INFORMATION METRICS FOR LONG-TIME ERRORS IN SPLITTING SCHEMES FOR STOCHASTIC DYNAMICS AND PARALLEL KINETIC MONTE CARLO

KONSTANTINOS GOURGOULIAS†, MARKOS A. KATSOULAKIS†, AND LUC REY-BELLET†

Abstract. We propose an information-theoretic approach to analyze the long-time behavior of numerical splitting schemes for stochastic dynamics, focusing primarily on parallel kinetic Monte Carlo (KMC) algorithms. Established methods for numerical operator splittings provide error estimates in finite-time regimes, in terms of the order of the local error and the associated commutator. Path-space information-theoretic tools such as the relative entropy rate allow us to control long-time error through commutator calculations. Furthermore, they give rise to an a posteriori representation of the error which can thus be tracked in the course of a simulation. Another outcome of our analysis is the derivation of a path-space information criterion for comparison (and possibly design) of numerical schemes, in analogy to classical information criteria for model selection and discrimination. In the context of parallel KMC, our analysis allows us to select schemes with improved numerical error and more efficient processor communication. We expect that such a path-space information perspective on numerical methods will be broadly applicable in stochastic dynamics, for both the finite and the long-time regime.

Key words. parallel kinetic Monte Carlo, operator splitting schemes, information theory, relative entropy, relative entropy rate, long-time errors, graph connectivity, information criteria

AMS subject classifications. 65C05, 65C20, 82C20

DOI. 10.1137/15M1047271

1. Introduction. Recently, schemes that depend on operator splitting have found wide applicability within the domain of simulation of complex chemical reaction systems, biological systems, or those that can be modeled by appropriate Markov processes, for example, interacting particle systems. The recipe of splitting the system into components that can be simulated separately in an appropriate manner has led to more efficient algorithms, sometimes because some of the components can be solved explicitly, as in chemical reaction systems [1], and sometimes because the splitting allows for parallel computations [2, 3].

In parallel with the development of those algorithms, there has also been a growing amount of work toward the numerical analysis of splitting methods for stochastic dynamics in different contexts [1, 2, 4, 5, 6, 7]. In particular, for the case of parallel lattice kinetic Monte Carlo (PL-KMC), the authors in [2] developed a general framework, based on semigroup theory, that connects lattice decompositions to operator splitting. Then, in [4], error estimates were provided for bounded time intervals along with comparisons between different splitting schemes. One of the important contributions of the work was to highlight the connection of the error with the commutator associated with the splitting and how it affects the efficiency of the scheme.
Although classical techniques in numerical analysis, such as the study of the local error of the splitting scheme and expansions of the global error [8], work well in providing error estimates for bounded intervals, the information they provide is not of great use when the focus is on long-time results. Given that a common goal is sampling from a stationary distribution and convergence occurs for large simulation times, it thus makes sense to develop methodologies for the study of long-time errors. Approaches to tackling this problem are varied. For instance, in the case of SDEs, study of the long-time behavior has been done by employing Poisson equations [9]. For Lie–Trotter splittings, backward error analysis [10] has been used to study the performance of the schemes in capturing the stationary distribution when simulating Langevin dynamics (but see also [11]).

The main idea in this work is information-theoretical in nature, following similar successful approaches studying the irreversibility of numerical schemes [12], sensitivity analysis [13], and quantifying the loss of information in coarse-graining of particle systems [14]. In those, the authors use the relative entropy, along with other quantities derived from it, to both generate insights and provide computable quantities that are useful during a simulation. Besides that, approaching the problem from information theory still allows one to infer results about more classical metrics of error. For instance, one can derive upper bounds for the weak error of specific observables through the use of variational inequalities [15].

Our goal is to use another derived quantity, the relative entropy on path space per unit time, or relative entropy rate (RER), to quantify the long-time loss of information when using a splitting scheme. For our comparison, we fix a time step $\Delta t$ and then compare the $\Delta t$-skeleton chain arising from the exact process with the discrete chain we get from the approximate process. Through rigorous asymptotics, we provide an a posteriori error expansion of RER in terms of $\Delta t$ and connect RER with quantities central to the classical analysis of splitting schemes, like the commutator and the order of the local error of the splitting method. After deriving computable estimators from our a posteriori expansions for the highest-order term coefficients, we estimate them with the use of SPPARKS [3], a parallel KMC simulator, and use them to compare two well-known splitting schemes, the Lie and Strang splittings. Also, we illustrate how a practitioner can use the RER as an information criterion for selecting schemes that takes into account both long-time accuracy and communication cost. We then proceed to link the connectivity of the exact process with the RER asymptotics, which in turn allows for greater generality in the study of different operator splittings.

The plan for the following sections is as follows. In section 2, we provide the necessary background for KMC, PL-KMC, construction, and analysis of operator splitting schemes. Section 3 introduces the pathwise relative entropy and relative entropy per unit time, which are the principal tools used in this work. In section 4 we discuss the use of the RER as a metric for studying the long-time loss of information that operator splitting schemes can have and motivate the use of asymptotic expansions for its study. Section 5 is particularly important, as we study schemes through the RER in the context of stochastic particle systems and continue to section 6 with some discussion about time-step selection and the balance between error and communication in parallel KMC. Then, in section 7, we highlight some connections between the proposed framework and model selection with information criteria. Section 8 studies the RER for operator splitting schemes in a more general setting with the use of ideas from graph theory. Finally, in section 9, we demonstrate that the RER can also be applied in transient regimes, before the simulation has converged to stationarity.
2. Background. Consider that the stochastic process of interest is an ergodic continuous time Markov Chain (CTMC) \( X_t \) on a finite, but possibly still significantly large, state space \( S \). This stochastic process can be completely defined by its transition rates, \( q(\sigma, \sigma') \), which describe the probability of an update from state \( \sigma \) to state \( \sigma' \) in an infinitesimal period of time. That is,

\[
P(X_{t+\Delta t} = \sigma' | X_t = \sigma) = P_{\Delta t}(\sigma, \sigma') = q(\sigma, \sigma') \Delta t + o(\Delta t), \sigma \neq \sigma'.
\]

KMC works by simulating the embedded Markov Chain \( Y_n = X_{t_n} \), with jump times \( t_n, t_n \sim \exp(\lambda) \). The parameter \( \lambda(\sigma) \) is the total rate when the system is at state \( \sigma \),

\[
\lambda(\sigma) = \sum_{\sigma' \neq \sigma, \sigma' \in S} q(\sigma, \sigma').
\]

This allows us to write the transition probabilities of the embedded Markov Chain \( p(\sigma, \sigma') = q(\sigma, \sigma') / \lambda(\sigma) \). We can also define the infinitesimal generator \( L \) that corresponds to the Markov chain as follows. First, consider \( f \): bounded and continuous function on the state space \( S \). Then, \( L \) acts on \( f \) at the state \( \sigma \) as

\[
L[f](\sigma) = \sum_{\sigma' \in S} q(\sigma, \sigma') (f(\sigma') - f(\sigma)).
\]

Note that \( L[\delta_\sigma](\cdot) = q(\cdot, \sigma') \) for all states \( \sigma, \sigma' \), where \( \delta_\sigma(\cdot) = \delta(\cdot, \sigma') \) is a Dirac probability measure. We shall also use the notation \( L^k \) for the resulting operator after \( k \) successive compositions of \( L \). Because \( L^k[\delta_\sigma](\cdot) = L^{k-1}[L[\delta_\sigma]](\cdot) \), we see that, for any \( k \), \( L^k[\delta_\sigma](\cdot) \) is a computable object that depends on the transition rates.

Under fairly general conditions \([16]\), the transition probability of the Markov process can be written as in semigroup form, i.e., \( P_t(\sigma, \sigma') = e^{tL}\delta_\sigma(\cdot) \). In the case of interest to us, \( L \) is going to be a bounded operator and such operators allow for a representation of the semigroup with a series expansion.

**Lemma 2.1.** Let \( L \) be a linear and bounded operator, \( L : C_b(S) \to C_b(S) \), with \( C_b(S) \) being the set of continuous and bounded functions on the space \( S \). Then \( L \) generates a uniformly continuous semigroup \( e^{tL} \) which we can express in power series form.

\[
e^{tL} = \sum_{k=0}^{\infty} \frac{t^k}{k!}L^k.
\]

**Proof.** This is a classical result for which many references exist; see, for example, Chapter 1, p. 2, of Pazy \([17]\).

Thus, making use of Lemma 2.1, we can write the transition probability as

\[
P_t(\sigma, \sigma') = e^{tL}\delta_\sigma(\cdot) = \sum_{k=0}^{\infty} \frac{t^k}{k!}L^k[\delta_{\sigma'}](\cdot), \; \sigma, \sigma' \in S.
\]

2.1. Constructing approximations by semigroup splitting. We will now give the foundations of approximations by splitting methods, as applied to the simulation of CTMCs, and proceed with how those ideas are applied in the case of PL-KMC.

