y) \hat{\gamma}(dy) - \log \int_{\mathbb{R}^d} \exp[g(y)] \mu(dy).$$

Let $\alpha$ be any point in $\mathbb{R}^d$. For $k \in \mathbb{N}$ we evaluate this inequality for the bounded continuous function

$$g_k(y) = \begin{cases} \langle \alpha, y \rangle & \text{if } \|y\| \leq k \\ k \langle \alpha, y/\|y\| \rangle & \text{if } \|y\| \geq k. \end{cases}$$

This function satisfies $|g_k(y)| \leq \|\alpha\| \|y\|$ for all $y \in \mathbb{R}^d$, and since $\int_{\mathbb{R}^d} \|y\| \hat{\gamma}(dy) < \infty$, the Lebesgue Dominated Convergence Theorem implies that

$$\lim_{k \to \infty} \int_{\mathbb{R}^d} g_k(y) \hat{\gamma}(dy) = \int_{\mathbb{R}^d} \langle \alpha, y \rangle \hat{\gamma}(dy) = \langle \alpha, \int_{\mathbb{R}^d} y \hat{\gamma}(dy) \rangle = \langle \alpha, \beta \rangle.$$

By assumption on $\mu$, $\int_{\mathbb{R}^d} \exp(\langle \alpha, y \rangle) \mu(dy) < \infty$ for all $\alpha \in \mathbb{R}^d$, and this is equivalent to the bound $\int_{\mathbb{R}^d} \exp[r \|y\|] \mu(dy) < \infty$ for all $r > 0$. Hence again by the Lebesgue Dominated Convergence Theorem

$$\lim_{k \to \infty} \log \int_{\mathbb{R}^d} \exp[g_k(y)] \mu(dy) = \log \int_{\mathbb{R}^d} \exp(\langle \alpha, y \rangle) \mu(dy) = H(\alpha).$$

It follows that

$$R(\gamma \| \mu) \geq \lim_{k \to \infty} \left( \int_{\mathbb{R}^d} g_k(y) \hat{\gamma}(dy) - \log \int_{\mathbb{R}^d} \exp[g_k(y)] \mu(dy) \right) = \langle \alpha, \beta \rangle - H(\alpha),$$

and since $\alpha \in \mathbb{R}^d$ is arbitrary, that

$$R(\gamma \| \mu) \geq \sup_{\alpha \in \mathbb{R}^d} \{ \langle \alpha, \beta \rangle - H(\alpha) \} = L(\beta).$$

This proves (C.15) and completes the proof of part (f) of Lemma 1.4.3. ■

### C.6 Proof of Part (g) of Lemma 6.2.3

In this section we prove under Condition 6.2.1 that there exists a stochastic kernel $\gamma(dy|x)$ on $\mathbb{R}^d$ given $\mathbb{R}^d$ satisfying for $x$ and $\beta$ in $\mathbb{R}^d$

$$R(\gamma(\cdot|x) || \mu(\cdot|x)) = L(x, \beta) \quad \text{and} \quad \int_{\mathbb{R}^d} y \gamma(dy|x) = \beta. \quad (C.16)$$

We show in fact that a stochastic kernel $\gamma(dy|x)$ can be chosen to be a measurable function $\gamma(dy|x, \beta)$ of both $x$ and $\beta$ in $\mathbb{R}^d$. By part (b) of Lemma 6.2.3 $L(x, \beta)$ is a lower semicontinuous function of $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$ and thus is measurable on $\mathbb{R}^d \times \mathbb{R}^d$. 
We now turn to the construction of a stochastic kernel solving (C.16). Given \( x \in \mathbb{R}^d \) we first consider \( \beta \in \mathbb{R}^d \) satisfying \( L(x, \beta) < \infty \). By Lemma C.5.1 there exists a unique probability measure \( \gamma^*(dy) = \gamma^*(dy|x, \beta) \) satisfying

\[
R(\gamma^*(\cdot)\|\mu(\cdot|x)) = L(x, \beta) \quad \text{and} \quad \int_{\mathbb{R}^d} y \gamma^*(dy) = \beta.
\]

On the other hand, if \( \beta \in \mathbb{R}^d \) satisfies \( L(x, \beta) = \infty \), then by part (f) of Lemma 6.2.3 any \( \gamma \in \mathcal{P}(\mathbb{R}^d) \) having mean \( \beta \) — in particular, \( \delta_\beta \) — must satisfy \( R(\gamma(\cdot)\|\mu(\cdot|x)) = \infty \). Thus for \( x \) and \( \beta \) in \( \mathbb{R}^d \)

\[
\gamma(dy|x, \beta) \doteq \begin{cases} 
\gamma^*(dy|x, \beta) & \text{if } L(x, \beta) < \infty \\
\delta_\beta(dy) & \text{if } L(x, \beta) = \infty
\end{cases}
\]

satisfies

\[
R(\gamma(\cdot|x, \beta)\|\mu(\cdot|x)) = L(x, \beta) \quad \text{and} \quad \int_{\mathbb{R}^d} y \gamma(dy|x, \beta) = \beta.
\]

Our goal is to show that \( \gamma(dy|x, \beta) \) is a stochastic kernel on \( \mathbb{R}^d \) given \( \mathbb{R}^d \times \mathbb{R}^d \).

By Theorem A.5.3 this can be carried out by exhibiting \( \gamma(dy|x, \beta) \), for each \( x \) and \( \beta \) in \( \mathbb{R}^d \), as the weak limit of a sequence \( \{\gamma^n(dy|x, \beta), n \in \mathbb{N}\} \) of stochastic kernels on \( \mathbb{R}^d \) given \( \mathbb{R}^d \times \mathbb{R}^d \). Fix a sequence \( \{\epsilon_n, n \in \mathbb{N}\} \) in \((0, 1)\) which converges to 0, and for \( n \in \mathbb{N} \) define the stochastic kernel

\[
\mu^n(dy|x) \doteq \frac{1}{1 + \epsilon_n} \left[ \mu(dy|x) + \epsilon_n \rho(dy) \right],
\]

where

\[
\rho(A) \doteq \frac{1}{(2\pi)^{d/2}} \int_A \exp(-\|y\|^2/2) \, dy
\]

for Borel subsets \( A \) of \( \mathbb{R}^d \). For \( n \in \mathbb{N} \) and \( x, \alpha, \) and \( \beta \) in \( \mathbb{R}^d \) we also introduce

\[
H^n(x, \alpha) \doteq \log \int_{\mathbb{R}^d} \exp(\langle \alpha, y \rangle) \mu^n(dy|x), \quad \tilde{H}^n(x, \alpha) \doteq \log \int_{\mathbb{R}^d} \exp(\langle \alpha, y \rangle) \left[ \mu(dy|x) + \epsilon_n \rho(dy) \right],
\]

and

\[
L^n(x, \beta) \doteq \sup_{\alpha \in \mathbb{R}^d} \{ \langle \alpha, \beta \rangle - H^n(x, \alpha) \}.
\]

Since \( H^n(x, \alpha) = \tilde{H}^n(x, \alpha) - \log(1 + \epsilon_n) \) and \( \tilde{H}^n(x, \alpha) \downarrow H(x, \alpha) \) as \( n \to \infty \), it follows that if \( L(x, \beta) < \infty \), then for all \( n \in \mathbb{N} \)

\[
L^n(x, \beta) = \sup_{\alpha \in \mathbb{R}^d} \{ \langle \alpha, \beta \rangle - \tilde{H}^n(x, \beta) \} + \log(1 + \epsilon_n) \leq L(x, \beta) + \log 2 < \infty.
\]

For each \( n \in \mathbb{N} \) and \( x \in \mathbb{R}^d \) the support of \( \mu^n(dy|x) \) is all of \( \mathbb{R}^d \). Hence by parts (d) and (e) of Lemma 6.2.3, \( \text{dom} L^n(x, \cdot) = \mathbb{R}^d \) and for each \( \beta \in \mathbb{R}^d \) there exists a unique \( \alpha = \alpha^n(x, \beta) \in \mathbb{R}^d \) satisfying \( \nabla_\alpha H(x, \alpha^n(x, \beta)) = \beta \). For \( \alpha \in \mathbb{R}^d \) we denote by \( \mu^n_\alpha(dy|x) \) the probability measure on \( \mathbb{R}^d \) which is absolutely continuous with respect to \( \mu^n(dy|x) \) and satisfies

\[
\frac{d\mu^n_\alpha(dy|x)}{d\mu^n(dy|x)}(y) \doteq \exp(\langle \alpha, y \rangle) \cdot \frac{1}{\int_{\mathbb{R}^d} \exp(\langle \alpha, y \rangle) \mu^n(dy|x)}.\]
Lemma C.5.1 implies that $\gamma(dy) = \mu^\gamma_{\alpha^n(x,\beta)}(dy|x)$ is the unique probability measure on $\mathbb{R}^d$ satisfying

$$R(\gamma(\cdot)||\mu^n(\cdot|x)) = L^n(x,\beta) \quad \text{and} \quad \int_{\mathbb{R}^d} y \, \gamma(dy) = \beta.$$ 

For each $n \in \mathbb{N}$ we claim that

$$\gamma^n(dy|x,\beta) = \begin{cases} \mu^\gamma_{\alpha^n(x,\beta)}(dy|x) & \text{if } L(x,\beta) < \infty \\ \delta_\beta(dy) & \text{if } L(x,\beta) = \infty \end{cases}$$

is a stochastic kernel on $\mathbb{R}^d$ given $\mathbb{R}^d \times \mathbb{R}^d$ and that for each $x$ and $\beta$ in $\mathbb{R}^d$ $\gamma^n(dy|x,\beta) \Longrightarrow \gamma(dy|x,\beta)$. This will complete the proof of part (g).

The function mapping $(x,\beta) \mapsto \delta_\beta(dy)$ obviously defines a stochastic kernel on $\mathbb{R}^d$ given $\mathbb{R}^d \times \mathbb{R}^d$. In addition, $L(x,\beta)$ is a measurable function of $(x,\beta) \in \mathbb{R}^d \times \mathbb{R}^d$, and for fixed $\alpha \in \mathbb{R}^d$ $\mu^\gamma_{\alpha^n}(dy|x)$ is a stochastic kernel on $\mathbb{R}^d$ given $\mathbb{R}^d$. Hence in order to prove that $\gamma^n(dy|x,\beta)$ is a stochastic kernel on $\mathbb{R}^d$ given $\mathbb{R}^d \times \mathbb{R}^d$, it suffices to show that the function mapping $(x,\beta) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto \alpha^n(x,\beta)$ is measurable. Since for each $x \in \mathbb{R}^d$ the support of $\mu^\gamma(dy|x)$ is all of $\mathbb{R}^d$, $L^n(x,\cdot)$ is differentiable on $\mathbb{R}^d$ [Lemma 6.2.3 (d)–(e)]. The equality $\nabla_{\alpha} H^n(x,\alpha^n(x,\beta)) = \beta$ combined with Legendre–Fenchel duality [Theorem D.2.6 (d)] and part (b) of Theorem D.2.5 imply that $\alpha^n(x,\beta) = \nabla_{\beta} L^n(x,\beta)$. We conclude that the function mapping $(x,\beta) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto \alpha^n(x,\beta)$ is measurable.

For $x$ and $\beta$ in $\mathbb{R}^d$ satisfying $L(x,\beta) = \infty$, $\gamma^n(dy|x,\beta) = \delta_\beta(dy) = \gamma(dy|x,\beta)$ for all $n \in \mathbb{N}$. Hence in order to complete the proof of part (g), it remains to show that if $L(x,\beta) < \infty$, then $\mu^\gamma_{\alpha^n}(dy|x) \Longrightarrow \gamma^\star(dy|x,\beta)$. Since $H(x,\alpha) < \infty$ for each $\alpha \in \mathbb{R}^d$, it follows that

$$\sup_{n \in \mathbb{N}} \int_{\mathbb{R}^d} \exp\langle \alpha, y \rangle \mu^n(dy|x) < \infty.$$ 

In addition

$$\sup_{n \in \mathbb{N}} R(\mu^\gamma_{\alpha^n}(\cdot|x)||\mu^n(\cdot|x)) = \sup_{n \in \mathbb{N}} L^n(x,\beta) \leq L(x,\beta) + \log 2 < \infty.$$ 

Hence by part (d) of Lemma 1.4.3 the sequence $\{\mu^\gamma_{\alpha^n}(dy|x), n \in \mathbb{N}\}$ is tight and uniformly integrable. Given any subsequence, there exists a subsubsequence and $\sigma(dy|x,\beta) \in \mathcal{P}(\mathbb{R}^d)$ such that [Theorem A.3.19 (a)]

$$\mu^\gamma_{\alpha^n}(dy|x) \Longrightarrow \sigma(dy|x,\beta) \quad \text{and} \quad \lim_{n \to \infty} \int_{\mathbb{R}^d} y \, \mu^\gamma_{\alpha^n}(dy|x) = \int_{\mathbb{R}^d} y \, \sigma(dy|x,\beta) = \beta.$$ 

