From Large Deviations to Statistical Mechanics: What Is the Most Likely Way for an Unlikely Event To Happen?

Department of Mathematics CEMAT's Open Seminar

Instituto Superior Técnico, Lisbon

28 May 2014

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

rsellis@math.umass.edu
http://www.math.umass.edu/~rsellis

Abstract

This talk is an introduction to the theory of large deviations, which studies the asymptotic behavior of probabilities of rare events. The talk is accessible to a general mathematical audience including graduate students. The theory of large deviations has its roots in the work of Ludwig Boltzmann, the founder of statistical mechanics. In 1877 he did the first large deviation calculation in science when he showed that large deviation probabilities of the empirical vector could be expressed in terms of the relative entropy function. In this talk Boltzmann's insight is applied to prove a conditional limit theorem that addresses a basic issue arising in mathematics, statistical mechanics, and other applications. What is the most

likely way for an unlikely event to happen? This question is answered in the context of n tosses of a cubic die and other random experiments involving finitely many outcomes. Let X_i denote the outcome of the i'th toss and define $S_n = X_1 + ... + X_n$. If the die were fair, then one would expect that for large n, S_n/n should be close to the theoretical mean of 3.5. Given that n is large but that S_n/n is close to a number z not equal to 3.5, the problem is to compute, in the limit n to infinity, the probability of obtaining $k = 1, 2, 3, 4, 5, 6$ on a single toss. Interestingly, this conditional limit theorem is intimately related to statistical mechanics because it gives a rigorous derivation, for a random ideal gas, of a basic construction due to Gibbs; namely, to derive the form of the canonical ensemble from the microcanonical ensemble. A related conditional limit theorem for the distribution of X_1, X_2 illustrates the phenomenon of propagation of chaos.

Our Lives Are Large Deviations

Statistically, the probability of any one of us being here is so small that you'd think the mere fact of existing would keep us all in a contented dazzlement of surprise. We are alive against the stupendous odds of genetics, infinitely outnumbered by all the alternates who might, except for luck, be in our places.

Even more astounding is our statistical improbability in physical terms. The normal, predictable state of matter throughout the universe is randomness, a relaxed sort of equilibrium, with atoms and their particles scattered around in an amorphous muddle. We, in brilliant contrast, are completely organized structures, squirming with information at every covalent bond. We make our living by catching electrons at the moment of their excitement by solar photons, swiping the energy released at the instant of each jump and storing it up in intricate loops for ourselves. We violate probability, by our nature. To be able to do this systemically, and in such wild varieties of form, from viruses to whales, is extremely unlikely; to have sustained the effort successfully for the several billion years of our existence, without drifting back into randomness, was nearly a mathematical impossibility.

Lewis Thomas, *The Lives of a Cell*
(New York: Viking Press, 1974), p. 141

"Life, by Any Reasonable Measure, Is Impossible"

Art is a way of saying what it means to be alive, and the most salient feature of existence is the unthinkable odds against it. For every way that there is of being here, there are an infinity of ways of not being here. Historical accident snuffs out whole universes with every clock tick. Statistics declare us ridiculous. Thermodynamics prohibits us. Life, by any reasonable measure, is impossible, and my life—this, here, now—infinitely more so. Art is a way of saying, in the face of all that impossibility, just how worth celebrating it is to be able to say anything at all.

Richard Powers, *Conjunctions*,
quoted in John Leonard, "Mind Painting,"
The New York Review of Books,
11 January 2001, p. 47.

"Something Just Short of Infinity to One"

This is the kind of question Henry liked to put to himself when he was a schoolboy: what are the chances of this particular fish, from that shoal, off that continental shelf ending up in the pages, no, on this page of this copy of the *Daily Mirror*? Something just short of infinity to one. Similarly, the grains of sand on a beach, arranged just so. The random ordering of the world, the unimaginable odds against any particular condition, still please him. Even as a child, and especially after Aberfan[1], he never believed in fate or providence, or the future being made by someone in the sky. Instead, at every instant, a trillion trillion futures; the pickiness of pure chance and physical laws seemed like freedom from the scheming of a gloomy god.

Ian McEwan, *Saturday*
(New York: Nan A. Talese, 2005), pp. 228–229

[1] On 21 October 1966, 144 people, 116 of them children, were killed when thousands of tons of coal waste slid onto the village of Aberfan in South Wales.

The theory of large deviations studies the asymptotic behavior of probabilities of rare events in certain random systems. The main focus is on events whose probabilities decay exponentially fast as the size of the system goes to ∞. The theory has been applied to a wide range of problems in which detailed information on rare events is required. One is often interested not only in the probability of rare events but also in the characteristic behavior of the system as the rare event occurs. For example, in applications to queueing theory and communication systems, the rare event could represent an overload or breakdown of the system. In this case, large deviation methodology can lead to an efficient redesign of the system so that the overload or breakdown does not occur. In applications to statistical mechanics the theory of large deviations gives precise, exponential-order estimates that are perfectly suited for asymptotic analysis.

In this talk I will illustrate the basic ideas of the theory of large deviations by studying a crooked gambling game. As I will indicate later, these ideas are closely related to statistical mechanics. They originated with Ludwig Boltzmann, one of founders of statistical mechanics, who used these ideas to calculate the equilibrium distribution of a random ideal gas. I will comment on Boltzmann's work at the end of the talk. Because of this historical fact, the talk should really be called ”From Statistical Mechanics to Large Deviations” rather than the reverse. In order to emphasize the beauty and flexibility of the theory of large deviations, I will motive the results using a formal notation. Complete proofs are given in sections 3–5 of my lecture notes for École de Physique Les Houches. An updated version of these lecture notes is available at the URL http://www.math.umass.edu/~rsellis/pdf-files/Les-Houches-lectures.pdf (see the handout).

Main problem of the talk. Here is a quick overview. A more detailed description will be given in a few minutes. We toss a fair cubic die n times. Each of the six faces of the die has the same probability of occurring (namely, 1/6), and the individual tosses are independent, which means that no toss has any influence on any of the other tosses. Define S_n to be the sum of the n tosses. The quantity S_n/n takes values in the interval $[1,6]$. According to the law of large numbers

$$\lim_{n\to\infty} P_n\{S_n/n \sim 3.5\} = 1.$$

Therefore, for any $z \neq 3.5$, we have $P_n\{S_n/n \sim z\} \to 0$. In fact, this is a large deviation event that converges to 0 exponentially fast as $n \to \infty$.

The main problem of the talk refers to a *crooked* gambling game. Given $z \neq 3.5$, suppose that $S_n/n \sim z$. Conditioned on this large deviation event, we ask the following question. What are the probabilities that each of the numbers 1, 2, 3, 4, 5, 6 appears on a toss of the die? For reasons that I explain later we consider this in the limit $n \to \infty$. Thus the problem is to calculate for $1 \leq k \leq 6$

$$\rho_k^* = \lim_{n\to\infty} P_n\{X_1 = k \,|\, S_n/n \sim z\}.$$

As we will see later [Theorem 1 Restated, part (b)], in a well-defined sense the vector $\rho^* = (\rho_1^*, \ldots, \rho_6^*)$ is the most likely way for the unlikely event $\{S_n/n \sim z\}$ to happen. A related problem is to calculate for $1 \leq k, \ell \leq 6$

$$\rho^*_{k,\ell} = \lim_{n \to \infty} P_n\{X_1 = k, X_2 = \ell \,|\, S_n/n \sim z\}$$

and its generalization for $X_1, X_2, \ldots, X_r$, $r \geq 3$.

We will analyze this problem in the following more general context of a game having α different outcomes. The dice game corresponds to $\alpha = 6$, which is the number of possible outcomes on each toss of the die. As in the dice game, the individual plays of the general game are independent. This is modeled by choosing P_n in the last bullet to be product measure.