As mentioned earlier, the transition probability of the CTMC of interest can be written as \( e^{tL}\delta_\sigma(\cdot) \). The goal is for us to design a splitting scheme that can
approximate the action of $e^{tL}$. In our context, this leads to a new CTMC. One way to build such a scheme is to start with a splitting of the infinitesimal generator $L$ (2.3) into components $L_1, L_2$ with $L = L_1 + L_2$. Then, if we consider a positive $T$ and by using the Trotter product formula [18], we have

$$e^{TL} = \lim_{n \to \infty} (e^{T/nL_1} e^{T/nL_2})^n. \tag{2.6}$$

Correspondingly, if we now fix $n \in \mathbb{N}$ and set $\Delta t = T/n$, we can write approximations of $e^{TL}$ by using (2.6). For example, two such approximations are

$$e^{TL} \simeq (e^{\Delta t L_1} e^{\Delta t L_2})^n \quad \text{(Lie)},$$

$$e^{TL} \simeq (e^{\Delta t/2L_1} e^{\Delta t/2L_2})^n \quad \text{(Strang)}. \tag{2.7}$$

Therefore for a one-step transition from $t = 0$ to $\Delta t$, (2.7) can be written as

$$e^{L\Delta t} \simeq e^{\Delta t L_1} e^{\Delta t L_2}, \quad e^{L\Delta t} \simeq e^{\Delta t/2L_1} e^{\Delta t/2L_2}. \tag{2.8}$$

Operator splittings can also be carried out with multiple components, such as $L = L_1 + L_2 + L_3 + L_4$. Such a splitting is used for two-dimensional (2D) lattice decompositions in SPPARKS [3]. All arguments can be simply extended to those cases, but we stick to two components, $L_1, L_2$, for notational convenience.

Throughout this work, we use $P_{\Delta t}(\sigma, \sigma')$ to denote the probability $e^{L\Delta t}\delta_{\sigma'}(\sigma)$ and $Q_{\Delta t}(\sigma, \sigma')$ for the approximations arising from splittings of the semigroup. Since $L$ is a bounded operator, we can express $P_{\Delta t}$ as expansion (2.5). If we pick $L_1, L_2$ so that they are also bounded, then we can express $Q_{\Delta t}$ as an expansion too. For example, for the Lie splitting

$$\exp(\Delta t L_1) \exp(\Delta t L_2)\delta'_\sigma(\sigma) = \sum_{k=0}^{\infty} \frac{\Delta t^k}{k!} \left( k! \sum_{m=0}^{k} \frac{L_{1m}^m}{m!} \cdot \frac{L_{2k-m}^{k-m}}{(k-m)!} \right) \delta_{\sigma'}(\sigma), \tag{2.9}$$

which can be showed by multiplying the semigroup expansions of $\exp(\Delta t L_1)$ and $\exp(\Delta t L_2)$. Thus, if we use the notation

$$L_Q^k : = k! \sum_{m=0}^{k} \frac{L_{1m}^m}{m!} \cdot \frac{L_{2k-m}^{k-m}}{(k-m)!}, \tag{2.10}$$

we can write (2.9) in the form

$$Q_{\Delta t}(\sigma, \sigma') = \sum_{k=0}^{\infty} \frac{\Delta t^k}{k!} L_Q^k [\delta_{\sigma'}](\sigma). \tag{2.11}$$

By the definition of $L_Q^k$ in (2.10), $L_Q^0 = I$, $L_Q^1 = L$, $L_Q^2 = (L_1^2 + L_2^2 + 2L_1L_2)$, and so on, for the case of the Lie splitting. By a similar argument, we can write an expansion like (2.11) for other operator splitting approximations. In general, $L_Q$ is not a generator of a Markov process and, in that case, $L_Q^k$ is not equal $L_Q$ after $k$ compositions but is defined in the context of the expansion in (2.11). The slight abuse of notation allows us to compare the expansion of the exact process (2.5) with expansions of the approximating schemes of the form (2.11).
One way to compare the accuracy of using $Q_{\Delta t}$ as opposed to $P_{\Delta t}$ is to calculate the local error between expansion (2.5) and (2.11). As an example, here are the corresponding relations for the Lie and Strang splittings. We use $Q_{\Delta t}^{\text{Lie}}, Q_{\Delta t}^{\text{Strang}}$ for Lie and Strang, respectively. We will also use the notation $[L_1, L_2] := L_1 L_2 - L_2 L_1$ to denote the operator that captures the failure of $L_1$ and $L_2$ to commute. By using the expansions (2.5), (2.11), we can show that

\begin{align}
(2.12) \quad P_{\Delta t}(\sigma, \sigma') &= Q_{\Delta t}^{\text{Lie}}(\sigma, \sigma') + \frac{1}{2} [L_1, L_2] \delta_{\sigma'}(\sigma) \Delta t^2 + O(\Delta t^3), \\
(2.13) \quad P_{\Delta t}(\sigma, \sigma') &= Q_{\Delta t}^{\text{Strang}}(\sigma, \sigma') + \frac{1}{24} ([L_1, [L_1, L_2]] - 2[L_2, [L_2, L_1]]) \delta_{\sigma'}(\sigma) \Delta t^3 \\
&\quad + O(\Delta t^4).
\end{align}

From relations (2.12) and (2.13), we observe that the Strang splitting has a better local error compared to Lie ($\Delta t^3$ versus $\Delta t^2$). Therefore, if we prescribe an error tolerance, the Strang scheme will be able to accommodate a larger $\Delta t$ than the Lie scheme. With a larger $\Delta t$, we will be able to take larger steps with the same tolerance during the simulation, and this is especially important for parallel KMC, as we strive for balance between error accumulation and efficiency.

To be able to discuss more general operator splitting approximations to $P_{\Delta t}$, we introduce the following helpful lemma.

**Lemma 2.2** (local order of error and commutator). Let $P_{\Delta t}(\sigma, \sigma') = e^{t\Delta t} \delta_{\sigma'}(\sigma)$ and $Q_{\Delta t}(\sigma, \sigma')$ an approximation of $P_{\Delta t}$ via a splitting scheme. Then, there is a function $C : S \times S \to \mathbb{R}$ and an integer $p$, $p > 1$, such that

\begin{equation}
(2.14) \quad P_{\Delta t}(\sigma, \sigma') = Q_{\Delta t}(\sigma, \sigma') + C(\sigma, \sigma') \Delta t^p + o(\Delta t^p).
\end{equation}

We will refer to $C(\sigma, \sigma') = ([L_p - L_Q^p] \delta_{\sigma'}(\sigma)$ as the commutator and to $p$ as the order of the local error.

**Proof.** The result is immediate by using representations (2.5), (2.11), since for $\sigma, \sigma' \in S$,

\[ P_{\Delta t}(\sigma, \sigma') - Q_{\Delta t}(\sigma, \sigma') = \sum_{k=0}^{\infty} \frac{\Delta t^k}{k!} (L^k - L_Q^k) \delta_{\sigma'}(\sigma). \]

Then, $p$ is the smallest nonnegative integer such that $L^p \neq L_Q^p$. This of course implies that $L^k = L_Q^k$ for $k < p$. \qed

Equations (2.12) and (2.13) are examples of this lemma for the cases of the Lie and Strang splittings, respectively. Although in the case of Lie we were able to write the form of $L_Q^k$ explicitly for all $k$ (equation (2.10)), this is not a requirement and we only need to know $L_Q^p$ to compute the commutator and that object arises naturally when subtracting the two expansions, (2.5) and (2.11).

**Remark 2.3.** Relation (2.14) is central to the numerical analysis of splitting schemes, as it is the starting point to the derivation of upper bounds for the local and global error [2, 4, 5]. Even though our focus in this manuscript is on operator splitting schemes for parallel KMC, as long as an expression for the local error such as (2.14) exists, a similar analysis can be carried out for other types of schemes.

As we will see in the follow-up, the commutator has many desired properties. Since it is equal to $(L^p - L_Q^p) \delta_{\sigma'}(\sigma)$, and both $L^p[\delta_{\sigma'}](\sigma)$ and $L_Q^p[\delta_{\sigma'}](\sigma)$ depend on
the known transition rates $q$, the commutator is a computable object for every pair of states $(\sigma, \sigma')$. We will see in section 5.1 that for parallel KMC the work required in order to compute the commutator can scale appropriately with the system size.

2.2. PL-KMC and splitting schemes. We consider the case of PL-KMC as an application of the ideas in the previous section concerning approximations by semigroup splitting. Further discussion on the ideas of this section can be found in Arampatzis et al. [2, 4].

Our main motivating example for PL-KMC is an interacting particle system. Let $\Lambda \subset \mathbb{Z}^d$ be a square lattice with $N$ sites. At each site of it, $x \in \Lambda$, we define an order parameter $\sigma(x) \in \Sigma = \{0, 1, \ldots, K\}$. This parameter can be, for example, the species that occupies the lattice site $x$. For instance, in the Ising model, $\sigma(x) = 0$ would imply that the lattice site $x$ is empty and $\sigma(x) = 1$ that a particle occupies $x$. The CTMC of interest is $\{\sigma_t\}_{t \geq 0}$, with state space $S = \Sigma^\Lambda$. At every $t$, $\sigma_t$ represents a snapshot of the different occupancies of the lattice. We can describe the dynamics of such a system by looking at the individual spin changes at different lattice sites. Two more properties that are common among such systems and which we will also assume is that the transitions between states of $\sigma_t$ are localized and that they only involve a finite number of lattice sites per transition step. Localization implies that the probability that a certain transition will happen (the order parameter of a finite collection of lattice sites will change) only depends on the values of $\sigma$ on a neighborhood around those lattice sites. In other words, transitions depend on local (neighborhood) rather than global (whole lattice) information (see Figure 1).