Since $\mu^n(dy|x) \Longrightarrow \mu(dy|x)$, the lower semicontinuity of $R(\cdot|\cdot)$ implies that

$$R(\sigma(\cdot|x,\beta)||\mu(\cdot|x)) \leq \liminf_{n \to \infty} R(\mu^\gamma_{\alpha^n}(\cdot|x)||\mu^n(\cdot|x)) \leq \limsup_{n \to \infty} L^n(x,\beta) \leq L(x,\beta),$$

and since the mean of $\sigma(dy|x,\beta)$ equals $\beta$, part (f) of Lemma 6.2.3 implies in fact that $R(\sigma(\cdot|x)||\mu(\cdot|x)) = L(x,\beta)$. According to Lemma C.5.1, $\gamma^\star(dy|x,\beta)$ is the unique probability measure on $\mathbb{R}^d$ satisfying this equality and having mean $\beta$. Hence $\sigma(dy|x,\beta) = \gamma^\star(dy|x,\beta)$. We conclude that any subsequence of $\{\mu^\gamma_{\alpha^n}(dy|x), n \in \mathbb{N}\}$ has a subsubsequence converging weakly to $\gamma^\star(dy|x,\beta)$. An argument by contradiction applied to an arbitrary subsequence shows that the entire sequence converges weakly to this measure. The proof of part (g) of Lemma 6.2.3 is complete. ■
C.7 Proof of Proposition 6.3.4

Part (a) of Proposition 6.3.4 gives a large collection of stochastic kernels $\mu(dy|x)$ for which Conditions 6.2.1 and 6.3.2 hold. Parts (b) and (c) show how one can construct, from a stochastic kernel $\mu(dy|x)$ for which the conditions hold, additional stochastic kernels for which the conditions hold. We now prove these three parts of the proposition. Let us recall that for any stochastic kernel $\mu(dy|x)$ on $\mathcal{R}^d$ given $\mathcal{R}^d$ satisfying Condition 6.2.1, we have defined for $x$, $\alpha$, and $\beta$ in $\mathcal{R}^d$

$$H(x, \alpha) \doteq \log \int_{\mathcal{R}^d} \exp(\alpha, y) \mu(dy|x) \quad \text{and} \quad L(x, \beta) \doteq \sup_{\alpha \in \mathcal{R}^d} \{\langle \alpha, \beta \rangle - H(x, \alpha)\}.$$ 

**Part (a).** Equation (6.18) defines a stochastic kernel $\mu(dy|x)$ on $\mathcal{R}^d$ given $\mathcal{R}^d$ in terms of a probability measure $\mu$ that has a finite moment generating function. A short calculation shows that

$$H(x, \alpha) = \langle \alpha, b(x) \rangle + H(\Sigma(x)\alpha),$$

where

$$H(\alpha) \doteq \log \int_{\mathcal{R}^d} \exp(\alpha, y) \mu(dy).$$

Under the assumptions on $\mu$, $\Sigma(x)$, and $b(x)$, it follows that for each $\alpha \in \mathcal{R}^d$

$$\sup_{x \in \mathcal{R}^d} H(x, \alpha) < \infty.$$ 

One easily checks that if $x_n \to x$, then $\mu(\cdot|x_n) \Rightarrow \mu(\cdot|x)$. Thus Condition 6.2.1 holds.

We now verify Condition 6.3.2 by first noting that for $x$ and $\beta$ in $\mathcal{R}^d$

$$L(x, \beta) = L\left(\left(\Sigma(x)^T\right)^{-1}[\beta - b(x)]\right),$$

where

$$L(\beta) \doteq \sup_{\beta \in \mathcal{R}^d} \{\langle \alpha, \beta \rangle - H(\alpha)\}.$$ 

Let any compact subset $\Delta$ of $\mathcal{R}^d$ be given as well as any points $\xi \in \Delta$, $v \in \Delta$, and $\gamma \in \mathcal{R}^d$. We define $\bar{\beta} \in \mathcal{R}^d$ by

$$\left(\Sigma(\xi)^T\right)^{-1}[\bar{\beta} - b(\xi)] = \left(\Sigma(v)^T\right)^{-1}[\gamma - b(v)].$$

With this definition we have $L(\xi, \bar{\beta}) = L(v, \gamma)$, and consequently the upper bound (6.16) in Condition 6.3.2 holds for any $\varepsilon > 0$. Furthermore,

$$\bar{\beta} - \gamma = \left[\Sigma(\xi)^T - \Sigma(v)^T\right]\left(\Sigma(v)^T\right)^{-1}[\gamma - b(v)] + [b(\xi) - b(v)].$$

Since $L(\cdot, \cdot)$ is uniformly superlinear [Lemma 6.2.3 (c)] and $\Sigma(x)$ and $b(x)$ are Lipschitz continuous, the last display implies that there exist constants $K < \infty$ and $\eta > 0$ such that

$$\|\bar{\beta} - \gamma\| \leq K\|\xi - v\| \left[1 + L(v, \gamma)\right] \quad \text{whenever} \quad \|\xi - v\| \leq \eta.$$
This yields the upper bound (6.17) in Condition 6.3.2, completing the verification of the condition. The proof of part (a) is complete.

**Part (b).** It is easily verified that Condition 6.2.1 holds for the stochastic kernel

\[ \mu(dy|x) = \sum_{j=1}^{N} \rho_j \mu_j(dy|x). \]

By an induction argument, we can reduce the verification of Condition 6.3.2 to the case \( N = 2 \). Thus we have two stochastic kernels \( \mu_1(dy|x) \) and \( \mu_2(dy|x) \) on \( \mathbb{R}^d \) given \( \mathbb{R}^d \) for which Conditions 6.2.1 and 6.3.2 hold as well as positive real numbers \( \rho_1 \) and \( \rho_2 \) summing to 1. For \( j = 1, 2 \) and \( x, \alpha, \) and \( \beta \) in \( \mathbb{R}^d \) define

\[ H_j(x, \alpha) = \log \int_{\mathbb{R}^d} \exp(\alpha, \beta) \mu_j(dy|x), \quad L_j(x, \beta) = \sup_{\alpha \in \mathbb{R}^d} \{ \langle \alpha, \beta \rangle - H_j(x, \alpha) \}, \]

and \( \mu(dy|x) = \rho_1 \mu_1(dy|x) + \rho_2 \mu_2(dy|x). \)

For \( x \) and \( \beta \) in \( \mathbb{R}^d \), \( L(x, \beta) \) will be expressed in terms of \( L_1(x, \beta), L_2(x, \beta), \) and

\[ R(u) = u \log \left( \frac{u}{\rho_1} \right) + (1-u) \log \left( \frac{1-u}{\rho_2} \right) \quad \text{for } u \in [0, 1]. \]

(C.17)

\( R(u) \) is the relative entropy of the probability vector \((u, 1-u)\) with respect to the probability vector \((\rho_1, \rho_2)\), and \( R(u) \geq 0 \) for each \( u \in [0, 1] \). Since

\[ H(x, \alpha) = \log (\rho_1 \exp[H_1(x, \alpha)] + \rho_2 \exp[H_2(x, \alpha)]), \]

Corollary D.4.1 implies that

\[ L(x, \beta) = \inf \{ R(u) + uL_1(x, \beta_1) + (1-u)L_2(x, \beta_2) : u \in [0, 1], \beta_1 \in \mathbb{R}^d, \beta_2 \in \mathbb{R}^d, u\beta_1 + (1-u)\beta_2 = \beta \}. \]

(C.18)

Let any compact subset \( \Delta \) of \( \mathbb{R}^d \) be given as well as any \( \varepsilon \in (0, 1) \) and points \( \xi \in \Delta, \, v \in \Delta, \) and \( \gamma \in \mathbb{R}^d \). The representation (C.18) yields \( u, \beta_1, \) and \( \beta_2 \) such that \( u\beta_1 + (1-u)\beta_2 = \gamma \) and

\[ R(u) + uL_1(v, \beta_1) + (1-u)L_2(v, \beta_2) \leq L(v, \gamma) + \varepsilon/3. \]

By hypothesis, Condition 6.3.2 holds for \( \mu_1(dy|x) \) and \( \mu_2(dy|x) \). Hence for each \( j = 1, 2 \), there exists \( \eta_j > 0 \) such that whenever \( \|\xi - v\| \leq \eta_j \), there are \( K_j < \infty \) and \( \beta_j \in \mathbb{R}^d \) for which

\[ L_j(\xi, \tilde{\beta}_j) - L_j(v, \beta_j) \leq \frac{\varepsilon}{3} [1 + L_j(v, \beta_j)] \]

and

\[ \|\tilde{\beta}_j - \beta_j\| \leq K_j \|\xi - v\| [1 + L_j(v, \beta_j)]. \]
C.8. A CONTINUITY PROPERTY OF CRAMÉR FUNCTIONS

We then define $\tilde{\beta} = u\tilde{\beta}_1 + (1 - u)\beta_2$ and $\eta = \eta_1 \land \eta_2$. Whenever $\|\xi - v\| \leq \eta$,

$$L(\xi, \tilde{\beta}) \leq R(u) + uL_1(\xi, \tilde{\beta}_1) + (1 - u)L_2(\xi, \beta_2)$$
$$\leq R(u) + u \left( L_1(v, \beta_1) + \frac{\varepsilon}{3} [1 + L_1(v, \beta_1)] \right)$$
$$\quad + (1 - u) \left( L_2(v, \beta_2) + \frac{\varepsilon}{3} [1 + L_2(v, \beta_2)] \right)$$
$$\leq L(v, \gamma) + \frac{2\varepsilon}{3} + \frac{\varepsilon}{3} \left( L(v, \gamma) + \frac{\varepsilon}{3} \right)$$
$$\leq L(v, \gamma) + \varepsilon \left[ 1 + L(v, \gamma) \right].$$

Thus the upper bound (6.16) in Condition 6.3.2 holds. In addition, whenever $\|\xi - v\| \leq \eta$,

$$\|\tilde{\beta} - \gamma\| = \|u(\tilde{\beta}_1 - \beta_1) + (1 - u)(\beta_2 - \beta_2)\|$$
$$\leq uK_1 \|\xi - v\| [1 + L_1(v, \beta_1)] + (1 - u)K_2 \|\xi - v\| [1 + L_2(v, \beta_2)]$$
$$\leq (K_1 \lor K_2) \|\xi - v\| [1 + uL_1(v, \beta_1) + (1 - u)L_2(v, \beta_2)]$$
$$\leq (K_1 \lor K_2) \|\xi - v\| \left[ 1 + L(v, \gamma) + \frac{\varepsilon}{3} \right]$$
$$\leq 2(K_1 \lor K_2) \|\xi - v\| \left[ 1 + L(v, \gamma) \right].$$

Therefore, with $K = 2(K_1 \lor K_2)$ the upper bound (6.17) in Condition 6.3.2 holds as well. This completes the verification of Condition 6.3.2 and hence the proof of part (b).

Part (c). For $\mu(\cdot | x) = (\mu_1 * \mu_2)(\cdot | x)$

$$H(x, \alpha) = H_1(x, \alpha) + H_2(x, \alpha) = \log \int_{\mathbb{R}^d} \exp(\alpha, y) \mu_1(dy | x) + \log \int_{\mathbb{R}^d} \exp(\alpha, y) \mu_2(dy | x).$$

It is easily verified that Condition 6.2.1 holds. According to Corollary D.4.2, for $x$ and $\beta$ in $\mathbb{R}^d$

$$L(x, \beta) = \inf\{ L_1(x, \beta_1) + L_2(x, \beta_2) : \beta_1 \in \mathbb{R}^d, \beta_2 \in \mathbb{R}^d, \beta_1 + \beta_2 = \beta \}, \quad (C.19)$$

where for $j = 1, 2$ and $\gamma \in \mathbb{R}^d$

$$L_j(x, \gamma) = \sup_{\alpha \in \mathbb{R}^d} \{ \langle \alpha, \gamma \rangle - H_j(x, \alpha) \}.$$

Using calculations similar to those given in the proof of part (b), one verifies Condition 6.3.2. The details are omitted. ■

C.8 A Continuity Property of Cramér Functions

Several results in the book assert continuity properties of functions that are defined as the Legendre–Fenchel transforms of cumulant generating functions. We refer to these Legendre–Fenchel transforms collectively as Cramér functions.
Part (b) of Lemma 6.5.2 states that under Condition 6.2.1 and part (a) of Condition 6.3.1 \( L(x, \beta) \) is a continuous function of \( (x, \beta) \in \mathbb{R}^d \times \Sigma \); i.e., if \( (\xi_n, \beta_n) \to (\xi, \beta) \), then

\[
\lim_{n \to \infty} L(\xi_n, \beta_n) = L(\xi, \beta). \tag{C.20}
\]

From the discussion at the beginning of Section 6.5, without loss of generality we assume that \( \Sigma \subseteq \text{ri}(\text{conv } S_\mu(x)) \) is a nonempty open convex subset of \( \mathbb{R}^d \). Hence, according to part (a) of Lemma 6.5.2, for all \( x \in \mathbb{R}^d \) \( \Sigma \) equals \( \text{int}(\text{dom } L(x, \cdot)) \), which is the (nonempty) interior of the effective domain of \( L(x, \cdot) \). Under Condition 7.2.1 and part (a) of Condition 7.2.2, part (b) of Lemma 7.5.2 asserts the same continuity property of \( L^{(1)}(x, \beta) \) and \( L^{(2)}(x, \beta) \), while under Conditions 6.2.1 and 6.3.2, part (b) of Lemma 6.6.2 asserts that \( L_\sigma(x, \beta) \) is a continuous function of \( (x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d \).

In all these cases, the continuity follows from Lemma C.8.1. For example, the continuity of \( L(x, \beta) \) asserted in part (b) of Lemma 6.5.2 is a consequence of the following calculation. Let \( \{(\xi_n, \beta_n), n \in \mathbb{N}\} \) be any sequence converging to \((\xi, \beta)\) and define \( g_n(\alpha) = H(\xi_n, \alpha) \) and \( g(\alpha) = H(\xi, \alpha) \) for \( \alpha \in \mathbb{R}^d \). According to part (a) of Lemma 6.2.3, these functions are all differentiable in \( \alpha \), and for any \( \alpha \)

\[
\lim_{n \to \infty} g_n(\alpha) = \lim_{n \to \infty} H(\xi_n, \alpha) = H(\xi, \alpha) = g(\alpha).
\]

In addition, the Legendre–Fenchel transforms \( g^* \) and \( g_n^* \) equal \( L(\xi_n, \cdot) \) and \( L(\xi, \cdot) \), respectively. The limit (C.20) now follows from Lemma C.8.1.