- $\alpha \geq 2$ an integer, the number of possible outcomes on each play of the game

Dice game: $\alpha = 6$

- $y_1 < y_2 < \ldots < y_\alpha$ real numbers, the possible outcomes on each play of the game

Dice game: $y_k = k$ for $1 \leq k \leq 6$

- $\rho_1, \rho_2, \ldots, \rho_\alpha$ positive numbers summing to 1; each y_k has probability ρ_k

Dice game: each $\rho_k = 1/6$

- $\Lambda = \{y_1, y_2, \ldots, y_\alpha\}$

Dice game: $\Lambda = \{1, 2, 3, 4, 5, 6\}$

- $\rho = (\rho_1, \rho_2, \ldots, \rho_\alpha)$ a probability vector or $\rho = \sum_{k=1}^{\alpha} \rho_k \delta_{y_k}$, a probability measure on Λ

Dice game: $\rho = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$ or $\rho = \sum_{k=1}^{6} \frac{1}{6}\delta_k$

- $\Omega_n = \Lambda^n$, the configuration space for n repetitions of the game

Dice game: $\Omega_n = \{1, 2, 3, 4, 5, 6\}^n$

- $\omega = (\omega_1, \omega_2, \ldots, \omega_n) \in \Omega_n$, where ω_j is the outcome of the j'th play
- $P_n = \rho^n$, finite product measure: $P_n\{\omega\} = \prod_{j=1}^{n} \rho_{k_j}$ if $\omega_j = y_{k_j}$

Dice game: $P_n\{\omega\} = 1/6^n$ for each $\omega \in \Omega_n$

- $X_j(\omega) = \omega_j$ for $\omega \in \Omega_n$ and $1 \leq j \leq n$. This is an i.i.d. sequence of random variables with common distribution $\rho = \sum_{k=1}^{\alpha} \rho_k \delta_{y_k}$.

Microscopic level of description. We play the game n times. Each $\omega \in \Lambda^n$ gives a microscopic description of the n plays. As in the next-to-last bullet, we define P_n to be the product measure ρ^n, and for $B \subset \Lambda^n$ we define

$$P_n\{B\} = \sum_{\omega \in B} P_n\{\omega\}.$$

For the dice game, since $P_n\{\omega\} = 1/6^n$, we have $P_n\{B\} = \text{card}(B)/6^n$. Although the microscopic level of description is precise, it is much too detailed to give useful information.

Macroscopic level of description: sample mean. There are various macroscopic levels of description, including the sample mean and the empirical vector. The simplest macroscopic level involves the sample mean. Given $\omega \in \Lambda^n$, define

$$S_n(\omega) = \sum_{j=1}^{n} X_j(\omega) = \sum_{j=1}^{n} \omega_j.$$

The **sample mean** S_n/n takes values in the closed interval $[y_1, y_\alpha]$. For the dice game we define $\bar{y} = 3.5 = \frac{1}{6}\sum_{k=1}^{6} k$. For the general game we define

$$\bar{y} = \sum_{k=1}^{\alpha} y_k \rho_k.$$

In both cases $\bar{y}$ equals $E^{P_n}\{X_1\}$, the mean of X_1, where E^{P_n} denotes expectation with respect to P_n. The **law of large numbers** implies that

$$\lim_{n \to \infty} P_n\{S_n/n \sim \bar{y}\} = 1.$$

The theory of large deviations (specifically, Cramér's Theorem) shows that there exists a function I mapping (y_1, y_α) into $[0, \infty)$ such that for any $z \in (y_1, y_\alpha)$

$$P_n\{S_n/n \sim z\} \approx \exp[-nI(z)] \text{ as } n \to \infty.$$

Notation. Let a be a sufficiently small number. The event $\{S_n/n \sim z\}$ is shorthand for the event

$$\begin{aligned}
&\{S_n/n \in [\bar{y} - a, \bar{y} + a]\} \text{ if } z = \bar{y}, \\
&\{S_n/n \in [z - a, z]\} \text{ if } y_1 < z < \bar{y}, \\
&\{S_n/n \in [z, z + a]\} \text{ if } \bar{y} < z < y_\alpha.
\end{aligned}$$

In each of the three cases we choose a so small that the respective interval is a subset of (y_1, y_α).

With reference to this notation, the large deviation estimate appearing just before the last paragraph means that

$$\lim_{a \to 0} \lim_{n \to \infty} \frac{1}{n} \log P_n\{S_n/n \sim z\} = -I(z).$$

We now combine the law of large numbers with the large deviation estimate. If $z = \bar{y}$, then $P_n\{S_n/n \sim \bar{y}\} \to 1$, and if $z \neq \bar{y}$, then $P_n\{S_n/n \sim z\} \to 0$. It follows that if $z = \bar{y}$, then $I(z) = 0$. On the other hand, if $z \neq \bar{y}$, then $I(z) > 0$, implying that

$$P_n\{S_n/n \sim z\} \approx \exp[-nI(z)] \to 0 \text{ exponentially fast.}$$

If $z \neq \bar{y}$, then we call the event $\{S_n/n \sim z\}$ a **large deviation event**. The function I is called a **rate function** or an **entropy function**.

For the dice game

$$P_n\{S_n/n \sim z\} = \frac{1}{6^n} \cdot \text{card}\{\omega \in \Lambda^n : S_n(\omega)/n \sim z\}.$$

Thus $I(z)$ records the multiplicity of microstates ω consistent with the macrostate z through the macroscopic variable S_n/n. This interpretation of the rate function is consistent with Boltzmann's insight of 1877 concerning the role of entropy in statistical mechanics, which uses probability theory to study systems consisting of large numbers of particles. Boltzmann's insight is the following. "Entropy is a bridge between a microscopic level, on which physical systems are defined in terms of the complicated interactions among the individual constituent particles, and a macroscopic level, on which the laws describing the behavior of the system are formulated."

Main problem of the talk. The main problem of the talk refers to a *crooked* game. Given $z \in (y_1, y_\alpha)$, $z \neq \bar{y}$, suppose that $S_n/n \sim z$. Conditioned on this large deviation event, we ask the following question. What are the $n \to \infty$ limits of the probabilities that each of the outcomes $y_1, y_2, \ldots, y_\alpha$ appears on a play of the game?

Part (a) of the next theorem gives the form of this limit. Parts (c) and (d) give the surprising generalizations to the $n \to \infty$ limits of the probabilities of the outcomes on two successive plays of the game and on r successive plays of the game. My goal in this talk is to give the main ideas of the proof of part (a). If I have time, then I will remark on the proofs of parts (c) and (d) and the relationship of part (d) to statistical mechanics.

Theorem 1. For $z \in (y_1, y_\alpha)$ the following results hold.

(a) For $1 \leq k \leq \alpha$

$$\rho_k^* = \lim_{n \to \infty} P_n\{X_1 = y_k \,|\, S_n/n \sim z\}$$

exists and for a suitable choice of β

$$\rho_k^* = \frac{1}{\text{Normalization}} \cdot \exp[\beta y_k]\, \rho_k.$$

(b) In the sense of part (b) of Theorem 1 Restated ($\lim_{n \to \infty} P_n\{L_n \sim \rho^* \,|\, S_n/n \sim z\} = 1$), $\rho^* = (\rho_1^*, \ldots, \rho_\alpha^*)$ is the most likely way for the unlikely event $\{S_n/n \sim z\}$ to happen.

(c) For $1 \leq k, \ell \leq \alpha$

$$\rho_{k,\ell}^* = \lim_{n \to \infty} P_n\{X_1 = y_k, X_2 = y_\ell \,|\, S_n/n \sim z\}$$

exists and equals $\rho^*_k \rho^*_\ell$. Thus, although X_1 and X_2 are not independent when conditioned on $S_n/n \sim z$, in the limit $n \to \infty$ we recover the independence with one-dimensional marginals ρ^*.

(d) For a positive integer $r \geq 3$ the limiting conditional distribution of $X_1, X_2, \ldots, X_r$, conditioned on $S_n/n \sim z$, equals the r-fold product measure ρ^{*r}; i.e., for any subset $B \subset \Omega_r$

$$\lim_{n\to\infty} P_n\{(X_1, X_2, \ldots, X_r) \in B \,|\, S_n/n \sim z\} = \rho^{*r}\{B\}.$$

Thus, although $X_1, X_2, \ldots, X_r$ are not independent when conditioned on $S_n/n \sim z$, in the limit $n \to \infty$ we recover the independence with one-dimensional marginals ρ^*.