We can formalize localization by looking at the implication for the transition rates of the process $\sigma_t$. Following the notation introduced in [2], let us assume that at time $t$, $\sigma_t = \sigma$. Now, we can express the transition rate for a jump to a new state $\sigma_{x,\omega}$ as

$$q(\sigma, \sigma_{x,\omega}) = q(x, \omega; \sigma),$$

where $x \in \Lambda$ and $\omega$ is an index of the set of all possible configurations, $S_x$, that correspond to an update at a lattice neighborhood $\Omega_x$ of the site $x$. When the only allowed transition is spin-flipping, that is, starting with $\sigma$, we can only go to states $\sigma'$ that differ in the order parameter of one lattice site $x$, we will write $\sigma'$ as $\sigma^x$ to denote the resulting state after the transition. It follows that for $\sigma_t$ we have an infinitesimal generator:

![Fig. 1. A checkerboard decomposition of a 2D lattice. Red sublattices correspond to group $G_1$ and white ones to $G_2$. For comparison, a nearest neighborhood region (n.n. region) is also shown (solid black cross). Transitions involving the center of that region only depend on the state of its nearest neighbors. So, if we pick the sublattices much larger than the size of an n.n. region, transitions in different sublattices belonging to the same group are independent. A site $x$ is said to belong to the boundary of its sublattice if part of its n.n. region is outside that sublattice (the green region is the collection of all such points for the first sublattice). If a transition occurs at such a site $x$, then an update needs to be made to the boundary information of all other sublattices for which $x$ belongs to an n.n. region.](image-url)
We can simulate the process $\sigma_t$ via standard KMC, as described in the beginning of section 2. Then the system would progress in time steps $t_n \sim \exp(\lambda(\sigma))$, where $\lambda(\sigma)$ is the total rate when the system is at state $\sigma$, as defined in (2.2). Since the total rate scales with the size of the lattice and the magnitude of the transition rates, a large or highly reactive model would be simulated slowly by classical KMC. The goal then, as realized in [2], is for a fixed $\Delta t > 0$ to design an approximation to the exact process $e^{\Delta t L}$ via a splitting method in such a way that allows for asynchronous computations.

To begin, we note that any decomposition of the lattice into nonoverlapping sublattices $\Lambda_i$ also induces a decomposition of the generator (2.16), that is,

$$L[f](\sigma) = \sum_{x \in \Lambda} \sum_{\omega \in S(x)} q(x, \omega; \sigma) \left( f(\sigma^{x, \omega}) - f(\sigma) \right).$$

Due to the localization of the system, we can decompose the lattice $\Lambda$ into $n$ sublattices, $\Lambda_i$, so that transitions in some sublattices are independent from transitions in others; see Figure 1. With two groups, $G_1 = \{\Lambda_i : i \text{ even}\}, G_2 = \{\Lambda_i : i \text{ odd}\}$, we can split $L$ into

$$L_j[f](\sigma) := \sum_{x \in G_j} \sum_{\omega \in S(x)} q(x, \omega; \sigma) \left( f(\sigma^{x, \omega}) - f(\sigma) \right), \quad j = 1, 2,$$

$$L[f](\sigma) = L_1[f](\sigma) + L_2[f](\sigma).$$

Thus, by the formulas in (2.18), we can use the ideas of the previous section to construct splitting approximations to $e^{L \Delta t}$. Those can also be interpreted as computation schedules for the parallel algorithm. Such schedules set two attributes of the simulation: (a) in what order to simulate the two groups asynchronously and (b) for how much time to simulate each group per time step (which the user controls with the $\Delta t$ parameter). A demonstration of how PL-KMC works is shown in Figure 2.

In general, the larger the $\Delta t$, the less different processes need to communicate to resolve inconsistencies during a run. This is a fact for any simulation algorithm
that can be expressed in the above operation splitting framework, e.g., SPPARKS and others [2]. Since communication is the usual bottleneck of PL-KMC algorithms, a practitioner would like to pick $\Delta t$ as large as possible, given a fixed tolerance. One of the important insights of the analysis in [4] is that the commutator controls this relationship. Simply put, a small $C(\cdot, \cdot)$ (as defined in Lemma 2.2) allows for a larger step size $\Delta t$.

3. Information metrics for comparing dynamics at long times. We will now introduce the main tools from information theory. In later sections, our focus will be to compare the exact process, $X_t$, and an approximation of it, $Y_t$, via their $\Delta t$-skeleton subprocesses. That is, given a fixed $\Delta t > 0$ and $M \in \mathbb{N}$, we look at the discrete-time Markov processes $X_{n\Delta t}$ and $Y_{n\Delta t}$ for $n \in \{0, \ldots, M\Delta t\}$, $T = M\Delta t$. For this reason, we now introduce those concepts for discrete-time processes.

Let us assume two discrete-time Markov processes $X_n$ and $Y_n$ on a countable state space $S$ with transition probabilities $P$ and $Q$, respectively. We also assume that for each process exists a corresponding unique stationary distribution $\mu_P$ and $\mu_Q$. Assuming $X_0 (Y_0)$ is distributed according to $\mu_P$ ($\mu_Q$), we can then calculate the probability of a specific path for each process. For example, if we fix a positive integer $M, T = M\Delta t$, and pick an $\vec{x} \in S^M$, then we have

$$P_{0:T}(\vec{x}) = P(X_T = x_M, \ldots, X_0 = x_0) = \mu_P(x_0)P(x_0, x_1)\cdots P(x_{M-1}, x_M).$$

Similarly, by changing $P$ to $Q$, we can calculate the path probability for $Y_n$.

Assuming one would prefer a path of length $T$ of the process $Y_n$ to infer results about a same length path of $X_n$, how much information about $X_n$ would be lost by such a method? This is a central question in coding theory and one way to quantify the information loss is through the idea of relative entropy,

$$R(Q_{0:T}|P_{0:T}) := \sum_{\vec{x} \in S^M} Q_{0:T}(\vec{x}) \log \frac{Q_{0:T}(\vec{x})}{P_{0:T}(\vec{x})}.$$  

Our definition here is with respect to the path measures $P_{0:T}, Q_{0:T}$, but we can apply the relative entropy to more general probability measures too. For this object to be properly defined, we need to have that $Q_{0:T}$ is absolutely continuous with respect to $P_{0:T}$, that is, $P_{0:T}(\vec{x}) = 0$ implies $Q_{0:T}(\vec{x}) = 0$. Other important properties of the RER are the following: (1) $R(Q_{0:T}|P_{0:T}) \geq 0$ for any $Q_{0:T}, P_{0:T}$ (Gibbs’ inequality); (2) $R(Q_{0:T}|P_{0:T}) = 0 \iff P_{0:T} = Q_{0:T}$. Note though that the relative entropy does not qualify as a metric in the classical sense, as it is not symmetric and does not satisfy the triangle inequality. It can, however, still be thought of as a distance between distributions and is useful as a building block for other information measures. For a more complete exposition on relative entropy and its properties, see Cover and Thomas [19].

Although the pathwise relative entropy is a suitable quantity to measure the similarity of the two path measures, it is computationally demanding to calculate, especially in the case of parallel KMC, where we do not have $Q_{0:T}$ and $P_{0:T}$ explicitly. For this reason, we look at a related object, the relative entropy per unit time, or RER. Given a probability measure $\nu_0$, $\nu_0(\vec{x}) = \nu_0(x_0), \vec{x} \in S^T$, the RER with respect to $\nu_0$ is defined as

$$H_{\nu_0}(Q|P) := \sum_{\vec{x} \in S^M} \nu_0(\vec{x})Q(x_0, x_1) \log \frac{Q(x_0, x_1)}{P(x_0, x_1)}.$$  

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.
would be observables. For instance, if that the RER is the right quantity to track can be given by considering time-averaged studying the pathwise relative entropy through the simpler RER.

In particular, when sampling from the stationary distribution corresponding to $Q$, that is, $\nu_0 = \mu_Q$, then $H_{\nu_i} = H_{\mu_Q} = H$ for all $i$. Then,

$$H(Q|P) = \sum_{x_0, x_1 \in S} \mu_Q(x_0)Q(x_0, x_1) \log \frac{Q(x_0, x_1)}{P(x_0, x_1)}.$$  

This also simplifies (3.3) to

$$R(Q_0:T|P_0:T) = M \cdot H(Q|P) + R(\mu_Q|\mu_P).$$

In (3.5), $R(\mu_Q|\mu_P)$ is the relative entropy of $\mu_Q$ with respect to $\mu_P$, capturing the loss of information between the exact and approximate stationary distribution. Note that $R(\mu_Q|\mu_P)$ does not depend on the length of the path. Instead, the term that quantifies the dependence on $T$ is $H(Q|P)$. Therefore, any difference between the two stationary measures becomes negligible for large times, which is a first advantage to studying the pathwise relative entropy through the simpler RER.