We suspect that the following lemma already appears in the literature, but have been unable to locate it. The closest that we could come to finding a continuity property is the paper [73]. The lemma is also applied in a related but slightly different context in Section 6.7.

**Lemma C.8.1.** Let \( \{g_n, n \in \mathbb{N}\} \) and \( g \) be convex functions which are finite and differentiable on \( \mathbb{R}^d \) and have respective Legendre–Fenchel transforms \( \{g_n^*, n \in \mathbb{N}\} \) and \( g^* \). We assume that \( \text{int}(\text{dom } g^*) \), the interior of the effective domain of \( g^* \), is nonempty and that for all \( \alpha \in \mathbb{R}^d \)

\[
\lim_{n \to \infty} g_n(\alpha) = g(\alpha).
\]

Then for any sequence \( \{\beta_n, n \in \mathbb{N}\} \) converging to \( \beta \in \text{int}(\text{dom } g^*) \)

\[
\lim_{n \to \infty} g_n^*(\beta_n) = g^*(\beta).
\]

**Remark C.8.2.** Under the assumption that \( g_n(\alpha) \to g(\alpha) \) for all \( \alpha \in \mathbb{R}^d \), the conclusion of the lemma is valid even if \( g_n \) and \( g \) are not differentiable on \( \mathbb{R}^d \). A proof can be based upon Theorem 1 in [73]. Related material appears in [94].

**Proof of Lemma C.8.1.** For any sequence \( \{\zeta_n, n \in \mathbb{N}\} \) converging to \( \zeta \in \mathbb{R}^d \)

\[
\lim_{n \to \infty} \inf \{g_n^*(\zeta_n) \geq g^*(\zeta) \}.
\]
This is immediate from the representations of $g^*_n$ and $g^*$ as Legendre–Fenchel transforms and from the assumed pointwise convergence of $g_n$ to $g$. In particular, for the sequence \( \{\beta_n, n \in \mathbb{N}\} \) appearing in the statement of this lemma

\[
\liminf_{n \to \infty} g^*_n(\beta_n) \geq g^*(\beta).
\]

In order to prove the reverse inequality

\[
\limsup_{n \to \infty} g^*_n(\beta_n) \leq g^*(\beta), \tag{C.21}
\]

we need the fact that for any \( \zeta \in \text{int}(\text{dom } g^*) \) there exists a sequence \( \{r_n, n \in \mathbb{N}\} \) such that

\[
\lim_{n \to \infty} r_n = \zeta \quad \text{and} \quad \lim_{n \to \infty} g^*_n(r_n) = g^*(\zeta).
\]

Indeed, Theorem D.2.7 and part (c) of Theorem D.2.6 guarantee the existence of \( \alpha \in \mathbb{R}^d \) such that

\[
\zeta = \nabla g(\alpha) \quad \text{and} \quad g^*(\zeta) = \langle \alpha, \zeta \rangle - g(\alpha).
\]

The points \( r_n \doteq \nabla g_n(\alpha) \) satisfy [Theorem D.2.4]

\[
\lim_{n \to \infty} r_n = \zeta \quad \text{and} \quad \lim_{n \to \infty} g^*_n(r_n) = \lim_{n \to \infty} (\langle \alpha, r_n \rangle - g_n(\alpha)) = \langle \alpha, \zeta \rangle - g(\alpha) = g^*(\zeta).
\]

We now prove (C.21) via contradiction. Thus suppose that there exists a sequence \( \{\beta_n, n \in \mathbb{N}\} \) such that \( \lim_{n \to \infty} \beta_n = \beta \in \text{int}(\text{dom } g^*) \) but

\[
\limsup_{n \to \infty} g^*_n(\beta_n) > g^*(\beta). \tag{C.22}
\]

We claim that there exists a closed ball \( C \) with nonempty interior and center \( \beta \) which for all sufficiently large \( n \) is contained in \( \text{int}(\text{dom } g^*_n) \). Indeed, we may choose points \( \{w^i, i = 1, 2, \ldots, 2^d\} \) in \( \text{int}(\text{dom } g^*) \) such that \( \beta \) lies in the interior of the convex hull of \( \{w^i\} \). As pointed out in the previous paragraph, for each \( i \in \{1, 2, \ldots, 2^d\} \) there exists a sequence \( \{w^i_n, n \in \mathbb{N}\} \) such that \( w^i_n \to w^i \) and

\[-\infty < \lim_{n \to \infty} g^*_n(w^i_n) = g^*(w^i) < \infty.\]

Thus for each \( i \in \{1, 2, \ldots, 2^d\} \), each \( \theta > 0 \), and all sufficiently large \( n \)

\[g^*(w^i) - \theta \leq g^*_n(w^i_n) \leq g^*(w^i) + \theta.\]

This implies that there exists a closed ball \( B \) with nonempty interior and center \( \beta \) and a positive integer \( N \) such that \( B \) is contained in the convex hull of \( \{w^i, i = 1, 2, \ldots, 2^d\} \) for all \( n \geq N \) and

\[-\infty < a \leq \inf\{g^*_n(y) : n \geq N, y \in B\} \leq \sup\{g^*_n(y) : n \geq N, y \in B\} \leq b < \infty.\]

The claim now follows for all \( n \geq N \) if we take \( C \) to be any closed ball with nonempty interior and center \( \beta \) which is a proper subset of \( B \). Since a convex function is continuous when restricted to the interior of its effective domain [Theorem D.2.2 (a)], each \( g^*_n \) restricted to \( C \) is continuous.
APPENDIX C. PROOFS OF A NUMBER OF RESULTS

Let $U$ be the closed unit ball in $\mathbb{R}^d$ centered at the origin and pick $\delta > 0$ so small that $C + \delta U \subset B$. Theorem D.2.3 states that for any two points $x$ and $y$ in $C$

$$\sup_{n \geq N} |g_n^*(x) - g_n^*(y)| \leq \frac{b-a}{\delta} \|x - y\|. \quad (C.23)$$

By the assumed lower bound (C.22), there exist $\varepsilon > 0$ and a subsequence of $\{\beta_n\}$ such that

$$g_n^*(\beta_n) \geq g^*(\beta) + \varepsilon.$$ 

In addition, since $\beta \in \text{int}(\text{dom} \ g^*)$, there exists a sequence $\{r_n, n \in \mathbb{N}\}$ such that

$$r_n \to \beta \quad \text{and} \quad \lim_{n \to \infty} g_n^*(r_n) = g^*(\beta).$$

We conclude that for all sufficiently large $n$

$$g_n^*(\beta_n) - g_n^*(r_n) \geq g^*(\beta) + \varepsilon/2 - g^*(\beta) = \varepsilon/2.$$

On the other hand, (C.23) implies that

$$\lim_{n \to \infty} (g_n^*(\beta_n) - g_n^*(r_n)) = 0.$$ 

This contradiction proves the upper bound (C.21) and completes the proof of the lemma. 

$\blacksquare$
Appendix D

Convex Functions

D.1 Introduction

Section D.2 of this appendix presents background material on convex functions. In Section D.3 we prove a theorem concerning the Legendre–Fenchel transform of compositions of convex functions. In the final section we derive from this theorem three interesting corollaries. Throughout this appendix references are made to the book of Rockafellar [79].

D.2 Background Material on Convex Functions

After making several definitions for a general linear space, we specialize to the case \( \mathbb{R}^d \). A subset \( \mathcal{X} \) of a linear space is called convex if \( sx + (1 - s)y \) lies in \( \mathcal{X} \) for all \( s \in (0, 1) \) and all \( x \) and \( y \) in \( \mathcal{X} \). A function \( f \) mapping a convex set \( \mathcal{X} \) into the extended real numbers is said to be convex if \( f(x) < \infty \) for at least one \( x \in \mathcal{X} \), \( f(x) > -\infty \) for all \( x \in \mathcal{X} \), and for all \( s \in (0, 1) \), \( x \in \mathcal{X} \), and \( y \in \mathcal{X} \)

\[
f(sx + (1 - s)y) \leq s f(x) + (1 - s) f(y).
\]

For the case \( \mathcal{X} = \mathbb{R}^d \) this definition coincides with the concept “proper convex function” as defined on page 24 of [79]. A function \( f \) mapping a convex set \( \mathcal{X} \) into the extended real numbers is said to be concave if \( -f \) is convex. The effective domain of a convex function on a convex set \( \mathcal{X} \) is the set

\[
\text{dom } f = \{ x \in \mathcal{X} : f(x) < \infty \}.
\]

A function \( f \) mapping a convex set \( \mathcal{X} \) into \( \mathbb{R} \) is said to be strictly convex on \( \mathcal{X} \) if for all \( s \in (0, 1) \) and distinct points \( x \) and \( y \)

\[
f(sx + (1 - s)y) < s f(x) + (1 - s) f(y).
\]

We need some definitions and notation concerning convex sets \( C \) in \( \mathbb{R}^d \). The closure of \( C \) is denoted by \( \text{cl } C \) and the interior of \( C \) by \( \text{int } C \). The affine hull of \( C \), \( \text{aff } C \), is
defined to be the intersection of all the affine sets that contain \( C \). We define the relative interior of \( C \), \( \text{ri} C \), as the interior that results when \( C \) is regarded as a subset of \( \text{aff} C \). The relative boundary of \( C \), \( \text{rbd} C \), is defined as the set difference \( \text{cl} C \setminus \text{ri} C \). If \( A \) is a subset of \( \mathbb{R}^d \), then the convex hull of \( A \), \( \text{conv} A \), is defined as the intersection of all the convex sets that contain \( A \). The closed convex hull of \( A \), \( \text{cc} A \), is defined as the intersection of all the closed convex sets that contain \( A \).

The first theorem states that the relative interior of the closed convex hull of a set \( A \) equals the relative interior of the convex hull of \( A \). This fact is needed in the proof of part (d) of Lemma 6.2.3.

**Theorem D.2.1.** For any subset \( A \) of \( \mathbb{R}^d \) \( \text{ri} (\text{cc} A) \) equals \( \text{ri} (\text{conv} A) \).

**Proof.** By Theorem 6.5 in [79]

\[
\text{cc} A = \bigcap \{ B : B \text{ closed convex in } \mathbb{R}^d, A \subset B \}
\]

\[
= \bigcap \{ \text{cl} C : C \text{ convex in } \mathbb{R}^d, A \subset C \}
\]

\[
= \text{cl} \left( \bigcap \{ C : C \text{ convex in } \mathbb{R}^d, A \subset C \} \right)
\]

\[
= \text{cl} (\text{conv} A).
\]

Theorem 6.3 in [79] now guarantees that \( \text{ri} (\text{cc} A) = \text{ri} (\text{cl} (\text{conv} A)) = \text{ri} (\text{conv} A) \). ■

Parts (a), (b), and (c) of the next theorem give continuity properties of convex functions on \( \mathbb{R}^d \), and part (d) is a refinement for convex functions on \( \mathbb{R} \).

**Theorem D.2.2.** (a) A convex function \( f \) on \( \mathbb{R}^d \) is continuous relative to \( \text{ri} (\text{dom} f) \).

(b) Let \( f \) be a finite convex function on \( \mathbb{R}^d \). Then \( f \) is continuous on \( \mathbb{R}^d \).

(c) Let \( f \) be a convex, lower semicontinuous function on \( \mathbb{R}^d \). Let \( y \) be a point in \( \text{rbd} (\text{dom} f) \) and \( x \) a point in \( \text{ri} (\text{dom} f) \). Then for each \( s \in (0,1] \) \( sx + (1-s)y \) lies in \( \text{ri} (\text{dom} f) \) and

\[
\lim_{s \to 0} f(sx + (1-s)y) = f(y).
\]

(d) Let \( f \) be a convex, lower semicontinuous function on \( \mathbb{R} \) which is finite on the closed bounded interval \([a,b] \). Then \( f \) is continuous on \([a,b] \).

**Proof.** (a) This is proved in Theorem 10.1 in [79].

(b) This follows from part (a).

(c) For \( y \in \text{rbd} (\text{dom} f) \), \( x \in \text{ri} (\text{dom} f) \), and \( s \in (0,1] \), Theorem 6.1 in [79] proves that \( sx + (1-s)y \in \text{ri} (\text{dom} f) \). The convexity of \( f \) yields

\[
\limsup_{s \to 0} f(sx + (1-s)y) \leq \limsup_{s \to 0} [sf(x) + (1-s)f(y)] = f(y),
\]

and the lower semicontinuity of \( f \) yields

\[
\liminf_{s \to 0} f(sx + (1-s)y) = \liminf_{s \to 0} f(y + s(x - y)) \geq f(y).
\]
(d) This follows from parts (a) and (c). \[\blacksquare\]

Theorem 10.4 in [79] proves that a convex function on \(\mathbb{R}^d\) is Lipschitz continuous relative to any closed bounded subset of \(\text{ri}(\text{dom } f)\). For an application in Section C.8 we need more precise information on the Lipschitz constant of \(f\). This information is readily available from the proof of Theorem 10.4 in [79] and is given in the next theorem.