In order to see that this result is plausible, we consider the special case $z = \bar{y} = \sum_{k=1}^{\alpha} y_k \rho_k$. In this case the law of large numbers implies that $P_n\{S_n/n \sim z\} \to 1$ as $n \to \infty$. Hence

$$\rho^*_k = \lim_{n\to\infty} P_n\{X_1 = y_k\} = \rho_k \text{ and } \rho^*_{k,\ell} = \lim_{n\to\infty} P_n\{X_1 = y_k, X_2 = y_\ell\} = \rho_k \rho_\ell.$$

There is a similar statement for part (d).

Strategy of proof of part (a). The quantity ρ^*_k is the limiting conditional probability of the event that $X_1 = y_k$ given the large deviation event $S_n/n \sim z$. The key to proving part (a) of Theorem 1 is to rewrite these two events in terms of a new macroscopic variable, the empirical vector, and to appeal to a large deviation estimate for the empirical vector discovered by Boltzmann and proved rigorously by Sanov in 1957.

Macroscopic level of description: empirical vector. The most elementary macroscopic level of description is in terms of the sample mean

$$S_n(\omega)/n = \frac{1}{n} \cdot \sum_{j=1}^{m} X_j(\omega) = \frac{1}{n} \cdot \sum_{j=1}^{m} \omega_j.$$

This macroscopic variable summarizes the α^n degrees of freedom in the microscopic description $\omega \in \Omega_n$ in terms of a single quantity. A more refined macroscopic level of description is in terms of the empirical vector. For $\omega \in \Omega_n$ and $y \in \Lambda$ define

$$L_n(y) = L_n(\omega, y) = \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j(\omega)}\{y\}.$$

Thus $L_n(\omega, y)$ counts the relative frequency with which y appears in the configuration ω; in symbols, $L_n(\omega, y) = n^{-1} \cdot \text{card}\{j \in \{1, \ldots, n\} : \omega_j = y\}$. We then define the empirical vector

$$\begin{aligned} L_n &= L_n(\omega) = (L_n(\omega, y_1), \ldots, L_n(\omega, y_\alpha)) \\ &= \frac{1}{n} \sum_{j=1}^{n} \left(\delta_{X_j(\omega)}\{y_1\}, \ldots, \delta_{X_j(\omega)}\{y_\alpha\}\right). \end{aligned}$$

L_n equals the sample mean of the i.i.d. random vectors $(\delta_{X_j(\omega)}\{y_1\}, \ldots, \delta_{X_j(\omega)}\{y_\alpha\})$. It takes values in the set of probability vectors

$$\mathcal{P}_\alpha = \left\{ \gamma = (\gamma_1, \gamma_2, \ldots, \gamma_\alpha) \in \mathbb{R}^\alpha : \gamma_k \geq 0, \sum_{k=1}^{\alpha} \gamma_k = 1 \right\}.$$

The macroscopic variables L_n and S_n/n are closely related. In fact, the mean of the empirical vector L_n equals the sample mean S_n/n; in symbols, for each $\omega \in \Omega_n$, $\sum_{k=1}^{\alpha} y_k L_n(y_k, \omega) = S_n(\omega)/n$. This fact is used in the proof of part (a) of Theorem 1, where I will say more about it.

The limiting behavior of L_n is straightforward to determine. Since the X_j have the common distribution ρ, for each $y_k \in \Lambda$

$$E^{P_n}\{L_n(y_k)\} = E^{P_n}\left\{ \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j}\{y_k\} \right\} = \frac{1}{n} \sum_{j=1}^{n} P_n\{X_j = y_k\} = \rho_k,$$

where E^{P_n} denotes expectation with respect to P_n. Hence by the law of large numbers for the sample means of i.i.d. random variables

$$\lim_{n\to\infty} P_n\{L_n \sim \rho\} = 1.$$

Notation. For $\gamma \in \mathcal{P}_\alpha$ the event $\{L_n \sim \gamma\}$ is shorthand for the event $\{L_n \in B(\gamma, \varepsilon)\}$ for $\varepsilon > 0$, where $B(\gamma, \varepsilon)$ is the open ball $\{\nu \in \mathcal{P}_\alpha : \|\gamma - \nu\| < \varepsilon\}$. In this definition $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^\alpha$. If $\gamma \neq \rho$, then we choose $\varepsilon > 0$ so small that $\rho \notin B(\gamma, \varepsilon)$.

It follows that for any $\gamma \in \mathcal{P}_\alpha$ not equal to ρ

$$\lim_{n\to\infty} P_n\{L_n \sim \gamma\} = 0.$$

Boltzmann's discovery, which was proved rigorously by Sanov, implies that these probabilities converge to 0 exponentially fast in n. The exponential decay rate is given in terms of the relative entropy, which we now define.

Definition 2. Relative Entropy. The relative entropy of $\gamma \in \mathcal{P}_\alpha$ with respect to ρ is defined by

$$I_\rho(\gamma) = \sum_{k=1}^{\alpha} \gamma_k \log \frac{\gamma_k}{\rho_k}.$$

The main property of the relative entropy that we need is that $I_\rho(\rho) = 0$ and that $I_\rho(\gamma) > 0$ for any $\gamma \in \mathcal{P}_\alpha$, $\gamma \neq \rho$ (Lem. 3).

The following two limit results are valid.

- **Law of large numbers.** $P_n\{L_n \sim \rho\} \to 1$.

- **Boltzmann–Sanov's large deviation estimate.** For $\gamma \in \mathcal{P}_\alpha, \gamma \neq \rho$, we have $I_\rho(\gamma) > 0$ and

$$P_n\{L_n \sim \gamma\} \approx \exp[-nI_\rho(\gamma)] \to 0 \text{ exponentially fast;}$$

i.e.,

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log P_n\{L_n \in B(\gamma, \varepsilon)\} = -I_\rho(\gamma).$$

Several properties of the relative entropy are given in the next lemma. The proof is typical of proofs of analogous results involving the relative entropy (see Propositions 7 and 8) in that we use a global, convexity-based inequality rather than calculus to determine where I_ρ attains its infimum over $\mathcal{P}_\alpha$. In the present case the global convexity inequality is that for $x \geq 0$, $x \log x \geq x - 1$ with equality if and only if $x = 1$.

Lemma 3. (a) For $\gamma \in \mathcal{P}_\alpha$, $I_\rho(\gamma) \geq 0$ and $I_\rho(\gamma) = 0 \iff \gamma = \rho$. Thus $I_\rho(\gamma)$ attains its infimum of 0 over $\mathcal{P}_\alpha$ at the unique measure $\gamma = \rho$.

(b) I_ρ is strictly convex on $\mathcal{P}_\alpha$.