### 3.1. Information metrics and observables.

Further justification for the fact that the RER is the right quantity to track can be given by considering time-averaged observables. For instance, if $f$ is a function of the state space, then such an observable would be

$$M \cdot F_M(\{X_n : n = 0, \ldots, M - 1\}) = \sum_{k=0}^{M-1} f(X_k).$$

An important performance metric for the approximation is the weak error:

$$\|E_{P[0,T]}[F_M] - E_{Q[0,T]}[F_M]\|, T = M\Delta t.$$  

In recent work [15], uncertainty quantification bounds have been developed for the weak error that are of the form

$$\Xi_-(Q[0,T]|P[0,T]; M \cdot F_M)/M \leq \|E_{P[0,T]}[F_M] - E_{Q[0,T]}[F_M]\| \leq \Xi_+(Q[0,T]|P[0,T]; M \cdot F_M)/M.$$  

The quantities $\Xi_+(Q[0,T]|P[0,T]; M \cdot F_M)$ are defined as goal-oriented divergences [15], taking into account the observable $F$, and such that $\Xi_+(Q[0,T]|P[0,T]; M \cdot F_M) = 0$, if $Q[0,T] = P[0,T]$ or $f$ is deterministic. Note that the bound in (3.7) is robust (see Theorem 3.4 in [20], as well as [21]): if we consider a positive $\eta$ and all $Q_{\Delta t}$ such that $R(Q_{\Delta t}|P_{\Delta t}) < \eta$, then the upper bound in (3.7) is attained.

Dividing (3.7) by $M$ and letting $M$ go to infinity gives an inequality with respect to the stationary measures $\mu_Q, \mu_P$ of the scheme, $Q_{\Delta t}$, and the exact process, $P_{\Delta t}$, respectively:
\[\xi_-(Q_{\Delta t}\|P_{\Delta t}; f) \leq \mathbb{E}_{\mu_0}[f] - \mathbb{E}_{\mu_F}[f] \leq \xi_+(Q_{\Delta t}\|P_{\Delta t}; f),\]

where \(\xi_\pm(Q_{\Delta t}\|P_{\Delta t}; f) = \lim_{M \to \infty} \Xi_\pm(Q_{0:T}\|P_{0:T}; F)/M\). But \(\xi_\pm\) also admit a variational representation as

\[
\begin{align*}
\xi_+(Q_{\Delta t}\|P_{\Delta t}; f) &= \inf_{c \geq 0} \left\{ \frac{1}{c} [\lambda_{Q_{\Delta t}, P_{\Delta t}}(c) + H(Q_{\Delta t}\|P_{\Delta t})] \right\}, \\
\xi_-(Q_{\Delta t}\|P_{\Delta t}; f) &= \sup_{c \geq 0} \left\{ \frac{1}{c} [\lambda_{Q_{\Delta t}, P_{\Delta t}}(-c) + H(Q_{\Delta t}\|P_{\Delta t})] \right\},
\end{align*}
\]

(3.9)

with \(\lambda_{Q_{\Delta t}, P_{\Delta t}}(c)\) in (3.9) to be the logarithm of the maximum eigenvalue of the matrix with entries \(P_{\Delta t}(x, y) \exp(c \cdot (f(y) - \mathbb{E}_{\mu_F}[f]))\) (see [21] for details). Especially when \(H(Q_{\Delta t}\|P_{\Delta t})\) is small and through the asymptotic expansion of \(\xi_\pm\), an upper bound for the weak error at stationarity can be given (following the ideas in [15, 21]):

\[
|\mathbb{E}_{\mu_0}[f] - \mathbb{E}_{\mu_F}[f]| \leq \sqrt{2 H(Q_{\Delta t}\|P_{\Delta t})} + O(H(Q_{\Delta t}\|P_{\Delta t})),
\]

(3.10)

and

\[
v_{\mu_F}(f) = \sum_{k=-\infty}^{\infty} \mathbb{E}_{\mu_F}[f(X_k)f(X_0)].
\]

(3.11)

Inequality (3.10) connects the long-time loss of accuracy that the weak error captures with the RER and \(v_{\mu_F}(f)\), which is the integrated auto-correlation function for the observable \(f\) and a quantity we can estimate during the simulation. As a consequence of (3.10), any further results on the asymptotic behavior of \(H(Q_{\Delta t}\|P_{\Delta t})\) with respect to \(\Delta t\) can be simply translated to the weak error point of view.

4. Long-time error behavior of splitting schemes. In this section, we compare the RER between two different processes. One of them will always be the \(\Delta t\)-skeleton process derived from the CTMC we wish to simulate, with transition probability

\[
P_{\Delta t}(\sigma, \sigma') = e^{\Delta t L} \delta_\sigma(\sigma).
\]

(4.1)

This exact \(\Delta t\)-process will be compared with the \(\Delta t\)-skeleton process derived from an operator splitting of (4.1). Such approximations will be denoted with \(Q_{\Delta t}\). We note here that the discretization (4.1) of the original Markov process with semigroup \(e^{\Delta t L}\) with respect to \(\Delta t\) is carried out only as a means to compare the original process with the approximations \(Q_{\Delta t}\). The transition kernel \(P_{\Delta t}\) is just a particular instance of the transition matrix of the continuous Markov process with semigroup \(P_t = e^{L t}\), so there is no approximation error in (4.1). In fact, using the \(\Delta t\)-skeleton corresponds to subsampling from the CTMC at every \(\Delta t\).

Our goal is to show the dependence of the RER to various quantities of interest that are usually computed for short-time error analysis. We will see that the commutator, the order of the local error, and other quantities make an appearance in the asymptotic results we develop. We limit our discussion to the case that \(\Delta t\) is in \((0, 1]\), as this is the interval where splitting schemes are most accurate. We also assume throughout this section that \(L\) is a bounded operator. We will often refer to the splittings previously discussed, Lie and Strang, which define discrete processes with transition probabilities

\[
\begin{align*}
Q^{\text{Lie}}_{\Delta t}(\sigma, \sigma') &= e^{\Delta t L_1} e^{L_2 \Delta t} \delta_\sigma'(\sigma), \\
Q^{\text{Strang}}_{\Delta t}(\sigma, \sigma') &= e^{\Delta t L_1/2} e^{L_2 \Delta t} e^{L_1 \Delta t/2} \delta_\sigma'(\sigma).
\end{align*}
\]

(4.2)
Here \( L \) is the original generator and \( L = L_1 + L_2 \) with \( L_1, L_2 \) assumed bounded as operators. For instance, in the case of parallel KMC, \( L_1, L_2 \) will be imposed by the domain decomposition of the lattice; see Figure 1.

Before we move on to the analysis, we need to address a last issue. As mentioned before, our main tool will be asymptotic expansions of the RER with respect to \( \Delta t \). We will then use those to do comparisons for different \( \Delta t \).

We will then use those to do comparisons for different \( \Delta t \) before, our main tool will be asymptotic expansions of the RER with respect to \( \Delta t \). It is important to first account for the scaling of RER with respect to that parameter. The situation can be best illustrated by the worst-case scenario, when the order of the local error between two Markov semigroups, \( Q_{\Delta t}^A, Q_{\Delta t}^B \), is equal to one.

**Lemma 4.1.** Let \( L_A, L_B \) be bounded generators of Markov processes, \( L_A \neq L_B \), with corresponding transition probabilities \( Q_{\Delta t}^A, Q_{\Delta t}^B \). Then,

\[
H(Q_{\Delta t}^B | Q_{\Delta t}^A) = O(\Delta t).
\]

**Proof.** The proof follows the ideas in Theorem 5.2. The argument is provided in the supplementary material (104727SupMat.pdf [local/web 126KB]).

**Remark 4.2.** Using Lemma 4.1, we can readily see that given an operator splitting scheme \( Q_{\Delta t} \) that approximates the exact \( P_{\Delta t} \), we expect a scaling at least of the type \( H(Q_{\Delta t} | P_{\Delta t}) = O(\Delta t) \). To correct for the \( \Delta t \) scaling, we will instead work with a \( \Delta t \)-normalized RER. That is, we redefine the RER as

\[
H(Q_{\Delta t} | P_{\Delta t}) := \frac{1}{\Delta t} \sum_{\sigma, \sigma'} \mu_Q(\sigma) Q_{\Delta t}(\sigma, \sigma') \log \left( \frac{Q_{\Delta t}(\sigma, \sigma')}{P_{\Delta t}(\sigma, \sigma')} \right).
\]

We wish to use the RER (equation (4.3)) to study the long-time loss of information between \( Q_{\Delta t} \) and \( P_{\Delta t} \). However, in the case of parallel KMC, those are difficult to calculate explicitly, hence we turn to asymptotic expansions instead. We will see that the terms in those expansions depend on the transition rates and, under suitable ergodic assumptions, can be estimated during the simulation.

**5. RER analysis for parallel KMC.** We will now study an example from a class of interacting particle systems, limiting our discussion to the Lie and Strang splittings. Given two states \( \sigma, \sigma' \in S \) and \( x \) lattice site, \( \sigma(x) \in \{0, 1\} \), we have that the transition rates \( q \) are

\[
q(\sigma, \sigma') = \begin{cases} q(\sigma, \sigma^+) > 0, & \sigma' = \sigma^+, \\ 0, & \text{else.} \end{cases}
\]

The rates in (5.1) provide a particular example of an adsorption/desorption system. Other mechanisms can be incorporated into (5.1), such as diffusion or reactions with multiple components or with particles that have many degrees of freedom [2].