**Theorem D.2.3.** Let \(f\) be a convex function on \(\mathbb{R}^d\) such that the interior of \(\text{dom } f\) is nonempty. Let \(C\) be any closed bounded subset of \(\text{int}(\text{dom } f)\) and let \(U\) be the closed unit ball in \(\mathbb{R}^d\) centered at the origin. Then there exists \(\delta > 0\) such that \(C + \delta U \subset \text{int}(\text{dom } f)\) and \(f\) restricted to \(C + \delta U\) is bounded and continuous. If
\[
\alpha_1 \doteq \inf \{ f(x) : x \in C + \delta U \}
\]
and
\[
\alpha_2 \doteq \sup \{ f(x) : x \in C + \delta U \},
\]
then for all \(x\) and \(y\) in \(C\)
\[
|f(x) - f(y)| \leq \frac{\alpha_2 - \alpha_1}{\delta} \|x - y\|.
\]

The next result is a strong convergence theorem for sequences of differentiable convex functions. It is proved in Theorem 25.7 in [79].

**Theorem D.2.4.** Let \(C\) be an open convex set in \(\mathbb{R}^d\) and \(f\) a convex function which is finite and differentiable on \(C\). Let \(\{f_n, n \in \mathbb{N}\}\) be a sequence of convex functions that are finite and differentiable on \(C\) and satisfy \(\lim_{n \to \infty} f_n(x) = f(x)\) for every \(x \in C\). Then for every \(x \in C\)
\[
\lim_{n \to \infty} \nabla f_n(x) = \nabla f(x).
\]

Let \(f\) be a convex function on \(\mathbb{R}^d\). If \(y\) is a point in \(\mathbb{R}^d\), then \(z \in \mathbb{R}^d\) is called a subgradient of \(f\) at \(y\) if
\[
f(x) \geq f(y) + \langle z, x - y \rangle \quad \text{for all } x \in \mathbb{R}^d.
\]

We define the subdifferential of \(f\) at \(y\) to be the set
\[
\partial f(y) \doteq \{ z \in \mathbb{R}^d : z \text{ is a subgradient of } f \text{ at } y \}.
\]

Parts (a) and (b) of the next theorem are proved in Theorems 23.4 and 25.1 in [79].

**Theorem D.2.5.** Let \(f\) be a convex function on \(\mathbb{R}^d\). The following conclusions hold.
(a) For \(y \in \text{ri}(\text{dom } f)\) \(\partial f(y)\) is nonempty, while for \(y \notin \text{dom } f\) \(\partial f(y)\) is empty.
(b) If \(f\) is differentiable at \(y \in \text{dom } f\), then \(\nabla f(y)\) is the unique subgradient of \(f\) at \(y\).

Given \(f\) a convex, lower semicontinuous function on \(\mathbb{R}^d\) we define its Legendre–Fenchel transform
\[
f^*(y) \doteq \sup_{x \in \mathbb{R}^d} \{ \langle x, y \rangle - f(x) \} \quad \text{for } y \in \mathbb{R}^d.
\]

A convex, lower semicontinuous function and its Legendre–Fenchel transform are dual in a sense that is made precise in the following result, which is proved in Theorems 12.2 and 23.5 in [79].
Theorem D.2.6. Let \( f \) be a convex, lower semicontinuous function on \( \mathbb{R}^d \) with Legendre–Fenchel transform \( f^* \). The following conclusions hold.

(a) \( f^* \) is a convex, lower semicontinuous function on \( \mathbb{R}^d \).

(b) \( \langle x, y \rangle \leq f(x) + f^*(y) \) for all \( x \) and \( y \) in \( \mathbb{R}^d \).

(c) \( \langle x, y \rangle = f(x) + f^*(y) \) if and only if \( y \in \partial f(x) \).

(d) \( y \in \partial f(x) \) if and only if \( x \in \partial f^*(y) \).

(e) \( f^{**} = f \); i.e., \( f(x) = \sup_{y \in \mathbb{R}^d} \{ \langle x, y \rangle - f^*(y) \} \) for all \( x \in \mathbb{R}^d \).

The range of the gradient of a differentiable convex function on \( \mathbb{R}^d \) is described as follows in Corollary 26.41 in [79].

Theorem D.2.7. Let \( f \) be a differentiable convex function on \( \mathbb{R}^d \). Then for any \( \zeta \in \text{ri}(\text{dom} \ f^*) \) there exists \( y \in \mathbb{R}^d \) such that \( \nabla f(y) = \zeta \).

Lemma 6.2.3 establishes a number of properties of the cumulant generating function \( H(x, \alpha) \) of a stochastic kernel and a number of properties of the Legendre–Fenchel transform \( L(x, \beta) \). The next theorem, a special case of Theorem 26.3 in [79], shows that the strict convexity of \( H(x, \cdot) \) on \( \mathbb{R}^d \) implies a smoothness property of \( L(x, \cdot) \).

Theorem D.2.8. If \( f \) is strictly convex on \( \mathbb{R}^d \), then \( \text{int}(\text{dom} \ f^*) \) is nonempty and \( f^* \) is differentiable on \( \text{int}(\text{dom} \ f^*) \).

The last result in this section describes how various finiteness assumptions on a convex, lower semicontinuous function are reflected in growth properties of its Legendre–Fenchel transform.

Theorem D.2.9. Let \( f \) be a convex, lower semicontinuous function on \( \mathbb{R}^d \). The following implications hold.

(a) If \( 0 \in \text{dom} \ f \), then for all \( \beta \in \mathbb{R}^d \) \( f^*(\beta) \geq -f(0) > -\infty \); i.e., \( f^* \) is bounded below.

(b) If \( 0 \in \text{int}(\text{dom} \ f) \), then for each \( M < \infty \) the set \( \{ \beta \in \mathbb{R}^d : f^*(\beta) \leq M \} \) is compact; i.e., \( f^* \) has compact level sets.

(c) \( f \) is finite on \( \mathbb{R}^d \) if and only if \( f^* \) is superlinear.

Proof. (a) For all \( \beta \in \mathbb{R}^d \)
\[
f^*(\beta) \geq \langle 0, \beta \rangle - f(0) = -f(0) > -\infty.
\]

(b) We denote \( \{ \beta \in \mathbb{R}^d : f^*(\beta) \leq M \} \) by \( Z(M) \). \( Z(M) \) is closed since \( f^* \) is lower semicontinuous. If \( \beta \) is in \( Z(M) \), then for any \( \alpha \in \mathbb{R}^d \)
\[
\langle \alpha, \beta \rangle \leq f(\alpha) + f^*(\beta) \leq f(\alpha) + M.
\]
Choose \( R > 0 \) such that \( \{ \alpha \in \mathbb{R}^d : \|\alpha\| \leq R \} \) is a subset of \( \text{int}(\text{dom} \ f) \). By part (a) of Theorem D.2.2, there exists a constant \( \Gamma \) independent of \( \beta \in Z(M) \) such that
\[
\sup_{\|\alpha\| \leq R} \langle \alpha, \beta \rangle = R \|\beta\| \leq \sup_{\|\alpha\| \leq R} f(\alpha) + M \leq \Gamma < \infty.
\]
This implies that $Z(M)$ is bounded and thus compact.

(c) If $f$ is finite on $\mathbb{R}^d$, then $f$ is continuous [Theorem D.2.2 (b)]. For $M \in (0, \infty)$ and nonzero $\beta \in \mathbb{R}^d$, define $\alpha \doteq M\beta/\|\beta\|$. Then

$$f^*(\beta) \geq \langle \alpha, \beta \rangle - f(\alpha) = M\|\beta\| - f(M\beta/\|\beta\|),$$

and so for all sufficiently large $N \in (0, \infty)$

$$\inf_{\{\beta \in \mathbb{R}^d : \|\beta\| = N\}} \frac{1}{\|\beta\|} f^*(\beta) \geq M - \frac{1}{N} \sup_{\{\beta \in \mathbb{R}^d : \|\beta\| = M\}} f(\alpha) \geq \frac{M}{2}.$$ 

Since $M$ is arbitrary, it follows that $f^*$ is superlinear. On the other hand, if $f^*$ is superlinear, then for any $\alpha \in \mathbb{R}^d$ there exists a compact set $K$ such that

$$f(\alpha) = \sup_{\beta \in \mathbb{R}^d} \{\langle \alpha, \beta \rangle - f^*(\beta)\} = \sup_{\beta \in K} \{\langle \alpha, \beta \rangle - f^*(\beta)\}.$$ 

Since $f^*$ is lower semicontinuous and thus bounded below on $K$, $f(\alpha) < \infty$. ■

D.3 A Theorem on the Legendre–Fenchel Transform of Compositions of Convex Functions

Theorem D.3.5 states a general representation for the Legendre–Fenchel transform of a composition of convex functions. This result, which is of independent interest, allows us to derive a number of basic formulas in convex analysis. In particular, we recover known formulas for the Legendre–Fenchel transform of the sum and of the maximum of convex functions [Corollaries D.4.2 and D.4.3]. We also obtain a new formula for the Legendre–Fenchel transform of $\log(\sum_{j=1}^N \rho_j \exp[h_j])$, where $N \geq 2$ is an integer, $\rho_1, \rho_2, \ldots, \rho_N$ are positive numbers summing to 1, and $h_1, h_2, \ldots, h_N$ are finite convex functions on $\mathbb{R}^d$ [Corollary D.4.1]. Formula (C.18) is a special case of this corollary. More generally, let $N \geq 1$ be an integer and $F$ a finite convex function on $\mathbb{R}^N$. Under suitable conditions, Theorem D.3.5 expresses the Legendre–Fenchel transform of $F(h_1, h_2, \ldots, h_N)$ in terms of the individual Legendre–Fenchel transforms $F^*$ and $h_1^*, h_2^*, \ldots, h_N^*$.

We next state the first set of conditions under which Theorem D.3.5 will be proved.

Condition D.3.1.

(a) $F$ is a finite convex function on $\mathbb{R}^N$ with Legendre–Fenchel transform $F^*$. For each $j \in \{1, 2, \ldots, N\}$ and for fixed values of all the coordinates $t_i, i \neq j$, $F(t_1, \ldots, t_j, \ldots, t_N)$ is a nondecreasing function of $t_j \in \mathbb{R}$ and

$$\lim_{t_j \to \infty} F(t_1, \ldots, t_j, \ldots, t_N) = \infty.$$ 

(b) $h_1, h_2, \ldots, h_N$ are convex, lower semicontinuous functions on $\mathbb{R}^d$ and $\cap_{j=1}^N \text{dom } h_j \neq \emptyset$. 
For \( j \in \{1, 2, \ldots, N\} \) let \( y_j \) be a point in \( \mathbb{R} \cup \{\infty\} \). If \( y_j = \infty \) for some \( j \in \{1, 2, \ldots, N\} \), then we define \( F(y_1, y_2, \ldots, y_N) = \infty \). With this definition, part (a) of Condition D.3.1 guarantees that \( F \) is a convex, lower semicontinuous, nondecreasing function on \((\mathbb{R} \cup \{\infty\})^N \). We claim that under Condition D.3.1 the composite function mapping
\[
\alpha \in \mathbb{R}^d \rightarrow F(h_1(\alpha), h_2(\alpha), \ldots, h_N(\alpha))
\]
is a convex, lower semicontinuous function on \( \mathbb{R}^d \). Indeed, this function is finite for \( \alpha \in \bigcap_{j=1}^N \text{dom } h_j \). The convexity of the composite function is immediate from the fact that each \( h_j \) is convex and that \( F \) is convex on \((\mathbb{R} \cup \{\infty\})^N \) and is nondecreasing in each coordinate. Let us prove that the composite function is lower semicontinuous for \( N = 2 \); general \( N \) is handled similarly. Suppose that \( \{\alpha_n, n \in \mathbb{N}\} \) is a sequence in \( \mathbb{R}^d \) converging to \( \alpha \in \mathbb{R}^d \). Since \( h_1 \) and \( h_2 \) are lower semicontinuous, there exist a subsequence of \( \{\alpha_n\} \) and points \( b_1 \) and \( b_2 \) in \( \mathbb{R} \cup \{\infty\} \) such that
\[
\lim_{n \to \infty} h_1(\alpha_n) = b_1 \geq h_1(\alpha) \quad \text{and} \quad \lim_{n \to \infty} h_2(\alpha_n) = b_2 \geq h_2(\alpha).
\]
Since \( F \) is lower semicontinuous and nondecreasing on \((\mathbb{R} \cup \{\infty\}) \times (\mathbb{R} \cup \{\infty\}) \)
\[
\liminf_{n \to \infty} F(h_1(\alpha_n), h_2(\alpha_n)) \geq F(b_1, b_2) \geq F(h_1(\alpha), h_2(\alpha)).
\]
This proves the lower semicontinuity of the composite function for \( N = 2 \).

We will prove in theorem D.3.5 that for \( \beta \in \mathbb{R}^d \) the Legendre–Fenchel transform of \( F(h_1, h_2, \ldots, h_N) \) equals
\[
M(\beta) = \inf\left\{ F^*(\lambda) + \sum_{j=1}^N \lambda_j h_j^*(\beta_j) : \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in \mathbb{R}^N, \beta_j \in \mathbb{R}^d \text{ for } j = 1, 2, \ldots, N, \sum_{j=1}^N \lambda_j \beta_j = \beta \right\}.
\]
This will be carried out under Condition D.3.1 and under either of the following two conditions. The proof under Condition D.3.3 will involve a mollification that allows a reduction to the proof under Condition D.3.2.

**Condition D.3.2.** \( M(\beta) \) is finite for all \( \beta \in \mathbb{R}^d \).

**Condition D.3.3.** For each \( j \in \{1, 2, \ldots, N\} \) \( h_j \) is finite on \( \mathbb{R}^d \).

In preparation for the statement of Theorem D.3.5, we prove the following facts about \( F^* \) and \( M(\beta) \).