Partial proof of part (a). If $\gamma = \rho$, then the definition of the relative entropy shows that $I_\rho(\gamma) = 0$. We now prove that $I_\rho(\gamma) \geq 0$ for all $\gamma \in \mathcal{P}_\alpha$. For $x \geq 0$ the graph of the strictly convex function $x \log x$ has the tangent line $y = x - 1$ at $x = 1$. Hence $x \log x \geq x - 1$ with equality if and only if $x = 1$. It follows that for any $\gamma \in \mathcal{P}_\alpha$

$$\frac{\gamma_k}{\rho_k} \log \frac{\gamma_k}{\rho_k} \geq \frac{\gamma_k}{\rho_k} - 1$$

with equality if and only if $\gamma_k = \rho_k$. Multiplying this inequality by ρ_k and summing over k yields

$$I_\rho(\gamma) = \sum_{k=1}^{\alpha} \gamma_k \log \frac{\gamma_k}{\rho_k} \geq \sum_{k=1}^{\alpha} (\gamma_k - \rho_k) = 0.$$

Remainder of the proof. (a) In order to prove that $I_\rho(\gamma) = 0$ if and only if $\gamma = \rho$, we must show that if $I_\rho(\gamma) = 0$, then $\gamma = \rho$. Assume that $I_\rho(\gamma) = 0$. Then

$$\begin{aligned} 0 &= \sum_{k=1}^{\alpha} \gamma_k \log \frac{\gamma_k}{\rho_k} \\ &= \sum_{k=1}^{\alpha} \left(\gamma_k \log \frac{\gamma_k}{\rho_k} - (\gamma_k - \rho_k) \right) \\ &= \sum_{k=1}^{\alpha} \rho_k \left(\frac{\gamma_k}{\rho_k} \log \frac{\gamma_k}{\rho_k} - \left(\frac{\gamma_k}{\rho_k} - 1 \right) \right). \end{aligned}$$

We now use the facts that $\rho_k > 0$ and that for $x \geq 0$, $x \log x \geq x - 1$ with equality if and only if $x = 1$. It follows that for each k, $\gamma_k = \rho_k$ and thus that $\gamma = \rho$. This completes the proof that $I_\rho(\gamma) \geq 0$ and $I_\rho(\gamma) = 0$ if and only if $\gamma = \rho$, which is the first assertion in the proposition.

(b) Since

$$I_\rho(\gamma) = \sum_{k=1}^{\alpha} \rho_k \frac{\gamma_k}{\rho_k} \log \frac{\gamma_k}{\rho_k},$$

the strict convexity of I_ρ is a consequence of the strict convexity of $x \log x$ for $x \geq 0$. ■

We are now ready to give Boltzmann's discovery, which we state using a heuristic notation. The formal calculations used to motivate the next result can easily be turned into a rigorous proof of an asymptotic theorem, known as Sanov's Theorem (page 13). From Boltzmann's momentous discovery both the theory of large deviations and the Gibbsian formulation of equilibrium statistical mechanics grew.

Theorem 4. Boltzmann's Discovery (1877). For $\gamma \in \mathcal{P}_\alpha$

$$P_n\{L_n \sim \gamma\} \approx \exp[-nI_\rho(\gamma)] \;\text{ as } n \to \infty.$$

If $\gamma \neq \rho$, then $I_\rho(\gamma) > 0$, and so $P_n\{L_n \sim \gamma\} \to 0$ exponentially fast.

Proof. We assume that γ is in the range of L_n. In this case $P_n(L_n = \gamma) > 0$. By elementary combinatorics

$$\begin{aligned} P_n\{L_n = \gamma\} &= P_n\left\{\omega \in \Omega_n : L_n(\omega) = \frac{1}{n}(n\gamma_1, n\gamma_2, \ldots, n\gamma_\alpha)\right\} \\ &= P_n\{\text{card}\{\omega_j = y_1\} = n\gamma_1, \ldots, \text{card}\{\omega_j = y_\alpha\} = n\gamma_\alpha\} \\ &= \frac{n!}{(n\gamma_1)!(n\gamma_2)!\cdots(n\gamma_\alpha)!}\, \rho_1^{n\gamma_1}\, \rho_2^{n\gamma_2} \,\cdots\, \rho_\alpha^{n\gamma_\alpha}. \end{aligned}$$

We now use Stirling's formula in the weak form $k! = k^k e^{-k} e^{\text{O}(\log k)}$. An elementary calculation yields

$$\frac{1}{n}\log P_n\{L_n = \gamma\} = -I_\rho(\gamma) + \text{O}\left(\frac{\log n}{n}\right).$$

The term $\text{O}(\log n/n)$ converges to 0 as $n \to \infty$. Hence multiplying both sides of the last display by n and exponentiating yields the results. ■

Here is the detailed calculation. Stirling's formula in the weak form $n! = n^n e^{-n} e^{\text{O}(\log n)}$ yields

$$\begin{aligned} &\frac{1}{n}\log P_n\{L_n = \gamma\} \\ &\quad = \frac{1}{n}\log\left(\frac{n!}{(n\gamma_1)!(n\gamma_2)!\cdots(n\gamma_\alpha)!}\right) + \sum_{k=1}^{\alpha}\gamma_k \log\rho_k \\ &\quad = \frac{1}{n}\log\left(\frac{n^n e^{-n}}{(n\gamma_1)^{n\gamma_1}e^{-n\gamma_1}\cdots(n\gamma_\alpha)^{n\gamma_\alpha}e^{-n\gamma_\alpha}}\right) + \sum_{k=1}^{\alpha}\gamma_k\log\rho_k + \text{O}\left(\frac{\log n}{n}\right) \\ &\quad = \frac{1}{n}\log\left(\frac{1}{\gamma_1^{n\gamma_1}\cdots\gamma_\alpha^{n\gamma_\alpha}}\right) + \sum_{k=1}^{\alpha}\gamma_k\log\rho_k + \text{O}\left(\frac{\log n}{n}\right) \\ &\quad = -\sum_{k=1}^{\alpha}\gamma_k\log\gamma_k + \sum_{k=1}^{\alpha}\gamma_k\log\rho_k + \text{O}\left(\frac{\log n}{n}\right) \\ &\quad = -\sum_{k=1}^{\alpha}\gamma_k\log\frac{\gamma_k}{\rho_k} + \text{O}\left(\frac{\log n}{n}\right) \;=\; -I_\rho(\gamma) + \text{O}\left(\frac{\log n}{n}\right). \end{aligned}$$

The term O($\log n/n$) converges to 0 as $n \to \infty$. Hence multiplying both sides of the last display by n and exponentiating yields the results.
■

Theorem 4 is a local estimate, which we now convert into a global estimate. For C a Borel subset of $\mathcal{P}_\alpha$ define

$$I_\rho(C) = \inf_{\gamma \in C} I_\rho(\gamma).$$

For any $\gamma \in C$, $I_\rho(\gamma) \geq I_\rho(C)$. The range of $L_n(\omega)$ for $\omega \in \Omega_n$ is the set of probability vectors having the form k/n, where $k \in \mathbb{R}^\alpha$ has nonnegative integer coordinates summing to n; hence the cardinality of the range does not exceed $(n+1)^\alpha$. Since

$$P_n\{L_n \in C\} = \sum_{\gamma \in C} P_n\{L_n \sim \gamma\} \approx \sum_{\gamma \in C} \exp[-nI_\rho(\gamma)]$$

and

$$\exp[-nI_\rho(C)] \leq \sum_{\gamma \in C} \exp[-nI_\rho(\gamma)] \leq (n+1)^\alpha \exp[-nI_\rho(C)],$$

one expects that to exponential order the following result holds.

Corollary 5. $P_n\{L_n \in C\} \approx \exp[-nI_\rho(C)]$ as $n \to \infty$.

Here is a rigorous statement. Any open ball $B(\gamma, \varepsilon)$ satisfies the condition on C in the first sentence of the corollary.

Corollary 6. Let C be a Borel subset of $\mathcal{P}_\alpha$ satisfying cl(C) = cl(int(C)) (e.g., $C = B(\gamma, \varepsilon)$). If $\rho \notin$ cl(C), then $I_\rho(C) > 0$ and

$$\lim_{n \to \infty} \frac{1}{n} \log P_n\{L_n \in C\} = -I_\rho(C).$$

Hence $P_n\{L_n \in C\} \to 0$ exponentially fast.

Corollary 6 is a consequence of the following rigorous reformulation of Boltzmann's discovery, known as Sanov's Theorem, which expresses the large deviation principle for the empirical vectors L_n.

Sanov's Theorem (1957). The sequence of empirical vectors L_n satisfies the large deviation principle on $\mathcal{P}_\alpha$ with rate function I_ρ in the following sense.