Given a lattice \( \Lambda \) with \( N \) sites, we are interested in simulating the process \( \sigma_t = \{\sigma_t(x) : x \in \Lambda\} \) in parallel with an operator splitting method, so we apply the ideas in section 2.2 to that end. We first decompose the lattice into nonoverlapping sublattices (see Figure 1) and this induces a decomposition of the generator into new generators \( L_1, L_2 \) as in (2.18). Then, for any \( T > 0 \), the adsorption/desorption system can be simulated in \([0, T]\) using the parallel KMC algorithm. From the short-time error analysis, we can control the error by computing the commutator, \( C(\cdot, \cdot) \), and the order of the local error that corresponds to the operator splitting scheme we use. For example, we know that for the Lie splitting that order is \( p = 2 \) and
We are now able to write an asymptotic result for RER for the Lie and Strang operator

\[ L_x \] 

the definition of \( \tanh \) considers a \( \Delta \)

rates. First, it is computable for any pair \((x, y) \in \Lambda \times S\) as it only depends on the transition rates \( q \). We recall here that \( \sigma^{x,y} \) stands for the resulting state \( \sigma' \) after a spin-flip of an initial state \( \sigma \) at lattice sites \( x, y, x \neq y \). A full description of the above formula along with a proof can be found in the supplementary material (104727SupMat.pdf [local/web 126KB]).

**Remark 5.1.** Formula (5.2) for the Lie commutator has two important properties. First, it is computable for any pair \((\sigma, \sigma') \in S \times S\) as it only depends on the transition rates \( q \). Second, it is surely equal to zero if \( \sigma' \neq \sigma^{x,y} \) and \( \sigma' \neq \sigma^x \) for all \( x, y \in \Lambda, x \neq y \), due to the \( \delta_{\sigma'} \) appearing in the different sums. We will also see that the sum in (5.2) needs to be evaluated only for the neighboring lattice sites \( x, y \) that are not both in the same group. For instance, in Figure 1, we would only need to evaluate the sum over the green boundary regions of every sublattice, which makes the computation of the commutator much simpler (see Remark 6.1 for a complexity analysis). Those properties hold for commutators of other operator splitting schemes too; see [4] and section 8.

To study the asymptotic behavior of the RER, we will need to quantify the dependence of various combinations of \( P_{\Delta t} \) and \( Q_{\Delta t} \) to \( \Delta t \). To this end, we use the following facts, both of which stem from Lemma 2.2:

\[
\begin{align*}
P_{\Delta t}(\sigma, \sigma') - Q_{\Delta t}(\sigma, \sigma') &= C(\sigma, \sigma') \Delta t^p + o(\Delta t^p), \\
P_{\Delta t}(\sigma, \sigma') + Q_{\Delta t}(\sigma, \sigma') &= 2\delta_{\sigma'}(\sigma) + 2q(\sigma, \sigma') \Delta t + o(\Delta t) \\
&= 2Q_{\Delta t}(\sigma, \sigma') + C(\sigma, \sigma') \Delta t^p + o(\Delta t^p).
\end{align*}
\]

We are now able to write an asymptotic result for RER for the Lie and Strang operator splittings in parallel KMC under the assumption in relation (5.1).

**Theorem 5.2.** Let \( \Delta t \in (0, 1) \) and \( \sigma, \Delta t \) on the lattice \( \Lambda \) with transition probability \( P_{\Delta t}(\sigma, \sigma') = e^{L_{\Delta t}} \delta_{\sigma'}(\sigma) \) for \( \sigma, \sigma' \in S \). Then, let \( L_1, L_2 \) be a splitting of \( L \) based on a decomposition of the lattice \( \Lambda \). Assuming that property (5.1) holds for the rates, if there exists a state \( \sigma \in S \) and lattice sites distinct \( x, y \) such that the Lie commutator \( C(\sigma, \sigma^{x,y}) \neq 0 \), we have that

\[
H(Q_{\Delta t}^{\text{Lie}}|P_{\Delta t}) = O(\Delta t^2) \quad \text{(Lie)}.
\]

Similarly, if there exists a state \( \sigma \in S \) and distinct lattice sites \( x, y, z \) such that \( C(\sigma, \sigma^{x,y}) \neq 0 \),

\[
H(Q_{\Delta t}^{\text{Strang}}|P_{\Delta t}) = O(\Delta t^2) \quad \text{(Strang)}.
\]

**Proof.** We will first show the result for the Lie case and then note the differences in the proof for the Strang case. Thus, we denote \( Q_{\Delta t}^{\text{Lie}} \) by \( Q_{\Delta t} \) and \( \mu_{\text{Lie}} \) by \( \mu_{Q} \) and consider a \( \Delta t \in (0, 1) \). As we wish to construct an asymptotic expansion for the RER (equation ((4.3))), we first need to expand the logarithm. Given a positive \( x \) and by the definition of \( \tanh^{-1} \),
\[ \log(x) = 2 \tanh \left( \frac{x - 1}{x + 1} \right) = 2 \sum_{k=0}^{\infty} \frac{1}{2k+1} \left( \frac{x - 1}{x + 1} \right)^{2k+1}. \]

This expansion of the logarithm converges for every \( x > 0 \), as can be seen by applying the root convergence test. Thus, expanding the logarithm part of the RER, we get

\[ \Delta t \cdot H(Q_{\Delta t}|P_{\Delta t}) = -2 \sum_{\sigma,\sigma'} \mu_Q(\sigma)Q_{\Delta t}(\sigma,\sigma') \frac{P_{\Delta t}(\sigma,\sigma') - Q_{\Delta t}(\sigma,\sigma')}{Q_{\Delta t}(\sigma,\sigma') + P_{\Delta t}(\sigma,\sigma')} + 2 \sum_{\sigma,\sigma'} \mu_Q(\sigma)J(\Delta t; \sigma,\sigma'). \]

We will study the asymptotic behavior of both parts of the RER in (5.9). First, applying (5.11) to the denominator of the fraction in (5.9) and carrying out the simplifications, we have

\[ \Delta t \cdot H(Q_{\Delta t}|P_{\Delta t}) = -2 \sum_{\sigma,\sigma'} \mu_Q(\sigma) (P_{\Delta t}(\sigma,\sigma') - Q_{\Delta t}(\sigma,\sigma') + G(\Delta t; \sigma,\sigma')) + 2 \sum_{\sigma,\sigma'} \mu_Q(\sigma)J(\Delta t; \sigma,\sigma'). \]

Now, since \( Q_{\Delta t}, P_{\Delta t} \) are transition probabilities, \( \sum_{\sigma' \in S} P_{\Delta t}(\sigma,\sigma') - Q_{\Delta t}(\sigma,\sigma') = 0 \) for all \( \sigma \in S \), and thus the corresponding part of (5.11) is zero. To progress, we need to study the dependence on \( \Delta t \) of \( J, G \). First, for \( G \) in (5.11),

\[ G(\Delta t; \sigma,\sigma') = \frac{(P_{\Delta t}(\sigma,\sigma') - Q_{\Delta t}(\sigma,\sigma'))C(\sigma,\sigma')\Delta t^2}{(2Q_{\Delta t}(\sigma,\sigma') + \Delta t^2C(\sigma,\sigma') + o(\Delta t^2))} + o(\Delta t^2). \]

To expose the dependence of the numerator of (5.12) to \( \Delta t \), we use (5.3) to get

\[ G(\Delta t; \sigma,\sigma') = \frac{(C(\sigma,\sigma'))^2}{2Q_{\Delta t}(\sigma,\sigma') + \Delta t^2C(\sigma,\sigma') + o(\Delta t^2)} \Delta t^4 + o(\Delta t^2). \]

We wish to show that \( G(\Delta t; \sigma,\sigma') = O(\Delta t^2) \). From the explicit form of the commutator in (5.2) and Remark 5.1, we can see that we need to study \( G \) only in the cases that \( \sigma' = \sigma^x \) or \( \sigma' = \sigma^{x,y} \), given a state \( \sigma \) and lattice sites \( x, y \), since otherwise \( C(\sigma,\sigma') = 0 \). Let us consider \( \sigma' = \sigma^{x,y} \). Since the order of the local error is equal to two, from expansion (2.11) and the fact that \( L_Q[\delta_{\sigma^{x,y}}](\sigma) = L[\delta_{\sigma^{x,y}}](\sigma) \) and \( L[\delta_{\sigma^{x,y}}] = q(\sigma,\sigma^{x,y}) = 0 \) (see the property in (5.1)), we have

\[ Q_{\Delta t}(\sigma,\sigma^{x,y}) = \frac{\Delta t^2}{2} L_Q^2[\delta_{\sigma}](\sigma) + o(\Delta t^2). \]