**Lemma D.3.4.** (a) Under part (a) of Condition D.3.1, the effective domain of \( F^* \) is a subset of the nonnegative orthant
\[
(\mathbb{R}^N)_+ = \{(\lambda_1, \lambda_2, \ldots, \lambda_N) \in \mathbb{R}^N : \text{each } \lambda_j \geq 0\}.
\]
(b) Under Conditions D.3.1 and D.3.2, \( M(\beta) \) is a convex function of \( \beta \in \mathbb{R}^d \).
D.3. LEGENDRE–FENCHEL TRANSFORM OF COMPOSITIONS

Proof. (a) Suppose that $\text{dom} F^*$ has nonempty intersection with the complement of $(\mathbb{R}^N)_+$. Then there exists $\lambda \in \text{ri}(\text{dom} F^*)$ satisfying $\lambda_j < 0$ for some $j \in \{1, 2, \ldots, N\}$. The set $\partial F^*(\lambda)$ is nonempty [Theorem D.2.5 (a)] and so contains a point $\xi$; thus $\lambda \in \partial F(\xi)$ [Theorem D.2.6 (d)]. Let $u^j$ denote the unit vector in the $j$'th coordinate direction. Then for any $r > 0$

$$F(\xi - ru^j) \geq F(\xi) + r|\lambda_j| > F(\xi).$$

This violates the assumption that $F$ is nondecreasing in the $j$'th coordinate direction. Part (a) is proved.

(b) Using part (a), we replace $\mathbb{R}^N$ in the definition of $M(\beta)$ by $(\mathbb{R}^N)_+$, giving

$$M(\beta) = \inf \left\{ F^*(\lambda) + \sum_{j=1}^{N} \lambda_j h_j^*(\beta_j) : \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in (\mathbb{R}^N)_+, \beta_j \in \mathbb{R}^d \text{ for } j = 1, 2, \ldots, N, \sum_{j=1}^{N} \lambda_j \beta_j = \beta \right\}. \quad (D.2)$$

We prove that for any $s \in (0, 1)$, $\varphi \in \mathbb{R}^d$, and $\psi \in \mathbb{R}^d$

$$M(s\varphi + (1 - s)\psi) \leq s M(\varphi) + (1 - s) M(\psi).$$

Given $\varepsilon > 0$, there exist $\lambda \in (\mathbb{R}^N)_+$, $\beta_j \in \mathbb{R}^d$ for $j = 1, 2, \ldots, N$, $\mu \in (\mathbb{R}^N)_+$, and $\gamma_j \in \mathbb{R}^d$ for $j = 1, 2, \ldots, N$ such that

$$\sum_{j=1}^{N} \lambda_j \beta_j = \varphi, \quad F^*(\lambda) + \sum_{j=1}^{N} \lambda_j h_j^*(\beta_j) \leq M(\varphi) + \varepsilon$$

and

$$\sum_{j=1}^{N} \mu_j \gamma_j = \psi, \quad F^*(\mu) + \sum_{j=1}^{N} \mu_j h_j^*(\gamma_j) \leq M(\psi) + \varepsilon.$$

Let $\Upsilon$ denote the set of $j \in \{1, 2, \ldots, N\}$ for which either $\lambda_j > 0$ or $\mu_j > 0$. We define

$$\sigma_j = \begin{cases} \frac{s\lambda_j \beta_j}{s\lambda_j + (1 - s)\mu_j} + \frac{(1 - s)\mu_j \gamma_j}{s\lambda_j + (1 - s)\mu_j} & \text{if } j \in \Upsilon \\ \xi_j & \text{if } j \in \{1, 2, \ldots, N\} \setminus \Upsilon \end{cases}$$

where in the last line $\xi_j$ is any vector in $\text{dom} h_j^*$.

Using

$$\sum_{j=1}^{N} (s\lambda_j + (1 - s)\mu_j) \sigma_j = s\varphi + (1 - s)\psi$$

and the convexity of $F^*$ and $h_1^*, h_2^*, \ldots, h_N^*$, we calculate

$$s M(\varphi) + (1 - s) M(\psi) + \varepsilon$$

$$\geq s F^*(\lambda) + (1 - s) F^*(\mu) + \sum_{j=1}^{N} \left[ s \lambda_j h_j^*(\beta_j) + (1 - s) \mu_j h_j^*(\gamma_j) \right]$$

and

$$\geq s F^*(\lambda) + (1 - s) F^*(\mu) + \varepsilon.$$
\[ s F^*(\lambda) + (1 - s) F^*(\mu) \]
\[ + \sum_{j \in T} (s\lambda_j + (1 - s)\mu_j) \left( \frac{s\lambda_j}{s\lambda_j + (1 - s)\mu_j} h_j^*(\beta_j) + \frac{(1 - s)\mu_j}{s\lambda_j + (1 - s)\mu_j} h_j^*(\gamma_j) \right) \]
\[ \geq F^*(s\lambda + (1 - s)\mu) + \sum_{j=1}^N (s\lambda_j + (1 - s)\mu_j) h^*(\sigma_j) \]
\[ \geq M(s\varphi + (1 - s)\psi). \]

The convexity of \( M(\beta) \) now follows by sending \( \varepsilon \to 0 \). This completes the proof of the lemma.

We now state the main result.

**Theorem D.3.5.** We assume Condition D.3.1 and either Condition D.3.2 or D.3.3. Then for each \( \beta \in \mathbb{R}^d \) the Legendre-Fenchel transform \([F(h_1, h_2, \ldots, h_N)]^*(\beta)\) equals the function \( M(\beta) \) given in (D.2); i.e.,

\[
[F(h_1, h_2, \ldots, h_N)]^*(\beta) = \inf \left\{ F^*(\lambda) + \sum_{j=1}^N \lambda_j h_j^*(\beta_j) : \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in (\mathbb{R}^N)_+, \right. \\
\left. \beta_j \in \mathbb{R}^d \text{ for } j = 1, 2, \ldots, N, \sum_{j=1}^N \lambda_j \beta_j = \beta \right\}.
\]

**Proof.** We first give the proof assuming Conditions D.3.1 and D.3.2. Under these conditions \( M(\beta) \) is a finite convex function of \( \beta \in \mathbb{R}^d \) [Lemma D.3.4 (b)] and so is lower semicontinuous [Theorem D.2.2 (b)]. In addition, \( F(h_1(\alpha), h_2(\alpha), \ldots, h_N(\alpha)) \) is a convex, lower semicontinuous function of \( \alpha \in \mathbb{R}^d \). Hence by Legendre-Fenchel duality [Theorem D.2.6] it suffices to prove that for each \( \alpha \in \mathbb{R}^d \)

\[
\sup_{\beta \in \mathbb{R}^d} \{ \langle \alpha, \beta \rangle - M(\beta) \} = F(h_1(\alpha), h_2(\alpha), \ldots, h_N(\alpha)).
\]

Substituting equation (D.2) yields

\[
\sup_{\beta \in \mathbb{R}^d} \{ \langle \alpha, \beta \rangle - M(\beta) \}
\]
\[ = \sup \left\{ \langle \alpha, \beta \rangle - F^*(\lambda) - \sum_{j=1}^N \lambda_j h_j^*(\beta_j) : \beta \in \mathbb{R}^d, \right. \\
\left. \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in (\mathbb{R}^N)_+, \beta_j \in \mathbb{R}^d \text{ for } j = 1, 2, \ldots, N, \sum_{j=1}^N \lambda_j \beta_j = \beta \right\}
\]
\[ = \sup \left\{ \sum_{j=1}^N \lambda_j \left( \sup_{\beta_j \in \mathbb{R}^d} \{ \langle \alpha, \beta_j \rangle - h_j^*(\beta_j) \} \right) - F^*(\lambda) : \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in \mathbb{R}^N \right\}.
\]
\[= \sup \left\{ \sum_{j=1}^{N} \lambda_j h_j(\alpha) - F^*(\lambda) : \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in \mathbb{R}^N \right\} \]
\[= F(h_1(\alpha), h_2(\alpha), \ldots, h_N(\alpha)).\]

This completes the proof of the theorem under Conditions D.3.1 and D.3.2.

We now prove the theorem under Conditions D.3.1 and D.3.3. If for all \( t \in \mathbb{R}^N \) \( F(t) \) equals a constant \( c \), then for all \( \beta \in \mathbb{R}^d \)
\[ [F(h_1, h_2, \ldots, h_N)]^*(\beta) = M(\beta); \]
in fact, both sides of this equation equal \(-c\) if \( \beta = 0 \) and equal \( \infty \) for all nonzero \( \beta \in \mathbb{R}^d \). Under the assumption that \( F \) is not a constant function, we will prove the last display using a mollification argument that allows us to reduce to the proof given in the previous paragraph.

For \( \delta \in (0, 1] \) and \( \alpha \in \mathbb{R}^d \) we define
\[ h_{j, \delta}(\alpha) = h_j(\alpha) + \delta \| \alpha \|^2 \]
and
\[ M_\delta(\beta) = \inf \left\{ F^*(\lambda) + \sum_{j=1}^{N} \lambda_j h^*_{j, \delta}(\beta_j) : \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in (\mathbb{R}^N)_+, \right. \]
\[ \beta_j \in \mathbb{R}^d \text{ for } j = 1, 2, \ldots, N, \sum_{j=1}^{N} \lambda_j \beta_j = \beta \} \].

We will prove that \( M_\delta(\beta) \) is finite for all \( \beta \in \mathbb{R}^d \), that
\[ \lim_{\delta \to 0} [F(h_1, h_2, \ldots, h_N, \delta)]^*(\beta) = [F(h_1, h_2, \ldots, h_N)]^*(\beta), \]
and that
\[ \lim_{\delta \to 0} M_\delta(\beta) = M(\beta). \]

Given these three facts, the desired representation for \([F(h_1, h_2, \ldots, h_N)]^*\) is readily obtained. Indeed, since for all \( \delta \in (0, 1] \) \( F, h_{1, \delta}, h_{2, \delta}, \ldots, h_{N, \delta}, \text{ and } M_\delta \) satisfy Conditions D.3.1 and D.3.2, by the first part of the proof
\[ [F(h_{1, \delta}, h_{2, \delta}, \ldots, h_{N, \delta})]^*(\beta) = M_\delta(\beta). \]
Sending \( \delta \to 0 \) yields
\[ [F(h_1, h_2, \ldots, h_N)]^*(\beta) = M(\beta), \]
which is what we want to show.

We first prove that \( M_\delta(\beta) \) is finite for all \( \beta \in \mathbb{R}^d \). By assumption \( F \) is not a constant function, and so there exists a nonzero \( \lambda \in \mathbb{R}^N \) such that \( F^*(\lambda) < \infty \). Since a convex function has at least one point in its effective domain for which the subdifferential is
nonempty, each \( h_{j,\delta} \) is superlinear. Hence \( h_{j,\delta}^* \) is finite on \( \mathbb{R}^d \) [Theorem D.2.9 (c)], and so for any \( \beta_1, \beta_2, \ldots, \beta_N \) in \( \mathbb{R}^d \) satisfying \( \sum_{j=1}^N \lambda_j \beta_j = \beta \)

\[
M_\delta(\beta) \leq F^*(\lambda) + \sum_{j=1}^N \lambda_j h_{j,\delta}^*(\beta_j) < \infty.
\]

We now show that \( M_\delta(\beta) > -\infty \). The finiteness of each \( h_j \) on \( \mathbb{R}^d \) [Condition D.3.3] and part (a) of Theorem D.2.9 imply that

\[
h_{j,\delta}^*(\beta) \geq -h_{j,\delta}(0) = -h_j(0).
\]

Thus

\[
M_\delta(\beta) \geq \inf \left\{ F^*(\lambda) - \sum_{j=1}^N \lambda_j \left( V^N_{i=1} h_i(0) \right) : \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in (\mathbb{R}^N)_+ \right\}.
\]

Since \( F \) is finite on \( \mathbb{R}^N \), \( F^* \) is superlinear [Theorem D.2.9 (c)], and therefore \( M_\delta(\beta) > -\infty \).