(a) **Large deviation upper bound.** For any closed subset F of $\mathcal{P}_\alpha$

$$\limsup_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in F\} \leq -I_\rho(F).$$

(b) **Large deviation lower bound.** For any open subset G of $\mathcal{P}_\alpha$

$$\liminf_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in G\} \geq -I_\rho(G).$$

Comments on the Proof. For $\gamma \in \mathcal{P}_\alpha$ and $\varepsilon > 0$, $B(\gamma,\varepsilon)$ denotes the open ball with center γ and radius ε and $\overline{B}(\gamma,\varepsilon)$ denotes the corresponding closed ball. Since $\mathcal{P}_\alpha$ is a compact subset of $\mathbb{R}^\alpha$, any closed subset F of $\mathcal{P}_\alpha$ is automatically compact. By a standard covering argument it is not hard to show that the large deviation upper bound holds for any closed set F provided that one obtains the large deviation upper bound for any closed ball $\overline{B}(\gamma,\varepsilon)$:

$$\limsup_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in \overline{B}(\gamma,\varepsilon)\} \leq -I_\rho(\overline{B}(\gamma,\varepsilon)).$$

Likewise, the large deviation lower bound holds for any open set G provided one obtains the large deviation lower bound for any open ball $B(\gamma,\varepsilon)$:

$$\liminf_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in B(\gamma,\varepsilon)\} \geq -I_\rho(B(\gamma,\varepsilon)).$$

The bounds in the last two displays can be proved via combinatorics and Stirling's formula as in the heuristic proof of Theorem 5; one can easily adapt the calculations given in section 1.4 of my 1985 book, *Entropy, Large Deviations, and Statistical Mechanics.* The details are omitted. ■

Proof of Corollary 6 from Sanov's Theorem. We apply the large deviation upper bound to $\overline{C} = \text{cl}(C)$ and the large deviation lower bound to $C^\circ = \text{int}(C)$. Since $\overline{C} \supset C \supset C^\circ$, it follows that $I_\rho(\overline{C}) \leq I_\rho(C) \leq I_\rho(C^\circ)$ and that

$$\begin{aligned}
-I_\rho(\overline{C}\,) &\geq \limsup_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in \overline{C}\,\} \\
&\geq \limsup_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in C\} \\
&\geq \liminf_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in C\} \\
&\geq \liminf_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in C^\circ\} \\
&\geq -I_\rho(C^\circ).
\end{aligned}$$

The continuity of I_ρ on $\mathcal{P}_\alpha$ implies that $I_\rho(C^\circ) = I_\rho(\overline{C^\circ})$. Since by hypothesis $\overline{C^\circ} = \overline{C}$, we conclude that the extreme terms in this display are equal to each other and to $I_\rho(C)$ and thus that

$$\limsup_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in C\} = \liminf_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in C\} = -I_\rho(C).$$

The desired limit follows. ■

We now turn to the proof of part (a) of Theorem 1. For $z \in (y_1, y_\alpha)$ and for $1 \le k \le \alpha$ our goal is to evaluate

$$\rho_k^* = \lim_{n\to\infty} P_n\{X_1 = y_k \,|\, S_n/n \sim z\}.$$

Notation. For γ a probability vector in $\mathcal{P}_\alpha$, define

$$\langle\gamma\rangle = \sum_{k=1}^{\alpha} y_k \gamma_k.$$

This quantity equals the mean of the probability measure $\sum_{k=1}^{\alpha} \gamma_k \delta_{y_k}$.

If $z = \bar{y} = \sum_{k=1}^{\alpha} y_k \rho_k = \langle\rho\rangle$, then by the law of large numbers $P_n\{S_n \sim z\} \to 1$ as $n \to \infty$. As we have already seen, in this case

$$\rho_k^* = \lim_{n\to\infty} P_n\{X_1 = y_k \,|\, S_n/n \sim z\} = \lim_{n\to\infty} P_n\{X_1 = y_k\} = \rho_k.$$

In the next theorem we evaluate ρ_k^* when $z \neq \bar{y}$. The limit in part (a) of Theorem 1 involves X_1 and S_n/n, which are not both symmetric functions of X_j. The basic idea in proving this limit is to re-express the limit in terms of the empirical vector L_n, which is a symmetric function of X_j, and then to use Boltzmann's discovery in the form of Corollary 5. Part (b) motivates the statement that ρ^* is the most likely way for the unlikely event $\{S_n/n \sim z\}$ to happen.

Theorem 1 Restated [part (a)]. For any $z \in (y_1, y_\alpha)$ such that $z \neq \bar{y}$, the following results hold.

(a) For $1 \le k \le \alpha$, $\rho_k^* = \lim_{n\to\infty} P_n\{X_1 = y_k \,|\, S_n/n \sim z\}$ exists and has the form

$$\rho_k^* = \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j] \rho_j} \cdot \exp[\beta y_k]\, \rho_k,$$

where $\beta = \beta(z)$ is the unique value of β satisfying $\langle\rho^*\rangle = z$. Hence $\rho^* = (\rho_1^*, \rho_2^*, \ldots, \rho_\alpha^*) \in \mathcal{P}_\alpha$.

(b) The probability vector ρ^* in part (a) has the property that

$$\lim_{n\to\infty} P_n\{L_n \sim \rho^* \,|\, S_n/n \sim z\} = 1.$$

(c) Define $A = \{\gamma \in \mathcal{P}_\alpha : \langle\gamma\rangle \sim z\}$. Then ρ^* in part (a) has the property that

$$\lim_{n\to\infty} P_n\{L_n \sim \rho^* \,|\, L_n \in A\} = 1.$$

Proof. We prove that (b) ⇒ (a), then prove that (c) ⇒ (b), then prove (c).

(b) ⇒ (a). We assume that

$$\lim_{n\to\infty} P_n\{L_n \sim \rho^* \,|\, S_n/n \sim z\} = 1.$$

Then for all large n we have with probability close to 1 that

$$\begin{aligned}
\rho_k^* &= E^{P_n}\{\rho_k^* \,|\, S_n/n \sim z\} \;\approx\; E^{P_n}\{L_n(y_k) \,|\, S_n/n \sim z\} \\
&= \frac{1}{n}\sum_{j=1}^{n} E^{P_n}\{\delta_{X_j}(y_k) \,|\, S_n/n \sim z\} \\
&= \frac{1}{n}\sum_{j=1}^{n} P_n\{X_j = y_k \,|S_n/n \ \sim z\} \\
&= P_n\{X_1 = y_k \,|\, S_n/n \sim z\}.
\end{aligned}$$

The last line follows by symmetry. This completes the proof of part (a) from part (b).

(c) $\Rightarrow$ (b). We assume that

$$\lim_{n\to\infty} P_n\{L_n \sim \rho^* \,|\, L_n \in A\} = 1,$$

where

$$A = \{\gamma \in \mathcal{P}_\alpha \,:\, \langle \gamma \rangle \sim z\}.$$

Part (b) follows if we can prove that the events $\{L_n \in A\}$ and $\{S_n/n \sim z\}$ coincide. In order to see this, we note that $L_n \in A$ if and only if the mean $\langle L_n \rangle$ of the empirical vector satisfies $\langle L_n \rangle \sim z$. A straightforward calculation shows that $\langle L_n \rangle = S_n/n$, which implies that

$$L_n \in A \Longleftrightarrow \langle L_n \rangle \sim z \Longleftrightarrow S_n/n \sim z.$$

Thus the quantities whose limits are evaluated in parts (b) and (c) coincide.