Thus, applying (5.14) to the denominator of (5.13),

\[ G(\Delta t; \sigma,\sigma^{x,y}) = \frac{(C(\sigma,\sigma^{x,y}))^2}{\Delta t^2 \cdot (L_Q^2[\delta_{\sigma^{x,y}}](\sigma) + C(\sigma,\sigma^{x,y})) + o(\Delta t^2)} \Delta t^4 + o(\Delta t^2) \]

\[ = \frac{(C(\sigma,\sigma^{x,y}))^2}{L_Q^2[\delta_{\sigma^{x,y}}](\sigma) + C(\sigma,\sigma^{x,y})} \Delta t^2 + o(\Delta t^2). \]
By similar calculations, we can show that \( G(\sigma, \sigma^x) = O(\Delta t^3) \), if \( C(\sigma, \sigma^x) \neq 0 \) for that \( x \in \Lambda \). Regardless, this would be a lower order, since \( \Delta t < 1 \). Thus, \( G(\Delta t; \sigma, \sigma') \) is indeed of order \( \Delta t^2 \). Next, we will account for \( J(\Delta t; \sigma, \sigma') \). If \( \sigma' = \sigma^{x,y} \), then

\[
(5.16) \quad J(\Delta t; \sigma, \sigma^{x,y}) = Q_{\Delta t}(\sigma, \sigma^{x,y}) \sum_{k=1}^{\infty} \frac{1}{2k + 1} \left( \frac{Q_{\Delta t}(\sigma, \sigma^{x,y}) - P_{\Delta t}(\sigma, \sigma^{x,y})}{Q_{\Delta t}(\sigma, \sigma^{x,y}) + P_{\Delta t}(\sigma, \sigma^{x,y})} \right)^{2k+1}.
\]

Because \( Q_{\Delta t}(\sigma, \sigma^{x,y}) = O(\Delta t^2) \) and \( Q_{\Delta t}(\sigma, \sigma^{x,y}) \pm P_{\Delta t}(\sigma, \sigma^{x,y}) = O(\Delta t^2) \), we get

\[
J(\Delta t; \sigma, \sigma^{x,y}) = O(\Delta t^2),
\]

since, for \( \sigma' = \sigma^x \), \( J(\Delta t; \sigma, \sigma^x) = O(\Delta t^4) \) and this is a lower order when \( \Delta t < 1 \). Therefore, \( H(Q_{\Delta t} | P_{\Delta t}) = O(\Delta t^4) \). Note that all of the terms of the series in (5.16) contribute a term of order \( \Delta t^2 \), so the coefficient of \( \Delta t^2 \) in the asymptotic expansion of the RER will be a result of the summation of all those terms.

Finally, we discuss the differences in our argument for the proof of the Strang case. First, the order of the local error for Strang is \( p = 3 \), so every time we use formula (5.3) in the proof, we would introduce a term of order \( \Delta t^3 \) instead of \( \Delta t^2 \). Then, using an expression for \( C(\cdot, \cdot) \) similar to (5.2) but for the Strang case, we would show that

\[
J(\Delta t; \sigma, \sigma^{x,y;z}) = O(\Delta t^3) = G(\Delta t; \sigma, \sigma^{x,y;z})
\]

for \( x, y, z \in \Lambda \) and \( x \neq y \neq z \). This would then give the result for Strang.

**5.1. Building biased a posteriori estimators for the RER.** Theorem 5.2 shows that the long-time accuracy with respect to the RER of the two operator splitting schemes, Lie and Strang, scales with \( \Delta t \) in the same way the global error does. However, it also exposes the first terms in the asymptotic expansion of the RER for Lie and Strang. Essentially,

\[
(5.17) \quad H(Q_{\Delta t}^{\text{Lie}} | P_{\Delta t}) = A \Delta t + o(\Delta t),
\]

\[
(5.18) \quad H(Q_{\Delta t}^{\text{Strang}} | P_{\Delta t}) = B \Delta t^2 + o(\Delta t^2),
\]

where \( A, B \) are the corresponding highest-order RER coefficients. Those have an explicit form that depends on the system one wishes to simulate and the commutator \( C(\sigma, \sigma') \) corresponding to the scheme. We focus on the case of the Lie splitting, though similar comments can also be made for Strang. For systems with transition rates satisfying the property in (5.1), the highest-order coefficient \( A \) appearing in (5.17) has the form

\[
(5.19) \quad A = \sum_{\sigma} \mu_{\text{Lie}}(\sigma) \sum_{x,y \in \Lambda} C_{\text{Lie}}(\sigma, x^{x,y}) F_{\text{Lie}}(\sigma, x^{x,y}),
\]

where \( C_{\text{Lie}} \) is the Lie commutator (see (2.12)) and \( F_{\text{Lie}} \) is a quantity that depends on the splitting (see (A.1) and (A.3) in the appendix for examples on how this \( F \) can look for different splittings). Both \( C \) and \( F \) can be expressed in terms of the transition rates of the process \( q \), i.e., they are computable for any state \( \sigma \) and \( x, y \in \Lambda \). Therefore, \( A \) in (5.19) can be estimated via an ergodic average when simulating with the Lie scheme and hence, for small \( \Delta t \), \( H(Q_{\Delta t}^{\text{Lie}} | P_{\Delta t}) \approx A \Delta t \).

At first glance, computing coefficient (5.19) involves work that scales with the size of the lattice. However, it was shown in Lemma 5.15 of [4] that the commutator only
depends on the boundary regions between sublattices (see Figure 1). We will continue this discussion in section 6, where we consider an adsorption-desorption system. We will also see that, apart from a comparison of the schemes in terms of the long-time loss of information, the estimators of RER can also be of use in tuning parameters of the scheme ($\Delta t$, domain decomposition, etc.). We will then consider the behavior of the RER when simulating other systems in section 8.

6. Error versus communication and time-step selection. In this section, we explore the balance between numerical error and processor communication in parallel KMC, in the context of a specific example. Let us assume a bounded 2D lattice, $\Lambda \subset \mathbb{Z}^2$ with $100 \times 100$ sites. At each site $x$, we have a spin variable, $\sigma(x) \in \Sigma = \{0, 1\}$, with $\sigma(x) = 0$ denoting an empty site and $\sigma(x) = 1$ an occupied one. Our model in this case is going to be an adsorption-desorption one, although the analysis would similarly apply for other mechanisms (diffusions, reactions, etc.; see [2] for more details).

The transition rates we will use correspond to spin-flip Arrhenius dynamics. Given a lattice site $x$, we may also define the nearest-neighbor set $\Omega_x = \{z \in \Lambda : |z - x| = 1\}$. The transitions rates are then

\begin{equation}
q(\sigma, \sigma^e) = q(x, \sigma) = c_1(1 - \sigma(x)) + c_2\sigma(x)e^{-\beta U(x)},
\end{equation}

\begin{equation}
U(x) = J_0 \sum_{y \in \Omega_x} \sigma(y) + h,
\end{equation}

where $c_1, c_2, -\beta, J_0$, and $h$ are constants that can be tuned to generate different dynamics. We recall that $\sigma^e$ denotes the result of a spin-flip at lattice position $x$ if we start from state $\sigma$. Note that the transition rates (6.1) have the property (5.1). When considering a jump from $\sigma$ to $\sigma^e$, $q$ only depends in the spin values of the sites close to $x$ (through $U(x)$). Since transitions are localized, we can thus employ a geometrical decomposition of the lattice, as described in section 2.1, and simulate the system in parallel. To accomplish this, we used Sandia Labs’ SPPARKS code, a kinetic Monte Carlo simulator [3].

From Table 1 and Remark 6.1, we can see that the cost of computing quantities that depend on the commutator scales as $O(N)$ for an $N \times N$ lattice. As the highest-order coefficients of the RER also depend on the commutator (see section 5.1), those also scale as $O(N)$. We can take advantage of the knowledge of the scaling by defining a per-particle RER (pp-RER). That is,

\begin{equation}
H_{pp}(Q_{\Delta t}|P_{\Delta t}) := \frac{1}{N}H(Q_{\Delta t}|P_{\Delta t}).
\end{equation}

This way, setting a tolerance for the pp-RER will have the same meaning across different system sizes. We confirmed that $O(N)$ is the right scaling of the pp-RER with respect to system size via simulation, as we saw that for increasing $N$, $H_{pp}(Q_{\Delta t}|P_{\Delta t}) \sim o(1)$.

To estimate the top-order coefficients of the pp-RER expansion, we simulated the system until convergence to the stationary distribution was established. After that, every sample simulated by SPPARKS [3] was used to calculate the estimates. Note that, in this case, we show an overestimate of $B$, so results for the Strang splitting will be even better than the ones presented in Figure 3. It is possible to get an estimator that converges to the exact value of $B$ by adding all of the positive terms in $L_s^3[\delta_\sigma](\sigma)$ to the denominator of (A.5). Figure 3 illustrates the difference in long-time accuracy between the two splittings. Since this is a logarithmic plot, most of the difference is made by Strang having a different order than Lie.
Fig. 3. Logarithmic scale: Comparison between $\Delta t$ and the estimate of the pp-RER for Lie and Strang. Estimates for the constants $A, B$ come from the simulation of a 2D Ising model on a $100 \times 100$ lattice with final time $T = 1000$. Simulation was done in parallel with SPPARKS.

### Table 1

Upper bounds (normalized by lattice size) on the number of lattice sites we need to evaluate the transition rates at in order to calculate the commutator for each operator splitting, assuming that a checkerboard decomposition into $m^2$ sublattices of an $N \times N$ lattice is used, as in Figure 1. The commutator also encodes the cost of communication between the processes. As $N$ grows, the cost of communication is smaller, as the processes spend more time simulating on the sublattices than updating each others’ boundaries.