The next step is to prove that for all \( \beta \in \mathbb{R}^d \)

\[
\lim_{\delta \to 0} [F(h_{1,\delta}, h_{2,\delta}, \ldots, h_{N,\delta})]^*(\beta) = [F(h_1, h_2, \ldots, h_N)]^*(\beta). \tag{D.4}
\]

By part (a) of Condition D.3.1 \( F \) is a finite convex function on \( \mathbb{R}^N \), and so \( F \) is continuous [Theorem D.2.2 (a)]. For each \( \alpha \in \mathbb{R}^d \) \( h_{j,\delta}(\alpha) \downarrow h_j(\alpha) \) as \( \delta \to 0 \), and since \( F \) is also nondecreasing

\[
F(h_{1,\delta}(\alpha), h_{2,\delta}(\alpha), \ldots, h_{N,\delta}(\alpha)) \downarrow F(h_1(\alpha), h_2(\alpha), \ldots, h_N(\alpha)) \text{ as } \delta \to 0.
\]

Thus

\[
\limsup_{\delta \to 0} [F(h_{1,\delta}, h_{2,\delta}, \ldots, h_{N,\delta})]^*(\beta) \leq [F(h_1, h_2, \ldots, h_N)]^*(\beta).
\]

On the other hand, for each \( \alpha \in \mathbb{R}^d \)

\[
\liminf_{\delta \to 0} [F(h_{1,\delta}, h_{2,\delta}, \ldots, h_{N,\delta})]^*(\beta) \geq \liminf_{\delta \to 0} [\langle \alpha, \beta \rangle - F(h_{1,\delta}(\alpha), h_{2,\delta}(\alpha), \ldots, h_{N,\delta}(\alpha))] = \langle \alpha, \beta \rangle - F(h_1(\alpha), h_2(\alpha), \ldots, h_N(\alpha)).
\]

It follows that

\[
\liminf_{\delta \to 0} [F(h_{1,\delta}, h_{2,\delta}, \ldots, h_{N,\delta})]^*(\beta) \geq \sup_{\alpha \in \mathbb{R}^d} \{ \langle \alpha, \beta \rangle - F(h_1(\alpha), h_2(\alpha), \ldots, h_N(\alpha)) \}
\]

\[
= [F(h_1, h_2, \ldots, h_N)]^*(\beta).
\]

This proves (D.4).
The final step is to prove that for all $\beta \in \mathbb{R}^d$

$$\lim_{\delta \to 0} M_\delta(\beta) = M(\beta).$$

Since for each $j \in \{1, 2, \ldots, N\}$, $\delta \in (0, 1]$, and $\alpha \in \mathbb{R}^d$ $h_{j,\delta}(\alpha) \geq h_j(\alpha)$, we have $h^*_j(\beta) \leq h^*_j(\delta)$ and thus $M_\delta(\beta) \leq M(\beta)$. It follows that

$$\limsup_{\delta \to 0} M_\delta(\beta) \leq M(\beta).$$

Fix any $\beta \in \mathbb{R}^d$ and let any sequence $\{\delta_k, k \in \mathbb{N}\}$ in $(0, 1]$ be given for which $\delta_k \to 0$ as $k \to \infty$. We complete the proof of Theorem D.3.5 by showing that there exists a subsequence of $\{\delta_k\}$ such that

$$\liminf_{k \to \infty} M_{\delta_k}(\beta) \geq M(\beta). \quad (D.5)$$

Without loss of generality we can assume that for the original sequence $\{\delta_k, k \in \mathbb{N}\}$ $B \doteq \liminf_{k \to \infty} M_{\delta_k}(\beta)$ is finite since otherwise there is nothing to prove. Let $\{\delta_k\}$ be a subsequence for which $M_{\delta_k}(\beta) \to B$. For each $k$ there exist $\lambda_{\delta_k} = (\lambda_{1,\delta_k}, \lambda_{2,\delta_k}, \ldots, \lambda_{N,\delta_k}) \in (\mathbb{R}^N)_+$ and points $\beta_{j,\delta_k} \in \mathbb{R}^d$ such that

$$\sum_{j=1}^N \lambda_{j,\delta_k} \beta_{j,\delta_k} = \beta \quad (D.6)$$

and

$$M_{\delta_k}(\beta) + \frac{1}{k} \geq F^*(\lambda_{\delta_k}) + \sum_{j=1}^N \lambda_{j,\delta_k} h^*_j(\beta_{j,\delta_k}). \quad (D.7)$$

As noted earlier in this proof, Theorem D.2.9 implies that

$$h^*_j(\beta_{j,\delta_k}) \geq -h_{j,\delta_k}(0) = -h_j(0) \quad (D.8)$$

and that $F^*$ is superlinear. Hence (D.7) yields

$$\sup_{k \in \mathbb{N}} \left\{ F^*(\lambda_{\delta_k}) - \sum_{j=1}^N \lambda_{j,\delta_k} \left( \sum_{i=1}^N h_i(0) \right) \right\} \leq \sup_{k \in \mathbb{N}} \left( M_{\delta_k}(\beta) + \frac{1}{k} \right) < \infty,$$

which in turn implies that the subsequence $\{\lambda_{\delta_k}\}$ lies in a compact subset of $(\mathbb{R}^N)_+$. There exists $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in (\mathbb{R}^N)_+$ such that along a subsequence $\lambda_{\delta_k} \to \lambda$.

Let $J$ denote the set of $j \in \{1, 2, \ldots, N\}$ for which $\lambda_j > 0$. We first prove (D.5) under the assumption that $J \neq \emptyset$. For each $j \in J$ and $\alpha \in \mathbb{R}^d$, since $\delta_k \in (0, 1]$, $h_{j,\delta_k}(\alpha) \leq h_{j,1}(\alpha)$ and so $h^*_j(\lambda_{\delta_k}) \leq h^*_j(\beta_{j,\delta_k})$. Together with $F^*(\lambda_{\delta_k}) = -F(0)$ and (D.8) for $i \neq j$, (D.7) implies the following. There exists $\Gamma < \infty$ such that for all sufficiently large $k$ in the subsequence along which $\lambda_{\delta_k} \to \lambda$

$$h^*_j(\beta_{j,\delta_k}) \leq h^*_j(\beta_{j,\delta_k}) \leq \Gamma.$$
The finiteness of $h_j$ on $\mathbb{R}^d$ and part (b) of Theorem D.2.9 imply that $h_{j,1}^*$ has compact level sets. The last display allows us to conclude that for all $j \in J$ there exists $\beta_j \in \mathbb{R}^d$ and a further subsequence such that $\beta_j \to \beta_j$ as $k \to \infty$ along this subsequence. In particular, in the case where $J = \{1, 2, \ldots, N\}$, (D.6) implies that $\sum_{j=1}^N \lambda_j \beta_j = \beta$.

Still under the assumption that $J \neq \emptyset$, assume also that $J$ is a proper subset of $\{1, 2, \ldots, N\}$. Setting $\beta_j = 0$ for $j \notin J$, we claim that the last equality in the previous paragraph is still valid; i.e.,

$$
\sum_{j=1}^N \lambda_j \beta_j = \sum_{j \in J} \lambda_j \beta_j = \beta.
$$

If this is not true, then $\sum_{j \in J^c} \lambda_j \beta_j$ equals $\gamma \neq \beta$ and by (D.6)

$$
\lim_{k \to \infty} \sum_{j \in J^c} \lambda_j \delta_j \beta_j \delta_k = \beta - \lim_{k \to \infty} \sum_{j \in J} \lambda_j \delta_k \beta_j \delta_k = \beta - \gamma \neq 0.
$$

In particular, for all sufficiently large $k$ we have $0 < \sum_{j \in J} \lambda_j \delta_k \to 0$. The function

$$
\varphi = (\bigvee_{j \in J} h_{j,1})^*
$$

satisfies $h_{j,1}^*(\beta) \geq \varphi(\beta)$ for each $j \in J^c$, $\delta_k \in [0, 1]$, and $\beta \in \mathbb{R}^d$, and since each $h_{j,1}$ is finite on $\mathbb{R}^d$, $\varphi$ is superlinear [Theorem D.2.9 (c)]. By (D.7), $F^*(\lambda \delta_k) \geq -F(0)$, (D.8) for $j \in J$, and the convexity of $\varphi$, there exists $\Gamma < \infty$ such that for all sufficiently large $k$

$$
\Gamma \geq \sum_{j \in J} \lambda_j \delta_k h_{j,1}^*(\beta_j \delta_k)
$$

$$
\geq \sum_{j \in J^c} \lambda_j \delta_k \varphi(\beta_j \delta_k) \geq \left( \sum_{j \in J} \lambda_j \delta_k \right) \varphi \left( \frac{\sum_{j \in J^c} \lambda_j \delta_k \beta_j \delta_k}{\sum_{j \in J} \lambda_j \delta_k} \right).
$$

The superlinearity of $\varphi$ implies that the last term in the display tends to $\infty$ as $k \to \infty$. This contradiction shows that $\sum_{j=1}^N \lambda_j \beta_j = \beta$. Hence this equality is valid whenever $J \neq \emptyset$.

The proof of (D.5) is now straightforward. Since for each $j \in J$ and $\alpha \in \mathbb{R}^d$

$$
\liminf_{k \to \infty} h_{j,1}^*(\beta_j \delta_k) \geq \liminf_{k \to \infty} \left( \langle \alpha, \beta_j \delta_k \rangle - h_j(\alpha) - \delta_k \|\alpha\|^2 \right)
$$

$$
= \langle \alpha, \beta_j \rangle - h_j(\alpha),
$$

we have

$$
\liminf_{k \to \infty} h_{j,1}^*(\beta_j \delta_k) \geq h_j^*(\beta_j).
$$

Equation (D.7), (D.8) for indices not in $J$, the lower semicontinuity of $F^*$, and $\sum_{j=1}^N \lambda_j \beta_j = \beta$ imply that

$$
\liminf_{k \to \infty} M_{\delta_k}(\beta) \geq \liminf_{k \to \infty} \left( F^*(\lambda \delta_k) + \sum_{j=1}^N \lambda_j h_{j,1}^*(\beta_j \delta_k) \right)
$$

$$
\geq F^*(\lambda) + \sum_{j=1}^N \lambda_j h_j^*(\beta_j)
$$

$$
\geq M(\beta).
$$
This is what we want to prove.

We now prove (D.5) under the assumption that $J = \emptyset$. We claim that in this case $\beta$ must equal 0. Indeed, by (D.6) $\sum_{j \in J} \lambda_{j,\delta_k} \beta_{j,\delta_k} = \beta$, and if $\beta \neq 0$, then as in (D.9) there exists $\Gamma < \infty$ such that for all sufficiently large $k$

$$
\Gamma \geq \left( \sum_{j \in J} \lambda_{j,\delta_k} \right) \varphi \left( \frac{\beta}{\sum_{j \in J} \lambda_{j,\delta_k}} \right).
$$

The superlinearity of $\varphi$ implies that the right hand side of this display tends to $\infty$ as $k \to \infty$. This contradiction shows that $\beta = 0$. The rest is easy. Since $J = \emptyset$, $\lambda_{\delta_k} \to \lambda = 0$, and since $F^*$ is lower semicontinuous, (D.7) and (D.8) yield

$$
\liminf_{k \to \infty} M_{\delta_k}(0) \geq F^*(0) \geq M(0).
$$

This completes the proof of Theorem D.3.5. □

### D.4 Three Examples

We present three interesting examples of Legendre–Fenchel transforms obtained from Theorem D.3.5. In each corollary it is assumed that $h_1, h_2, \ldots, h_N$ are finite convex functions on $\mathbb{R}^d$ and thus satisfy part (b) of Condition D.3.1 and Condition D.3.3. Let $\Delta_N$ denote the set of probability vectors in $\mathbb{R}^N$; i.e., the set of vectors $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in \mathbb{R}^N$ with nonnegative components summing to 1. If for fixed $x \in \mathbb{R}^d$ we identify $h_j(\alpha) = H_j(x, \alpha)$, then the first corollary yields formula (C.18).

**Corollary D.4.1.** Let $h_1, h_2, \ldots, h_N$ be finite convex functions on $\mathbb{R}^d$. Given positive numbers $\rho_1, \rho_2, \ldots, \rho_N$ summing to 1, we define for $(t_1, t_2, \ldots, t_N) \in \mathbb{R}^N$

$$
F(t_1, t_2, \ldots, t_N) \doteq \log \left( \sum_{j=1}^N \rho_j \exp[t_j] \right)
$$

and for $\lambda \in \Delta_N$

$$
R(\lambda \| \rho) \doteq \sum_{j=1}^N \lambda_j \log \left( \frac{\lambda_j}{\rho_j} \right).
$$

Then for each $\beta \in \mathbb{R}^d$ the Legendre–Fenchel transform of

$$
F(h_1(\alpha), h_2(\alpha), \ldots, h_N(\alpha)) \doteq \log \left( \sum_{j=1}^N \rho_j \exp[h_j(\alpha)] \right), \alpha \in \mathbb{R}^d,
$$

is given by

$$
\left[ \log \left( \sum_{j=1}^N \rho_j \exp[h_j] \right) \right]^* (\beta)
$$
\[ = \inf \left\{ R(\lambda \| \rho) + \sum_{j=1}^{N} \lambda_j h_j^*(\beta_j) : \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in \Delta_N, \right\} \]

\[ \beta_j \in \mathbb{R}^d \text{ for } j = 1, 2, \ldots, N, \sum_{j=1}^{N} \lambda_j \beta_j = \beta \right\}. \]

**Proof.** For \((t_1, t_2, \ldots, t_N) \in \mathbb{R}^N\)

\[ F(t_1, t_2, \ldots, t_N) = \log \left( \sum_{j=1}^{N} \rho_j \exp[t_j] \right) \]

satisfies part (a) of Condition D.3.1. \(F^*(\lambda)\) is given by

\[ F^*(\lambda) = \begin{cases} R(\lambda \| \rho) & \text{if } \lambda \in \Delta_N \\ \infty & \text{if } \lambda \in \mathbb{R}^N \setminus \Delta_N. \end{cases} \]

The first line is a consequence of the Donsker–Varadhan variational formula for the relative entropy [Lemma 1.4.3 (a)]. In order to prove the second line, let \(\{u^i, i = 1, 2, \ldots, N\}\) be the unit coordinate vectors in \(\mathbb{R}^N\). Since \(F^*(\lambda)\) is the Legendre–Fenchel transform of the cumulant generating function of \(\mu = \sum_{j=1}^{N} \rho_j \delta_{u^i}\), part (d) of Lemma 6.2.3 implies that \(F^*(\lambda) = \infty\) if \(\lambda \not\in \text{cl}(\text{conv } S_\mu) = \Delta_N\). Substituting the formula for \(F^*(\lambda)\) into equation (D.3) completes the proof of the corollary. 

The next corollary gives the formula for the Legendre–Fenchel transform of the sum of finite convex functions. This formula appears in [79, Thm. 16.4] under a weaker hypothesis. This corollary yields the equality in part (a) of Lemma 6.6.2 and equation (C.19) in Section C.2.