Here is a quick proof that $\langle L_n \rangle = S_n/n$. For each $\omega \in \Omega_n$

$$\begin{aligned}
\langle L_n(\omega) \rangle &= \sum_{k=1}^{\alpha} y_k L_n(\omega, y_k) \\
&= \sum_{k=1}^{\alpha} y_k \cdot \frac{1}{n}\text{card}\{j \in \{1,2,\ldots,n\} : X_j(\omega) = y_k\} \\
&= \frac{1}{n}\sum_{j=1}^{n} X_j(\omega) = S_n(\omega)/n.
\end{aligned}$$

Here is a second proof that $\langle L_n \rangle = S_n/n$. For each $\omega \in \Omega_n$

$$\begin{aligned}
\langle L_n(\omega) \rangle = \sum_{k=1}^{\alpha} y_k L_n(\omega, y_k) &= \sum_{k=1}^{\alpha} y_k \cdot \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j(\omega)}(y_k) \\
&= \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{\alpha} y_k \delta_{X_j(\omega)}(y_k) \\
&= \frac{1}{n}\sum_{j=1}^{n} X_j(\omega) \;=\; S_n(\omega)/n.
\end{aligned}$$

(c). We now prove part (c), which we write in the form

$$\lim_{n\to\infty} P_n\{L_n \in B(\rho^*,\varepsilon) \,|\, L_n \in A\} = 1,$$

where again

$$A = \{\gamma \in \mathcal{P}_\alpha : \langle\gamma\rangle \sim z\}.$$

Heart of the proof. The key is to use the large deviation estimate

$$P_n\{L_n \in C\} \approx \exp[-nI_\rho(C)] \text{ as } n \to \infty$$

for Borel subsets C of $\mathcal{P}_\alpha$, which is discussed in Corollary 5. According to this estimate, to exponential order

$$\begin{aligned} P_n\{L_n \in B(\rho^*,\varepsilon) \,|\, L_n \in A\} &= P_n\{L_n \in B(\rho^*,\varepsilon) \cap A\} \cdot \frac{1}{P_n\{L_n \in A\}} \\ &\approx \exp[\,-n(I_\rho(B(\rho^*,\varepsilon) \cap A) - I_\rho(A))\,]. \end{aligned}$$

Thus one should obtain the conditioned limit

$$\lim_{n\to\infty} P_n\{L_n \in B(\rho^*,\varepsilon) \,|\, L_n \in A\} = 1$$

if $I_\rho(B(\rho^*,\varepsilon) \cap A) = I_\rho(A)$. By sending $\varepsilon \to 0$, we see that ρ^* must satisfy

$$I_\rho(\{\rho^*\} \cap A) = I_\rho(A);$$

i.e., $\rho^* \in A$ and the infimum of I_ρ on A is attained at the unique point ρ^*. This is proved in Proposition 8.

For $z \in (y_1, \bar{y})$ the set A has the form

$$A = \{\gamma \in \mathcal{P}_\alpha : \langle\gamma\rangle \sim z\} = \{\gamma \in \mathcal{P}_\alpha : \langle\gamma\rangle \in [z-a, z]\}.$$

We motivate Proposition 8 by replacing the constraint $\langle\gamma \sim [z-a,z]$ by the equality constraint $\langle\gamma\rangle = z$, which is easier to handle. We also consider any $z \in (y_1, y_\alpha)$, $z \neq \bar{y}$.

Proposition 7. Let $z \in (y_1, y_\alpha)$ be given, $z \neq \bar{y}$. Then I_ρ attains its infimum over

$$\widetilde{A} = \{\gamma \in \mathcal{P}_\alpha : \langle\gamma\rangle = \textstyle\sum_{k=1}^{\alpha} y_k\gamma_k = z\}$$

at the unique point $\rho^* = (\rho_1^*, \rho_2^*, \ldots, \rho_\alpha^*)$ defined in part (a) of Theorem 1 Restated: for each $k = 1, 2, \ldots, \alpha$

$$\rho_k^* = \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j]\,\rho_j} \cdot \exp[\beta y_k]\,\rho_k,$$

where $\beta = \beta(z) \neq 0$ is the unique value of β satisfying $\langle\rho^*\rangle = \sum_{k=1}^{\alpha} y_k\rho_k^* = z$.

The form of ρ^*_k reflects the form of the constraint on $\sum_{k=1}^{\alpha} y_k \gamma_k$. The quantities y_k in the constraint appear in the exponents defining ρ^*_k. This is a special case of a general principle.

Why does the point ρ^* arise as the unique minimum point of $I_\rho(\gamma)$ for $\gamma \in \widetilde{A}$? We motivate this by considering the calculus problem of determining critical points of $I_\rho(\gamma)$ for $\gamma \in \widetilde{A}$; i.e., for γ subject to the constraints that $\sum_{k=1}^{\alpha} \gamma_k = 1$ and $\sum_{k=1}^{\alpha} y_k \gamma_k = z$. Let λ and $-\beta$ be Lagrange multipliers corresponding to these two constraints. Since $I_\rho(\gamma) = \sum_{j=1}^{\alpha} \gamma_j \log(\gamma_j/\rho_j)$, for each k we have

$$\begin{aligned} 0 &= \frac{\partial(I_\rho(\gamma) + \lambda\left(\sum_{j=1}^{\alpha} \gamma_j - 1\right) - \beta\left(\sum_{j=1}^{\alpha} y_j \gamma_j - z\right)}{\partial \gamma_k} \\ &= \log \gamma_k + 1 - \log \rho_k + \lambda - \beta y_k. \end{aligned}$$

It follows that

$$\gamma_k = \exp[-\lambda - 1]\ \exp[\beta y_k]\, \rho_k.$$

Now pick λ so that $\sum_{k=1}^{\alpha} \gamma_k = 1$ and pick $\beta = \beta(z)$ so that $\sum_{k=1}^{\alpha} y_k \gamma_k = z$. With these choices of λ and β, γ_k equals ρ^*_k.

We now have a candidate ρ^* for the minimum point of $I_\rho(\gamma)$ on $\widetilde{A}$. The proof that ρ^* is the unique such minimum point is based on properties of the relative entropy and does not use calculus. We recall that for each $k \in \{1, \ldots, \alpha\}$

$$\frac{\rho^*_k}{\rho_k} = \frac{1}{V} \cdot \exp[\beta y_k],$$

where V is the normalization $\sum_{j=1}^{\alpha} \exp[\beta y_j] \rho_j$. We assume that there exists a unique value of $\beta \in \mathbb{R}$ such that $\langle \rho^* \rangle = z$; if so, then $\rho^* \in \widetilde{A}$. This fact is proved in Lemma 10. For any $\gamma \in \widetilde{A}$

$$\begin{aligned} I_\rho(\gamma) &= \sum_{k=1}^{\alpha} \gamma_k \log \frac{\gamma_k}{\rho_k} = \sum_{k=1}^{\alpha} \gamma_k \log \frac{\gamma_k}{\rho^*_k} + \sum_{k=1}^{\alpha} \gamma_k \log \frac{\rho^*_k}{\rho_k} \\ &= I_{\rho^*}(\gamma) + \beta \sum_{k=1}^{\alpha} y_k \gamma_k - (\log V) \sum_{k=1}^{\alpha} \gamma_k \\ &= I_{\rho^*}(\gamma) + \beta z - \log V. \end{aligned}$$

Since $I_{\rho^*}(\gamma) \geq 0$ with equality if and only if $\gamma = \rho^*$ [Lemma 3], it follows that if $\gamma = \rho^*$, then $I_\rho(\rho^*) = \beta z - \log V$ and that if $\gamma \neq \rho^*$, then

$$I_\rho(\gamma) > \beta z - \log V = I_\rho(\rho^*).$$

This completes the proof that I_ρ attains its infimum over $\widetilde{A}$ at the unique vector ρ^*.

Contraction principle relating L_n and S_n/n. We have just proved that

$$\inf\{I_\rho(\gamma) : \gamma \in \widetilde{A}\} = \inf\{I_\rho(\gamma) : \gamma \in \mathcal{P}_\alpha, \langle\gamma\rangle = z\} = \beta z - \log V,$$

where $V = \sum_{j=1}^{\alpha} \exp[\beta y_j]\rho_j$. It is not hard to show that the function $I(z) = \beta z - \log V$ equals the Cramér rate function in the large deviation principle for S_n/n. This is an example of the contraction principle applied to the map $\gamma \in \mathcal{P}_\alpha \mapsto \langle\gamma\rangle \in [y_1, y_\alpha]$. This map has the property that for each $\omega \in \Omega_n$, $S_n(\omega)/n = \langle L_n(\omega)\rangle$.