<table>
<thead>
<tr>
<th>Upper bound of the commutator cost (normalized by number of sites, $N^2$)</th>
<th>Lie</th>
<th>Strang</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2(m + 1)/N$</td>
<td>$6(m + 1)/N$</td>
<td></td>
</tr>
</tbody>
</table>

Remark 6.1 (on the efficiency of computing the highest-order coefficients of the expansion of the RER for the Lie and Strang operator splittings.). In the case of a checkerboard decomposition of the lattice (see Figure 1), we can calculate in exactly how many sites we need to evaluate the rates in order to calculate the commutator. However, for our purposes, upper bounds will be more appropriate. Table 1 offers a comparison of those bounds when we decompose an $N \times N$ lattice into $m^2$ sublattices, assuming nearest neighbor interactions. Notice that the cost is larger for Strang due to the complexity of the corresponding commutator.

On a more practical note, a user of a splitting scheme may instead like to see the flipped relationship. That is, given a fixed tolerance, what is the maximum time window during which the simulation can run asynchronously? If we interpret tolerance as a fixed value of $H_{pp}(Q_{\Delta t}|P_{\Delta t})$ during the simulation, then the relationship with $\Delta t$ is the one in Figure 4. There we can see that if our error tolerance with respect to the pp-RER is $10^{-3}$, then any $\Delta t$ smaller than 0.7 works for the Strang splitting. To get within the same tolerance with Lie, $\Delta t$ has to be less than 0.02, a substantially smaller step-size for parallel computations. As is expected, a smaller step-size comes with larger communication cost and thus a longer computation for the same tolerance. This can be seen in Figure 5.

Remark 6.2. Figures 4 and 5 illustrate the very practical consequences of the theory. Interest in highly accurate splitting schemes in PL-KMC stems from a tolerance-
versus-communication point of view. A user of such a scheme would like for it to be as accurate as possible; therefore the step size, \( \Delta t \), should be relatively small. However, for the scheme to be efficient, \( \Delta t \) should be large enough for every processor to have a substantial amount of work to do before communications are in order. A good balance can be reached in between and a scheme that is more accurate allows for a larger \( \Delta t \) while holding the same error tolerance. Given that the RER captures long-time behavior, this is an important comparison between the schemes.

### 6.1. The pp-RER as an efficient diagnostic quantity for parallel KMC.

The discussion above about the pp-RER, (6.3), suggests the use of these estimates as efficient diagnostic quantities for comparing schemes. As discussed in the previous section, we can infer the scaling of the top-order coefficient of the RER by the properties of the commutator. Consequently, we can “normalize” the RER (as in (6.3)) by that scaling to derive a similarity measure that does not depend on system size. This
is significant as it allows practitioners to compare schemes and tune parameters ($\Delta t$, domain decomposition, etc.) on a system of smaller size and thus avoid further slowing down of the target simulation, which is crucial for complicated systems. Overall, our approach can be viewed as a diagnostic tool that allows us to compare different parallelization schemes based on operator splitting.

7. Some connections with model selection and information criteria. The interacting particle system application considered in section 6 allows us to look at the RER via a statistical lens. The goal is to compare two models, $Q^1_{\Delta t}, Q^2_{\Delta t}$, of the actual distribution $P_{\Delta t}$ by utilizing simulated data. From this standpoint, our methodology is nothing more than model selection. There is an abundance of literature toward tackling the comparison of different models, given a sufficiently large amount of data. A prominent example is the use of information criteria in the model selection literature, like Akaike [22] and Bayesian [23]. Those provide estimates for the information lost compared to a given data set by using one approximate model instead of another, without requiring knowledge of the true model.

The approach in this work is very similar in nature. As stated before, motivated by Theorem 5.2, we can express the RER in each case as

$$H(Q^i_{\Delta t}|P_{\Delta t}) = A_i \Delta t^{p_i} + o(\Delta t^{p_i}), p_i \geq 1, i \in \{1, 2\}.$$

For instance, in the case of the Lie splitting, $A_1 = A$ as defined in (A.1), $p_1 = 2$, and for Strang $A_2 = B, p_2 = 3$, as defined in (A.3). Given simulated data and for a small fixed $\Delta t$, we can estimate the coefficients $A_i$. Comparison of the schemes can now be done through

$$H(Q^1_{\Delta t}|P_{\Delta t}) - H(Q^2_{\Delta t}|P_{\Delta t}) = A_1 \Delta t^{p_1} - A_2 \Delta t^{p_2} + o\left(\Delta t^{\min(p_1, p_2)}\right).$$

(7.1)

The difference $A_1 \Delta t^{p_1} - A_2 \Delta t^{p_2}$ shares the properties of the information criteria previously mentioned while also introducing some new ones:

1. It is a computationally tractable quantity.
2. It compares the schemes in terms of long-time information loss (through $p_1, p_2$).
3. It takes into account communication cost of each scheme (through $A_1, A_2$ and associated commutators).

Thus, as an information criterion, RER differences like in (7.1) offer a different perspective through which to pick a splitting scheme over another. A new element in our approach, compared to the earlier vast literature on information criteria, is the use of RER instead of the standard relative entropy. Using RER allows us to compare stochastic dynamics models and in a data context, correlated time series.

8. Generalizations, connectivity, and relative entropy rate. Up to this point, we have analyzed the RER with respect to the leading order in $\Delta t$ for the case of a stochastic particle system (see Theorem 5.2). In this section, we study the RER in a more general setting and illustrate that it captures more details about the system and the scheme used than one would expect. We will also see how the order of the RER can change depending on those details, resulting in some cases in schemes of higher accuracy.

Definition 8.1 (restriction of a generator). Let us have a set $A$ with $A \subset S \times S$ and $L$ be an infinitesimal generator of a Markov process with associated transition...
In the case that the two states are disconnected, i.e.,

\begin{equation}
\tag{8.2}
\end{equation}

between the two states is defined as \( CTMP \) over a countable state space \( S \).

Moreover, since it does not have to be symmetric, that is, \( d \) is not a metric in the classical sense, it does not have to be symmetric, that is, \( d(\sigma, \sigma') \neq d(\sigma', \sigma) \) in general. However, it satisfies the triangle inequality. In addition, the distances depend only on the transition rates, i.e., they are time independent. We will refer to those distances \( L \) and we will use \( Path(\sigma \rightarrow \sigma') \) for the set of all paths from \( \sigma \) to \( \sigma' \). Thus, we are now able to define a distance between states by looking at the length of the shortest path that connects them.

**Definition 8.2 (distance between states).** Let \( q \) be the transition rates of a CTMC on a countable state space \( S \). Then, let \( \sigma, \sigma' \in S, \sigma \neq \sigma' \). The distance \( d_q \) between the two states is defined as

\begin{equation}
\tag{8.2}
\end{equation}

In the case that the two states are disconnected, i.e., \( Path(\sigma \rightarrow \sigma') = \emptyset \), then \( d(\sigma, \sigma') = +\infty \). Given those distances, one can also define the diameter of the space as

\[ \text{diam}(S) = \max_{(\sigma, \sigma') \in S \times S} \{ d(\sigma, \sigma') \}. \]

This notion of distance comes from graph theory and is known as the geodesic distance. When there is no ambiguity concerning the transition rates used, we will drop the \( q \) from the notation, using \( d \) instead of \( d_q \). \( d \) is not a metric in the classical sense, since it does not have to be symmetric, that is, \( d(\sigma, \sigma') \neq d(\sigma', \sigma) \) in general. However, it satisfies the triangle inequality. In addition, the distances depend only on the transition rates, i.e., they are time independent. We will refer to those distances as the connectivity of the state space for the Markov chain with transition rates \( q \).

The importance of using such a distance can be seen in the following result concerning compositions of the infinitesimal generator \( L \).

**Lemma 8.3.** Let \( L \) be an infinitesimal generator of a Markov process, with corresponding transition rates \( q \), and let \( \sigma' \) be some state of the process. Then,

\[ \{ \sigma : L^n[\delta_{\sigma'}](\sigma) \neq 0 \} \subseteq \{ \sigma : d(\sigma, \sigma') \leq n \} = B_n(\sigma'). \]

**Proof.** The proof is by induction. The argument can be found in supplementary material (10472SupMat.pdf [local/web 126KB]).

In other words, for a fixed state \( \sigma' \), if \( d(\sigma, \sigma') > n \), then \( L^n[\delta_{\sigma'}](\sigma) = 0 \). The set \( B_n(\sigma') \) contains all states that are connected with \( \sigma' \) with \( n - 2 \) or less in between states. We will also use the notation \( S_n(\sigma') := \{ \sigma : d(\sigma, \sigma') = n \} \).
Since our primary interest is in studying approximations based on splitting our generator \( L \) to \( L_1, L_2 \), it makes sense to have an extension of the previous result to compositions of \( L_1, L_2 \). The following lemma is the generalization of Lemma 8.3 to compositions of restrictions. We will use the notation \( L^k|_A \) to denote the \( k \)th composition of generator \( L \), where, instead of the original transition rates, we use \( q_A \).