**Corollary D.4.2.** Let \(h_1, h_2, \ldots, h_N\) be finite convex functions on \(\mathbb{R}^d\). For \((t_1, t_2, \ldots, t_N) \in \mathbb{R}^N\) we define

\[ F(t_1, t_2, \ldots, t_N) = \sum_{j=1}^{N} t_j. \]

Then for each \(\beta \in \mathbb{R}^d\) the Legendre–Fenchel transform of

\[ F(h_1(\alpha), h_2(\alpha), \ldots, h_N(\alpha)) = \sum_{j=1}^{N} h_j(\alpha), \quad \alpha \in \mathbb{R}^d, \]

is given by

\[ \left( \sum_{j=1}^{N} h_j \right)^* (\beta) = \inf \left\{ \sum_{j=1}^{N} h_j^*(\beta_j) : \beta_j \in \mathbb{R}^d \text{ for } j = 1, 2, \ldots, N, \sum_{j=1}^{N} \beta_j = \beta \right\}. \]
D.4. THREE EXAMPLES

Proof. For \((t_1, t_2, \ldots, t_N) \in \mathbb{R}^N\)

\[
F(t_1, t_2, \ldots, t_N) = \sum_{j=1}^{N} t_j
\]
satisfies part (a) of Condition D.3.1. Substituting

\[
F^*(\lambda) = \begin{cases} 
0 & \text{if each } \lambda_j = 1 \\
\infty & \text{if } \lambda \in \mathbb{R}^N \setminus \{1, 1, \ldots, 1\}
\end{cases}
\]

into equation (D.3) completes the proof of the corollary. ■

The last corollary gives the formula for the Legendre–Fenchel transform of the maximum of finite convex functions. This formula appears in [79, Theorem 16.5] under a weaker hypothesis.

Corollary D.4.3. Let \(h_1, h_2, \ldots, h_N\) be finite convex functions on \(\mathbb{R}^d\). For \((t_1, t_2, \ldots, t_N) \in \mathbb{R}^N\) we define

\[
F(t_1, t_2, \ldots, t_N) = \bigvee_{j=1}^{N} t_j.
\]

Then for each \(\beta \in \mathbb{R}^d\) the Legendre–Fenchel transform of

\[
F(h_1(\alpha), h_2(\alpha), \ldots, h_N(\alpha)) = \bigvee_{j=1}^{N} h_j(\alpha), \ \alpha \in \mathbb{R}^d,
\]
is given by

\[
\left(\bigvee_{j=1}^{N} h_j\right)^* (\beta) = \inf \left\{ \sum_{j=1}^{N} \lambda_j h_j^*(\beta_j) : \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in \Delta_N, \right. \\
\left. \beta_j \in \mathbb{R}^d \text{ for } j = 1, 2, \ldots, N, \sum_{j=1}^{N} \lambda_j \beta_j = \beta \right\}.
\]

Proof. For \((t_1, t_2, \ldots, t_N) \in \mathbb{R}^N\)

\[
F(t_1, t_2, \ldots, t_N) = \bigvee_{j=1}^{N} t_j
\]
satisfies part (a) of Condition D.3.1. Substituting

\[
F^*(\lambda) = \begin{cases} 
0 & \text{if } \lambda \in \Delta_N \\
\infty & \text{if } \lambda \in \mathbb{R}^N \setminus \Delta_N
\end{cases}
\]

into equation (D.3) completes the proof of the corollary. ■
Appendix E

Proof of Theorem 5.3.5 When Condition 5.4.1 Replaces Condition 5.3.1

E.1 Introduction

In Theorem E.2.3 we prove that when Condition 5.3.1 is replaced by Condition 5.4.1, all the conclusions of Theorem 5.3.5 remain true. As we point out in part (b) of Remark 6.3.6, under this replacement of conditions we are able to derive the Laplace principle in Theorem 6.3.3. The motivation for this extension is our desire to treat a wider class of random walks, including models analogous to diffusions with linear drift.

E.2 Proofs of Results

For easy reference we restate Condition 5.4.1. We recall the definition

\[ H_b(x, \alpha) = \log \int_{\mathbb{R}^d} \exp(\alpha, y - b(x)) \mu(dy|x). \]

Condition 5.4.1.

(a) There exists a locally Lipschitz continuous function \( b : \mathbb{R}^d \rightarrow \mathbb{R}^d \) such that for each \( \alpha \in \mathbb{R}^d \)

\[ \sup_{x \in \mathbb{R}^d} H_b(x, \alpha) < \infty. \]

(b) For any point \( x_0 \in \mathbb{R}^d \) and any continuous function \( y \) mapping \([0, 1]\) into \( \mathbb{R}^d \), the equation

\[ x(t) = x_0 + \int_0^t b(x(s)) \, ds + y(t) \]

has at least one solution \( \{x(t), t \in [0, 1]\} \).
The first result in which Condition 5.3.1 is assumed is Proposition 5.3.2. When this condition is replaced by Condition 5.4.1, the boundedness of the running costs assumed in equation (5.8) of the proposition no longer gives the uniform integrability of the sequence of control measures \( \{ \nu^n, n \in \mathbb{N} \} \) as expressed in equation (5.9). However, the boundedness of the running costs does yield an alternative that is just as useful. We state the replacement of Proposition 5.3.2 after introducing some notation.

For \( x \in \mathbb{R}^d \) we define a function \( \Theta_x \) mapping \( \mathcal{P}(\mathbb{R}^d) \) into \( \mathcal{P}(\mathbb{R}^d) \) by

\[
\Theta_x \gamma(dy) \doteq \gamma(dy + b(x)).
\]

\( \Theta_x \) shifts \( \gamma \) by the amount \( -b(x) \); e.g., if \( \gamma \) has mean \( v \), then \( \Theta_x \gamma \) has mean \( v - b(x) \). Let \( \mu(dy|x) \) be the stochastic kernel on \( \mathbb{R}^d \) given \( \mathbb{R}^d \) in terms of which the random walk model of Chapter 5 is defined. Defining the shifted stochastic kernel

\[
\tilde{\mu}(dy|x) \doteq \Theta_x \mu(dy|x) = \mu(dy + b(x)|x),
\]

we see that for each \( x \) and \( \alpha \) in \( \mathbb{R}^d \) \( H_b(x, \alpha) \) equals the cumulant generating function of \( \tilde{\mu}(\cdot|x) \); i.e.,

\[
H_b(x, \alpha) = \log \int_{\mathbb{R}^d} \exp(\alpha, y) \tilde{\mu}(dy|x). \tag{E.1}
\]

Fix \( n \in \mathbb{N} \). Recall that for the random walk model the control applied at random time \( j \in \{0, 1, \ldots, n - 1\} \) is a stochastic kernel \( \nu_j^n = \nu_j^n(dy|x_0, x_1, \ldots, x_j) \) on \( \mathbb{R}^d \) given \( \mathbb{R}^d \) \( (\mathbb{R}^d)^{j+1} \). For \( x_0, x_1, \ldots, x_j \) in \( \mathbb{R}^d \), we define the shifted control

\[
\tilde{\nu}_j^n(dy) \doteq \Theta_x \nu_j^n(dy|x_0, x_1, \ldots, x_j) = \nu_j^n(dy + b(x_j)|x_0, x_1, \ldots, x_j). \tag{E.2}
\]

The sequence \( \{\tilde{\nu}_j^n, j = 0, 1, \ldots, n - 1\} \) is called an admissible shifted control sequence. By analogy with equations (5.3) and (5.4), for \( n \in \mathbb{N} \) and \( t \in [0, 1] \) we define the stochastic kernel

\[
\tilde{\nu}^n(dy|t) \doteq \begin{cases} 
\tilde{\nu}_j^n(dy) & \text{if } t \in [j/n, (j + 1)/n), \ j = 0, 1, \ldots, n - 2 \\
\tilde{\nu}_{n-1}^n(dy) & \text{if } t \in [(n - 1)/n, 1],
\end{cases} \tag{E.3}
\]

where \( \tilde{\nu}_j^n(dy) = \tilde{\nu}_j^n(dy|\tilde{X}_0^n, \tilde{X}_1^n, \ldots, \tilde{X}_j^n) \). We also define a random probability measure on \( \mathbb{R}^d \times [0, 1] \) by

\[
\tilde{\nu}^n(A \times B) \doteq \int_B \tilde{\nu}^n(A|t) \, dt \tag{E.4}
\]

for Borel subsets \( A \) of \( \mathbb{R}^d \) and \( B \) of \( [0, 1] \). The term admissible shifted control measure will be used to refer to the random probability measure \( \tilde{\nu}^n \).

In order to prove the replacement of Proposition 5.3.2, the following invariance property of the relative entropy with respect to certain mappings is needed.

**Lemma E.2.1.** Let \( \mathcal{X} \) be a Polish space and \( \psi \) a one-to-one function mapping \( \mathcal{X} \) onto \( \mathcal{X} \) such that both \( \psi \) and the inverse mapping \( \psi^{-1} \) are measurable. Let \( \Delta_\psi \) denote the function mapping \( \mathcal{P}(\mathcal{X}) \) onto \( \mathcal{P}(\mathcal{X}) \) which is given by \( \Delta_\psi \alpha \doteq \alpha \circ \psi^{-1} \). Then for all probability measures \( \nu \) and \( \mu \) on \( \mathcal{X} \)

\[
R(\Delta_\psi \nu \| \Delta_\psi \mu) = R(\nu \| \mu).
\]
E.2. PROOFS OF RESULTS

Proof. According to part (g) of Lemma 1.4.3

\[
R(\Delta_\psi \nu \| \Delta_\psi \mu) = \sup_{\pi \in \Pi} \sum_{A \in \pi} \Delta_\psi \nu(A) \log \frac{\Delta_\psi \nu(A)}{\Delta_\psi \mu(A)}
\]

\[
= \sup_{\pi \in \Pi} \sum_{A \in \pi} \nu(\psi^{-1}(A)) \log \frac{\nu(\psi^{-1}(A))}{\mu(\psi^{-1}(A))},
\]

where \(\Pi\) denotes the class of all finite measurable partitions of \(\mathcal{X}\). For each \(\pi \in \Pi\) we define \(\psi^{-1}(\pi) = \{\psi^{-1}(A) : A \in \pi\}\). The properties of \(\psi\) guarantee that if \(\pi_1 \neq \pi_2\) are in \(\Pi\), then \(\psi^{-1}(\pi_1) \neq \psi^{-1}(\pi_2)\) and that each finite measurable partition of \(\mathcal{X}\) has the form \(\psi^{-1}(\pi)\) for some \(\pi \in \Pi\). As a consequence

\[
R(\Delta_\psi \nu \| \Delta_\psi \mu) = \sup_{\pi \in \Pi} \sum_{A \in \pi} \nu(\psi^{-1}(A)) \log \frac{\nu(\psi^{-1}(A))}{\mu(\psi^{-1}(A))}
\]

\[
= \sup_{\pi \in \Pi} \sum_{B \in \pi} \nu(B) \log \frac{\nu(B)}{\mu(B)}
\]

\[
= R(\nu \| \mu).
\]

This completes the proof. \(\blacksquare\)

We now state the replacement of Proposition 5.3.2.

**Proposition E.2.2.** In the random walk model of Chapter 5, we assume part (a) of Condition 5.4.1. For \(n \in \mathbb{N}\) and \(x \in \mathbb{R}^d\), consider any admissible control sequence \(\{\nu_j^n, j = 0, 1, \ldots, n-1\}\) such that

\[
\sup_{n \in \mathbb{N}} \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu_j^n(\cdot) \| \mu(\cdot | \tilde{X}_j^n)) \right\} \equiv \Delta < \infty,
\]  

(E.5)

where \(\nu_j^n(dy) = \nu_j^n(dy | \tilde{X}_0^n, \tilde{X}_1^n, \ldots, \tilde{X}_j^n)\). Let the sequence of admissible shifted control measures \(\{\tilde{\nu}_n, n \in \mathbb{N}\}\) on \(\mathbb{R}^d \times [0, 1]\) be defined by equations (E.2)-(E.4). Then \(\{\tilde{\nu}_n\}\) is tight and, in fact, has the uniform integrability property

\[
\lim_{C \to \infty} \sup_{n \in \mathbb{N}} \mathbb{E}_x \left\{ \int_{\{y \in \mathbb{R}^d : |y| > C\}} \|y\| \tilde{\nu}_n(dy \times dt) \right\} = 0.
\]

(E.6)

**Proof.** Using Lemma E.2.1, we rewrite (E.5) as

\[
\sup_{n \in \mathbb{N}} \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\tilde{\nu}_j^n(\cdot) \| \tilde{\mu}(\cdot | \tilde{X}_j^n)) \right\} \equiv \Delta < \infty.
\]

Because of part (a) of Condition 5.4.1, the same proof as in Proposition 5.3.2 yields the limit (E.6). This limit implies that

\[
\sup_{n \in \mathbb{N}} \mathbb{E}_x \left\{ \int_{\mathbb{R}^d \times [0, 1]} \|y\| \tilde{\nu}_n(dy \times ds) \right\}
\]

\[
= \sup_{n \in \mathbb{N}} \int_{\mathcal{P}(\mathbb{R}^d \times [0, 1])} \left( \int_{\mathbb{R}^d \times [0, 1]} \|y\| \gamma(dy \times dt) \right) \mathbb{P}_x \{\tilde{\nu}_n \in d\gamma\} < \infty,
\]

(E.7)
which, as in the proposition, implies that \( \tilde{\nu}^n \) is tight.

We now show that when Condition 5.4.1 replaces Condition 5.3.1 all the conclusions of Theorem 5.3.5 continue to hold.

**Theorem E.2.3.** In the random walk model of Chapter 5, we assume Condition 5.4.1. For each \( n \in \mathbb{N} \) and \( x \in \mathbb{R}^d \), consider any admissible control sequence \( \nu_j^n, j = 0, 1, \ldots, n-1 \) such that

\[
\sup_{n \in \mathbb{N}} \mathbb{E}_x \left\{ \frac{1}{n} \sum_{j=0}^{n-1} R(\nu_j^n(\cdot) \mu(\cdot, X_j^n)) \right\} < \infty. \tag{E.8}
\]

Let the sequences \( \{ \tilde{X}_n, n \in \mathbb{N} \} \) and \( \{ \tilde{X}_n, n \in \mathbb{N} \} \) be defined by formulas (5.6) and (5.5), respectively. Let the sequence of admissible control measures \( \{ \nu^n, n \in \mathbb{N} \} \) be defined by equations (5.3) and (5.4). Then all the conclusions of Theorem 5.3.5 continue to hold.