Statement and proof of Proposition 8. In Proposition 8 we prove that for $z \in (y_1, \bar{y})$, $I_\rho(\gamma)$ attains its infimum over A at the unique point ρ^*. The proof is similar for $z \in (\bar{y}, y_\alpha)$. In Lemmas 10 and 11 we show that there exists a unique value of $\beta = \beta(z) < 0$ such that $\sum_{k=1}^{\alpha} y_k \rho_k^* = z$.

Proposition 8. Let $z \in (y_1, \bar{y})$ be given. Then I_ρ attains its infimum over

$$A = \left\{\gamma \in \mathcal{P}_\alpha : \langle\gamma\rangle = \sum_{k=1}^{\alpha} y_k \gamma_k \in [z - a, z]\right\}$$

at the unique point $\rho^* = (\rho_1^*, \rho_2^*, \ldots, \rho_\alpha^*)$ defined in part (a) of Theorem 7: for each $k = 1, 2, \ldots, \alpha$

$$\rho_k^* = \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j} \cdot \exp[\beta y_k]\, \rho_k,$$

where $\beta = \beta(z) < 0$ is the unique value of β satisfying $\langle\rho^*\rangle = \sum_{k=1}^{\alpha} y_k \rho_k^* = z$.

Proof. Let

$$c(\beta) = \log\left(\sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j\right).$$

For each $k \in \{1, \ldots, \alpha\}$

$$\frac{\rho_k^*}{\rho_k} = \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j} \cdot \exp[\beta y_k] = \frac{1}{\exp[c(\beta)]} \cdot \exp[\beta y_k].$$

Hence for any $\gamma \in A$

$$\begin{aligned} I_\rho(\gamma) &= \sum_{k=1}^{\alpha} \gamma_k \log \frac{\gamma_k}{\rho_k} = \sum_{k=1}^{\alpha} \gamma_k \log \frac{\gamma_k}{\rho_k^*} + \sum_{k=1}^{\alpha} \gamma_k \log \frac{\rho_k^*}{\rho_k} \\ &= I_{\rho^*}(\gamma) + \beta \sum_{k=1}^{\alpha} y_k \gamma_k - c(\beta). \end{aligned}$$

Since $I_{\rho^*}(\rho^*) = 0$ and $\sum_{k=1}^{\alpha} y_k \rho_k^* = z$, it follows that

$$I_\rho(\rho^*) = I_{\rho^*}(\rho^*) + \beta \sum_{k=1}^{\alpha} y_k \rho_k^* - c(\beta) = \beta z - c(\beta).$$

Now consider any $\gamma \in A$, $\gamma \neq \rho^*$. Since $\beta < 0$, $\sum_{k=1}^{\alpha} y_k \gamma_k \leq z$, and $I_{\rho^*}(\gamma) \geq 0$ with equality if and only if $\gamma = \rho^*$, we obtain

$$\begin{aligned} I_\rho(\gamma) &= I_{\rho^*}(\gamma) + \beta \sum_{k=1}^{\alpha} y_k \gamma_k - c(\beta) \\ &> \beta \sum_{k=1}^{\alpha} y_k \gamma_k - c(\beta) \geq \beta z - c(\beta) = I_\rho(\rho^*). \end{aligned}$$

We conclude that for any $\gamma \in A$, $I_\rho(\gamma) \geq I_\rho(\rho^*)$ with equality if and only if $\gamma = \rho^*$. Thus I_ρ attains its infimum over A at the unique point ρ^*. The proof of the proposition is complete. ■

We complete the proof of part (a) of Theorem 1 by showing how that there exists a unique value of $\beta = \beta(z)$ such that $\langle \rho^* \rangle = z$, where

$$\rho_k^* = \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j} \cdot \exp[\beta y_k]\, \rho_k.$$

Lemma 9. The quantity

$$\langle \rho^* \rangle = \sum_{k=1}^{\alpha} y_k \rho_k^* = \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j} \cdot \sum_{k=1}^{\alpha} y_k \exp[\beta y_k]\, \rho_k$$

is a continuous, strictly increasing function of $\beta \in \mathbb{R}$ with range (y_1, y_α). For $\beta < 0$ the range is $(y_1, \bar{y})$, for $\beta = 0$ the range is $\bar{y}$, and for $\beta > 0$ the range is $(\bar{y}, y_\alpha)$. Hence for each $z \in (y_1, y_\alpha)$ there exists a unique value of $\beta \in \mathbb{R}$ such that $\langle \rho^* \rangle = z$.

Proof. As $\beta \to \infty$, $\langle \rho^* \rangle \to y_\alpha$, and as $\beta \to -\infty$, $\langle \rho^* \rangle \to y_1$. In the next lemma we show that

$$c(\beta) = \log\left(\sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j \right)$$

has the properties that $c'(\beta) = \langle \rho^* \rangle$ and $c''(\beta) > 0$ for all $\beta \in \mathbb{R}$. In addition $c'(0) = \langle \rho \rangle = \bar{y}$. It follows that for $\beta < 0$ the range of $\langle \rho^* \rangle = c'(\beta)$ is $(y_1, \bar{y})$, for $\beta = 0$ the range is $\bar{y}$, and for $\beta > 0$ the range is $(\bar{y}, y_\alpha)$. This property of the range of $\langle \rho^* \rangle$ yields the last assertion. ■

We now prove that $c(\beta)$ has the properties stated in the last proof.

Lemma 10. For $\beta \in \mathbb{R}$, $c(\beta)$ has the following properties.

(a) $c''(\beta) > 0$ for all β; i.e., c is strictly convex on $\mathbb{R}$.

(b) $c'(0) = \sum_{k=1}^{\alpha} y_k \rho_k = \bar{y}$.

(c) $c'(\beta) \to y_1$ as $\beta \to -\infty$ and $c'(\beta) \to y_\alpha$ as $\beta \to \infty$.

(d) c' is a one-to-one function mapping $\mathbb{R}$ onto the open interval (y_1, y_α), which is the interior of the smallest interval containing the set $\Lambda = \{y_1, y_2, \ldots, y_\alpha\}$.

Proof. (a) We define

$$\langle y \rangle_\beta = \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j} \cdot \sum_{k=1}^{\alpha} y_k \, \exp[\beta y_k]\, \rho_k$$

and

$$\langle (y - \langle y \rangle_\beta)^2 \rangle_\beta = \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j} \cdot \sum_{k=1}^{\alpha} (y_k - \langle y \rangle_\beta)^2 \, \exp[\beta y_k]\, \rho_k$$

and calculate

$$c'(\beta) = \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j} \cdot \sum_{k=1}^{\alpha} y_k \, \exp[\beta y_k]\, \rho_k = \langle y \rangle_\beta$$

and

$$\begin{aligned} c''(\beta) &= \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j} \cdot \sum_{k=1}^{\alpha} y_k^2 \, \exp[\beta y_k]\, \rho_k - \langle y \rangle_\beta^2 \\ &= \langle (y - \langle y \rangle_\beta)^2 \rangle_\beta \, > \, 0. \end{aligned}$$

The last line gives part (a). This calculation shows that $c'(\beta)$ equals the mean of the probability vector $\exp[\beta y_k]\rho_k / \sum_{j=1}^{\alpha} \exp[\beta y_j]\,\rho_j$, and $c''(\beta)$ equals the variance of this probability vector.

(b) This follows from the formula for $c'(\beta)$ in part (a).

(c) Since $y_1 < y_j$ for all $j = 2, \ldots, \alpha$,

$$\lim_{\beta \to -\infty} c'(\beta) = \lim_{\beta \to -\infty} \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta(y_j - y_1)]\,\rho_j} \cdot \sum_{k=1}^{\alpha} y_k \, \exp[\beta(y_k - y_1]\,\rho_k = y_1.$$

One similarly proves that $\lim_{\beta \to \infty} c'(\beta) = y_\alpha$.

(d) According to part (a), $c'(\beta)$ is a strictly increasing function of β. It follows from the limits in part (c) that c' is a one-to-one function mapping $\mathbb{R}$ onto the open interval (y_1, y_α). This completes the proof of the lemma. ■

Boltzmann introduced large deviation techniques into science when he calculated the equilibrium distribution of energy levels in a random ideal gas. This equilibrium distribution is known as the Maxwell–Boltzmann distribution. In fact, in Theorem 1 Restated I solve precisely the same problem solved by Boltzmann by essentially the same method. In the model introduced in this talk, we let X_j for $1 \le j \le n$ denote the energy levels of a random ideal gas consisting of non-interacting particles. The possible energy levels are denoted by $\Lambda = \{y_1, y_2, \ldots, y_\alpha\}$, which are an increasing sequence of real numbers. We assume that the X_j are i.i.d. with common distribution $\rho = \sum_{k=1}^{\alpha} \rho_k \delta_{y_k}$, where $\rho_k = 1/\alpha$ for each k. We also assume that the average energy $S_n/n \sim z$, where z is a fixed number in $(y_1, \ldots, y_\alpha)$. In Theorem 1 Restated I prove that

$$\lim_{n \to \infty} P_n\{L_n \sim \rho^* | L_n \in A\} = \lim_{n \to \infty} P_n\{L_n \sim \rho^* | \langle L_n \rangle \sim z\} = 1$$

if ρ^* is the probability measure defined in part (a) of that theorem. In this context ρ^* is the Maxwell–Boltzmann distribution.

In the talk I prove this assertion by using the global large deviation estimate in Corollary 5 to reduce the problem to calculating the probability measure $\rho^* \in \mathcal{P}_\alpha$ that minimizes the relative entropy $I_\rho(\gamma)$ over all probability measures $\gamma \in \mathcal{P}_\alpha$ having mean z. Boltzmann proves this assertion by an equivalent technique, which starts with the local large deviation estimate

$$P_n\{L_n \sim \gamma\} \approx \exp[-nI_\rho(\gamma)] \text{ as } n \to \infty.$$

He then argues that since $\langle L_n \rangle = S_n/n \sim z$ and since $L_n \sim \gamma$, the equilibrium measure ρ^* is that probability measure that has mean z and maximizes the probability in the last display over all $\gamma \in \mathcal{P}_\alpha$ having mean z, or equivalently the probability measure that minimizes the relative entropy $I_\rho(\gamma)$ over all $\gamma \in \mathcal{P}_\alpha$ having mean z. This is precisely the way I prove the assertion made at the end of the preceding paragraph.

I end this talk with a number of comments. In comment 3 I relate the limit in part (d) of Theorem 1 to statistical mechanics.

1. Part (c) of Theorem 1 states that for $1 \leq k, \ell \leq \alpha$

$$\lim_{n \to \infty} P_n\{X_1 = y_k, X_2 = y_\ell \,|\, S_n/n \sim z\} = \rho^*_k \rho^*_\ell.$$

This gives the surprising conclusion that although X_1 and X_2 are not independent when conditioned on $S_n/n \sim z$, in the limit $n \to \infty$ we recover the independence. One proves this limit result by rewriting the conditional distribution of X_1, X_2 conditioned on the event $\{S_n/n \sim z\}$ in terms of the empirical pair vector and appealing to the large deviation estimate for the empirical pair vector that is analogous to the Boltzmann–Sanov result.

This proof is based on standard ideas in information theory, which were pointed out to me by Neri Merhav, Department of Electrical Engineering at the Technion – Israel Institute of Technology in Haifa, Israel. I am grateful to him for sharing this proof with me on April 26, 2010 after my talk at the Technion. In order to simplify the notation we assume that n is a positive even integer. We start with the definition of the empirical pair vector. For $\omega \in \Omega_n$ and $k, \ell \in \{1, \ldots, \alpha\}$ define

$$\begin{aligned} L_{n,2}(\{y_k, y_\ell\}) &= L_{n,2}(\omega, \{y_k, y_\ell\}) \\ &= \frac{1}{n/2}\left(\sum_{j=1}^{n/2} \delta_{X_{2j-1}(\omega), X_{2j}(\omega)}\{y_k, y_\ell\}\right) \\ &= \frac{1}{n/2}\left(\sum_{j=1}^{n/2} \delta_{X_{2j-1}}(\omega)\{y_k\} \times \delta_{X_{2j}}(\omega)\{y_\ell\}\right) \end{aligned}$$

This counts the relative frequency with which the pair $\{y_k, y_\ell\}$ appears in the configuration $((\omega_1, \omega_2), (\omega_3, \omega_4), \ldots, (\omega_{n-1}, \omega_n))$. We then define the empirical pair vector

$$L_{n,2} = \{L_{n,2}(\{y_k, y_\ell\}), k, \ell = 1, \ldots, \alpha\}.$$

For each ω, $L_{n,2}$ takes values in the set $\mathcal{P}_{\alpha,2}$ consisting of all probability measures on $\Lambda \times \Lambda$. For $1 \leq j \leq n/2$ the sequence (X_{2j-1}, X_{2j}) is i.i.d. with common distribution $\rho \times \rho$. By Sanov's Theorem the sequence of empirical pair vectors $L_{n,2}$ satisfies the large deviation principle on $\mathcal{P}_{\alpha,2}$ with rate function

$$I_{\rho\times\rho}(\gamma) = \sum_{k,\ell=1}^{\alpha} \gamma_{k,\ell} \log \frac{\gamma_{k,\ell}}{\rho_k \times \rho_\ell}.$$

As in Lemma 3, for $\gamma \in \mathcal{P}_{\alpha,2}$, $I_{\rho\times\rho}(\gamma) \geq 0$ and $I_{\rho\times\rho}(\gamma) = 0 \Longleftrightarrow \gamma = \rho \times \rho$. In order to prove part (c) of Theorem 1, we rewrite, in terms of $L_{n,2}$, the conditional distribution of X_1, X_2 conditioned on the event $\{S_n/n \sim z\}$ and follow the pattern of the proof of part (a) of Theorem 1. The only changes are changes in notation.

2. Part (d) of Theorem 1 states that for any positive integer $r \geq 3$ the limiting conditional distribution of $X_1, X_2, \ldots, X_r$ conditioned on $S_n/n \sim z$ equals the r-fold product measure ρ^{*r}. One proves this as in the case $r = 2$ by rewriting the conditional distribution of $X_1, X_2, \ldots, X_r$ conditioned on $S_n/n \sim z$ in terms of the obvious generalization of the empirical pair measure and again following the pattern of the proof of part (a) of Theorem 1.

3. We recall that $\rho^* = (\rho_1^*, \rho_2^*, \ldots, \rho_k^*)$, where

$$\rho_k^* = \frac{1}{\text{Normalization}} \cdot \exp[\beta y_k]\, \rho_k.$$

   I would like to give a statistical mechanics interpretation of the result stated in the preceding item. We start by rewriting the limit in part (d) of Theorem 1 in the following nice form. For $r \geq 3$ and any subset B of Ω_r

$$\lim_{n\to\infty} P_n\{(X_1, X_2, \ldots, X_r) \in B \,|\, S_n/n \sim z\} \\ = \rho^{*r}\{B\} = \frac{1}{\text{Normalization}} \cdot \int_B \exp[\beta S_r]\, dP_r.$$

   This limit has an interesting interpretation for a simple physical system known as a discrete ideal gas. This system consists of n noninteracting particles, each of which can have an energy value $y_k \in \Lambda$ with probability $\rho_k = 1/\alpha$. For $\omega \in \Omega_n$, $S_n(\omega)$ denotes the total energy in ω. For this system the conditional measure $P_n\{\cdot \,|\, S_n/n \sim z\}$ defines the microcanonical ensemble on the energy shell consisting of all $\omega \in \Omega_n$ for which $S_n(\omega)/n \sim z$. The product measure ρ^{*r} defines the canonical ensemble. The limit in the last display implies that the two ensembles are equivalent.

4. The result stated in the preceding item is a special case of the general problem of the equivalence of the microcanonical and canonical ensembles for statistical models of turbulence and spin systems. A complete answer to this problem has been obtained by Bruce Turkington, myself, and co-workers in terms of concavity properties of the microcanonical entropy.