**Proof.** For each \( n \in \mathbb{N} \) we introduce a controlled process \( \{ \tilde{X}_j^n, j = 0, 1, \ldots, n \} \) taking values in \( \mathbb{R}^d \) which stands in the same relationship to \( \{ \tilde{X}_j^n, j = 0, 1, \ldots, n-1 \} \) that the controlled process \( \{ X_j^n, j = 0, 1, \ldots, n \} \) stands to \( \{ X_j^n, j = 0, 1, \ldots, n-1 \} \). We recall from Section 5.2 that for \( n \in \mathbb{N}, x \in \mathbb{R}^d, \) and \( j \in \{ 0, 1, \ldots, n-1 \} \) \( X_j^n = X_0^n = x \) and

\[
X_{j+1}^n = X_j^n + \frac{1}{n} \tilde{X}_j^n,
\]

where

\[
\tilde{P}_x \{ \tilde{X}_j^n \in dy \mid X_0^n = X_1^n = \ldots = X_{j-1}^n \} = \nu_j^n(dy \mid X_0^n = X_1^n = \ldots = X_{j-1}^n).
\]

We now define \( \tilde{T}_0^n = 0 \) and for \( j \in \{ 1, 2, \ldots, n \} \)

\[
\tilde{T}_j^n = X_j^n - \frac{1}{n} \sum_{i=0}^{j-1} b(X_i^n) - x \quad \text{and} \quad \tilde{\Psi}_j^n = \tilde{X}_j^n - b(X_j^n).
\]

The dynamics of the process \( \{ \tilde{T}_j^n, j = 0, 1, \ldots, n-1 \} \) are given by

\[
\tilde{T}_{j+1}^n = \tilde{T}_j^n + \frac{1}{n} \tilde{\Psi}_j^n,
\]

where

\[
P_x \{ \tilde{\Psi}_j^n \in dy \mid X_0^n = X_1^n = \ldots = X_j^n \} = P_x \{ \tilde{X}_j^n \in dy + b(X_j^n) \mid X_0^n = X_1^n = \ldots = X_j^n \}
\]

\[
= \nu_j^n(dy + b(X_j^n) \mid X_0^n = X_1^n = \ldots = X_j^n)
\]

\[
= \tilde{\nu}_j^n(dy \mid X_0^n = X_1^n = \ldots = X_j^n).
\]

In terms of the controlled process \( \{ \tilde{T}_j^n, j = 0, 1, \ldots, n \} \), we define the piecewise linear interpolation

\[
T^n(t) = \tilde{T}_j^n + \left( t - \frac{j}{n} \right) \tilde{\Psi}_j^n \quad \text{for} \ t \in [j/n, (j+1)/n], j = 0, 1, \ldots, n-1,
\]
and the piecewise constant interpolation

\[
\tilde{\Theta}^n(t) = \begin{cases} \check{\Theta}^n_j & \text{if } t \in \left[ j/n, (j + 1)/n \right), \ j = 0, 1, \ldots, n - 2 \\ \tilde{\Theta}^n_{n-1} & \text{if } t \in \left[ (n - 1)/n, 1 \right] \end{cases}
\]

The processes \( \check{\Theta}^n = \{ \check{\Theta}^n(t), t \in [0, 1] \} \) and \( \tilde{\Theta}^n = \{ \tilde{\Theta}^n(t), t \in [0, 1] \} \) are analogous to the processes \( \check{X}^n \) and \( \tilde{X}^n \) defined in equations (5.6) and (5.5), respectively.

Under part (a) of Condition 5.4.1, if (E.8) is valid, then Proposition E.2.2 guarantees that the sequence \( \{ \check{\nu}^n \} \) is tight and is uniformly integrable in the sense that (E.6) holds. Hence the same proof as in Theorem 5.3.5 allows us to derive all the conclusions of that theorem when the measures \( \{ \check{\nu}^n \} \) replace the measures \( \{ \nu^n \} \) and the processes \( \{ \check{\Theta}^n \} \) and \( \{ \tilde{\Theta}^n \} \) replace the processes \( \{ \check{X}^n \} \) and \( \{ \tilde{X}^n \} \). In particular, we conclude that every subsequence of

\[ \{ (\check{\nu}^n, \check{\Theta}^n, \tilde{\Theta}^n), n \in \mathbb{N} \} \]

has a subsubsequence that converges in distribution to a triple \( (\check{\nu}, \check{\Theta}, \tilde{\Theta}) \). There exists a probability space \( (\hat{\Omega}, \hat{\mathcal{F}}, \hat{P}_x) \) such that \( \check{\nu} \) can be realized as a stochastic kernel on \( \mathbb{R}^d \times [0, 1] \) given \( \hat{\Omega} \). As in Lemma 5.3.4, we have the \( \hat{P}_x \)-a.s. decomposition

\[ \check{\nu}(dy \times dt) = \check{\nu}(dy|t) \otimes dt, \quad (E.9) \]

where \( \check{\nu}(dy|t) \) is a stochastic kernel on \( \mathcal{P}(\mathbb{R}^d) \) given \([0, 1] \times \hat{\Omega} \). The limit \( \check{\Theta} = \{ \check{\Theta}(t), t \in [0, 1] \} \) maps \( \hat{\Omega} \) into \( \mathcal{C}([0, 1] : \mathbb{R}^d) \), and w.p.1 this process satisfies for every \( t \in [0, 1] \)

\[ \check{\Theta}(t) = \int_{\mathbb{R}^d \times [0,t]} y \check{\nu}(dy \times ds). \quad (E.10) \]

As \( n \to \infty \) along this subsequence, Proposition E.2.2 implies that the random variables

\[ \int_{\mathbb{R}^d \times [0,1]} \|y\| \check{\nu}^n(dy \times ds) \]

converge in distribution to an integrable random variable. By Prohorov’s Theorem and the Skorohod Representation Theorem, as \( n \to \infty \) along a further subsequence, we have w.p.1

\[ (\check{\nu}^n, \check{\Theta}^n, \tilde{\Theta}^n) \to (\check{\nu}, \check{\Theta}, \tilde{\Theta}) \text{ and } \limsup_{n \to \infty} \int_{\mathbb{R}^d \times [0,1]} \|y\| \check{\nu}^n(y \times ds) < \infty. \quad (E.11) \]

Consider any point \( \omega \in \hat{\Omega} \) for which (E.9), (E.10), and (E.11) are valid. By the definition of \( \check{\Theta}^n_j \) the relation

\[ \check{X}^n_j = x + \frac{1}{n} \sum_{i=0}^{j-1} b(\check{X}^n_i) + \check{\Theta}^n_j \]

holds for all \( j \in \{0, 1, \ldots, n\} \). This implies that for each \( n \) and all \( t \in [0,1] \)

\[ \check{X}^n(t) = x + \int_0^t b(\check{X}^n(s)) ds + \check{\Theta}^n(t) + c^n(t), \]
where for \( j \in \{0,1,\ldots,n-1\} \) and \( t \in [j/n, (j+1)/n) \)
\[
c^n(t) = -\int_{j/n}^{t} b(\bar{X}^n(s)) \, ds.
\]
We also know by Condition 5.4.1 that a solution to
\[
\bar{X}(t) = x + \int_0^t b(\bar{X}(s)) \, ds + \bar{Y}(t) \tag{E.12}
\]
exists, and since \( b \) is locally Lipschitz, the solution \( \bar{X} = \{\bar{X}(t), t \in [0,1]\} \) is unique. Since \( \bar{X} \) is continuous, there exists \( B \in (0,\infty) \) such that \( \|\bar{X}(t)\| \leq B/3 \) for all \( t \in [0,1] \). Following the standard convention, we let the minimum over the empty set be \( \infty \) and define
\[
\tau^n_B = \frac{1}{n} \min\{j \in \{0,1,\ldots,n\} : \|\bar{X}^n_j\| \geq B\}.
\]
If \( L \) denotes the Lipschitz constant of the restriction of \( b \) to \( \{x : \|x\| \leq B\} \), then Gronwall’s Inequality [Theorem A.6.4] implies that for all \( t \in [0,\tau^n_B) \cap [0,1] \)
\[
\|\bar{X}^n(t) - \bar{X}(t)\| \leq \left(\sup_{s \in [0,\tau^n_B]} \|\bar{Y}^n(s) - \bar{Y}(s)\| + \sup_{s \in [0,\tau^n_B]} \|c^n(s)\|\right) e^L.
\]
Since
\[
\lim_{n \to \infty} \sup_{s \in [0,1]} \|\bar{Y}^n(s) - \bar{Y}(s)\| = 0 \quad \text{and} \quad \lim_{n \to \infty} \sup_{s \in [0,\tau^n_B]} \|c^n(s)\| = 0,
\]
we conclude that for all sufficiently large \( n \) \( \|\bar{X}^n(t)\| \leq 2B/3 \) for all \( t \in [0,\tau^n_B) \cap [0,1] \).

We claim that for all sufficiently large \( n \) \( \tau^n_B = \infty \). Indeed, if this were not the case, then for some subsequence of \( n \in \mathbb{N} \) there would exist \( j \in \{0,1,\ldots,n-1\} \) such that \( \|\bar{X}^n_j\| \leq 2B/3 \) and \( \|\bar{X}^n_{j+1}\| \geq B \). Since \( \|b(x)\| \) is uniformly bounded on \( \{x : \|x\| \leq B\} \), this would imply that
\[
\liminf_{n \to \infty} \|\bar{Y}^n_{j+1} - \bar{Y}^n_j\| = \liminf_{n \to \infty} \left\|\bar{X}^n_{j+1} - \bar{X}^n_j - \frac{1}{n} b(\bar{X}^n_j)\right\| \geq B/3
\]
as \( n \to \infty \) along this subsequence. Since this would contradict the uniform convergence of \( \bar{Y}^n \) to the continuous function \( \bar{Y} \), we conclude that \( \tau^n_B = \infty \) for all sufficiently large \( n \).

We have shown that for all sufficiently large \( n \)
\[
\|\bar{X}^n(t) - \bar{X}(t)\| \leq \left(\sup_{t \in [0,1]} \|\bar{Y}^n(t) - \bar{Y}(t)\| + \sup_{t \in [0,1]} \|c^n(t)\|\right) e^L
\]
and
\[
\lim_{n \to \infty} \sup_{t \in [0,1]} \|c^n(t)\| = 0.
\]
Thus the sequence \( \{\bar{X}^n\} \) converges uniformly on \([0,1]\) to the continuous process \( \bar{X} \), and so \( \{\bar{X}^n\} \) converges uniformly on \([0,1]\) to the same limit. Recall that \( \tilde{\nu}^n \) is simply a shifted version of \( \nu^n \) in the sense that for any bounded measurable function \( f \) on \( \mathbb{R}^d \times [0,1] \)
\[
\int_{\mathbb{R}^d \times [0,1]} f(y, s) \tilde{\nu}^n(dy \times ds) = \int_{\mathbb{R}^d \times [0,1]} f(y - b(\bar{X}^n(s)), s) \nu^n(dy \times ds).
\]
Recall also that
\[
\limsup_{n \to \infty} \int_{\mathbb{R}^d \times [0,1]} \|y\| \tilde{\nu}^n(dy \times ds) < \infty.
\]
Since \(b\) is continuous, the convergence of \(\{\tilde{\nu}^n\}\) to \(\tilde{\nu}\) and \(\{\tilde{X}^n\}\) to \(\tilde{X}\) imply the convergence of \(\{\nu^n\}\) to a measure \(\nu\), and for any bounded continuous function \(f\) on \(\mathbb{R}^d \times [0,1]\)
\[
\int_{\mathbb{R}^d \times [0,1]} f(y, s) \tilde{\nu}(dy \times ds) = \int_{\mathbb{R}^d \times [0,1]} f(y - b(\tilde{X}(s)), s) \nu(dy \times ds). \tag{E.13}
\]
Since by Fatou's Lemma \(\|y\|\) is integrable with respect to \(\tilde{\nu}\) and thus with respect to \(\nu\), equations (E.10) and (E.13) yield for all \(t \in [0,1]\)
\[
\Upsilon(t) = \int_{\mathbb{R}^d \times [0,t]} y \tilde{\nu}(dy \times ds)
= \int_{\mathbb{R}^d \times [0,t]} y \nu(dy \times ds) - \int_0^t b(\tilde{X}(s)) ds.
\]
Hence we obtain from (E.12) that for all \(t \in [0,1]\)
\[
\tilde{X}(t) = x + \int_{\mathbb{R}^d \times [0,t]} y \nu(dy \times ds). \tag{E.14}
\]
The probability–1 decomposition \(\tilde{\nu}(dy \times dt) = \tilde{\nu}(dy|t) \otimes dt\) and equation (E.13) yield the probability–1 decomposition \(\nu(dy \times dt) = \nu(dy|t) \otimes dt\), where
\[
\nu(dy|t) = \tilde{\nu}(dy - b(\tilde{X}(t))|t).
\]
Since we have already shown that w.p.1
\[
(\nu^n, \tilde{X}^n, \tilde{X}^n) \to (\nu, \tilde{X}, \tilde{X})
\]
and that (E.14) holds, we conclude that all the conclusions of Theorem 5.3.5 are valid. The proof of Theorem E.2.3 is complete. \(\blacksquare\)
Bibliography


BIBLIOGRAPHY