**Lemma 8.4.** Let us have the state space \( S \) and \( S \times S = A \cup B, A \cap B = \emptyset \), along with generators \( L_1 = L|_A, L_2 = L|_B \). We fix \( \sigma' \in S \) and \( k, m \in \mathbb{N} \). Then,

\[
\{ \sigma : L^k [L^m \delta_{\sigma'}] (\sigma) \neq 0 \} \subseteq \{ \sigma : d(\sigma, \sigma') \leq k + m \}.
\]

**Proof.** The proof is an induction argument similar to that of Lemma 8.3; see supplementary materials (104727SupMat.pdf [local/web 126KB]). \(\square\)

Lemma 8.4 can be simply extended to more complicated compositions by the use of similar arguments. Thus, if every composition of \( L_1, L_2 \) is controlled in the sense of Lemma 8.4, then it is not difficult to see that the same control holds for collections of them of the same order, i.e., if we fix \( \sigma' \in S \) and \( k \in \mathbb{N} \),

\[
(8.3) \quad \{ \sigma : L^k \delta_{\sigma'}(\sigma) \} \subseteq \{ \sigma : d(\sigma, \sigma') < k \}.
\]

We can use restrictions of generators as building blocks for splitting schemes. A point often made in this work is the importance of the commutator in studying those schemes. Thus, it makes sense to have a relation between connectivity and the commutator.

**Lemma 8.5 (support of the commutator).** Let \( L \) be the generator of a Markov process and \( L_1, L_2 \) restrictions of that generator. Let also \( \Delta t > 0 \). Then, assume \( Q_{\Delta t} \) is an approximation of \( P_{\Delta t} \) by using a splitting scheme of order \( p \) with associated commutator \( C \). Then, for fixed \( \sigma' \in S \),

\[
\{ \sigma : C(\sigma, \sigma') \neq 0 \} \subseteq \{ \sigma : d(\sigma, \sigma') \leq p \}.
\]

**Proof.** In Lemma 2.2, we defined the commutator as \( C(\sigma, \sigma') = (L^p - L^p_2) \delta_{\sigma'}(\sigma) \). From Lemma 8.3, we have that if \( d(\sigma, \sigma') > p \), then \( L^p \delta_{\sigma'}(\sigma) = 0 \) and from (8.3), \( L^p_2 \delta_{\sigma'}(\sigma) = 0 \). This gives the result. \(\square\)

When the state space is finite, as in the case of stochastic particle systems on finite lattices, then the commutator \( C \) is a matrix indexed by the different states. An implication of Lemma 8.5 is that there is a reordering of the rows/columns that turns \( C \) into a banded matrix. Regardless, we can now prove a general result for the asymptotics of the RER.

**Theorem 8.6.** Consider \( \Delta t \in (0, 1) \) and let \( P_{\Delta t}(\sigma, \sigma') = e^{\Delta t \delta_{\sigma'}(\sigma)} \), \( Q_{\Delta t}(\sigma, \sigma') \) be an approximation of \( P_{\Delta t} \) based on a splitting scheme with \( L_1, L_2 \) restrictions of the generator \( L \) and \( \mu_0 \) the stationary measure corresponding to \( Q_{\Delta t} \). Then, if the splitting scheme is of order \( p \), we define the bounded diameter of the state space as \( \hat{k} \),

\[
\hat{k} = \min\{ \text{diam}(S), p \} = \min\{ \max_{\sigma, \sigma'} \{ d(\sigma, \sigma') \}, p \}.
\]

Then, if \( C(\sigma, \sigma') \neq 0 \) for at least one pair \( \sigma, \sigma' \in S \) such that \( d(\sigma, \sigma') = \hat{k} \), we have that

\[
H(Q_{\Delta t}|P_{\Delta t}) = O(\Delta t^{2p-(\hat{k}+1)}).
\]
Proof. The proof of this theorem is the generalization of the argument given for Theorem 5.2. Picking up from formula (5.13),

$$J(\Delta t; \sigma, \sigma') = \frac{(C(\sigma, \sigma'))^2}{2Q_{\Delta t}(\sigma, \sigma') + \Delta t^p C(\sigma, \sigma') + o(\Delta t^p)} \Delta t^{2p} + o(\Delta t^{2p-k}).$$

(8.4)

Our goal is to show that $J(\Delta t; \sigma, \sigma') = O(\Delta t^{2p-k})$ for some $(\sigma, \sigma')$ and that this is the highest order attainable. Next, let us have $(\sigma, \sigma') \in S \times S$ such that $d(\sigma, \sigma') = \hat{k}$. Then, from (2.11) and (8.3), we have that

$$Q_{\Delta t}(\sigma, \sigma') = \sum_{k=0}^{\infty} \frac{L^k_{Q}[\delta_{\sigma'}](\sigma)}{k!} \Delta t^k = O(\Delta t^k), \quad \Delta t \in (0, 1].$$

(8.5)

Thus from (8.4) and (8.5), we can expose the first term of the asymptotic expansion of $F$ as

$$J(\Delta t; \sigma, \sigma') = \begin{cases} 
\frac{(C(\sigma, \sigma'))^2}{2L^k_{Q}[\delta_{\sigma'}](\sigma)/k!} \Delta t^{2p-k} + o(\Delta t^{2p-k}), & \hat{k} < p, \\
\frac{(C(\sigma, \sigma'))^2}{2L_{Q}[\delta_{\sigma'}](\sigma)/k! + C(\sigma, \sigma')} \Delta t^p + o(\Delta t^p), & \hat{k} = p.
\end{cases}$$

(8.6)

Next, we need to address the contribution of the rest of the expansion used (see the proof of Theorem 5.2), that is,

$$G(\Delta t; \sigma, \sigma') = Q_{\Delta t}(\sigma, \sigma') \sum_{k=1}^{\infty} \frac{1}{2k+1} \left( \frac{Q_{\Delta t}(\sigma, \sigma') - P_{\Delta t}(\sigma, \sigma')}{Q_{\Delta t}(\sigma, \sigma') + P_{\Delta t}(\sigma, \sigma')} \right)^{2k+1}.$$

If $\hat{k} < p$, then $G(\Delta t; \sigma, \sigma') = O(\Delta t^{3p-2k})$, which are lower-order terms given that $\Delta t \leq 1$. However, if $\hat{k} = p$, $G(\Delta t; \sigma, \sigma') = O(\Delta t^p)$ and in fact every term of the series in $G$ is of that order.

Finally, $H(Q_{\Delta t}/P_{\Delta t})$ can never have higher order than $p-1$, as that would require $(\sigma, \sigma')$ such that $d(\sigma, \sigma') > p+1$ and then $C(\sigma, \sigma') = 0$ (from Lemma 8.5).

The assumption on the commutator in Theorem 8.6 is simple to check for parallel KMC, as we can write down the commutator $C(\sigma, \sigma')$ explicitly. For example, for Lie, $C(\sigma, \sigma')$ is given by (5.2), so checking the assumption is just a matter of calculation. Additionally, to find the bounded diameter $\hat{k} = \min\{\text{diam}(S), p\}$, it is sufficient to have lower bounds for the diameter, diam$(S)$, as the order of the local error of the scheme, $p$, will typically be much smaller. Example 8.1 shows a case where $p$ is close to diam$(S)$ and the implications this has for the RER.

8.1. Markov chain example. In order to illustrate the connectivity-RER relation, we are studying a simple example where we can compute the RER and all related quantities explicitly, either by hand or by any symbolic algebra system. All calculations of the RER in this example are not from sampling but by using definition (3.4).

We study the case of a Markov process with transition rate matrix, $Q$ and diam$(S) = 2$. We consider a positive $\Delta t$, $\Delta t < 1$, and

$$Q = \begin{pmatrix}
-3 & 1 & 2 \\
3 & -4 & 1 \\
1 & 0 & -1
\end{pmatrix}.$$
Given this, we can calculate the transition probability matrix of the Markov chain as the matrix exponential of $Q$, $P_{\Delta t}(\sigma, \sigma') = \exp(\Delta t Q)\delta_{\sigma'}(\sigma)$. Our system has diameter equal to two since $Q_{3,2} = 0$ but $Q_{3,1}, Q_{1,2} \neq 0$. We can construct approximations of $P_{\Delta t}$ by splitting $Q$ into components $A, B$ with $Q = A + B$, similarly to how we expressed the generator $L$ as $L_1 + L_2$. One way to do this is

$$A = \begin{pmatrix} -3 & 1 & 2 \\ 3 & -4 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & -1 \end{pmatrix}.$$

Thus, one approximation of $\exp(Q\Delta t)$ could be $\exp(A\Delta t)\exp(B\Delta t)$, which corresponds to the Lie splitting. From Theorem 8.6, since $\text{diam}(S) = p = 2$, we expect $H(Q_{\Delta t}^{\text{Lie}}|P_{\Delta t}) = O(\Delta t)^1$. This is indeed the case, as

$$H(Q_{\Delta t}^{\text{Lie}}|P_{\Delta t}) \simeq 0.124\Delta t - 0.0566\Delta t^2 + O(\Delta t^3).$$

The use of $\simeq$ comes from a truncation of the coefficients to three significant digits.

We can work similarly with the Strang splitting, now using $\exp(A\Delta t/2)\exp(B\Delta t)\cdot\exp(A\Delta t/2)$ as the approximation to $P_{\Delta t}$. The local order of the Strang splitting is $p = 3$, so we expect $H(Q_{\Delta t}^{\text{Strang}}|P_{\Delta t}) = O(\Delta t^{2.3-3}) = O(\Delta t^3)$ (see Theorem 8.6). This can be readily demonstrated by a calculation of the RER, followed by the derivation of its asymptotic expansion:

$$H(Q_{\Delta t}^{\text{Strang}}|P_{\Delta t}) \simeq 0.0279\Delta t^3 + 0.000672\Delta t^4 + O(\Delta t^5).$$

9. Quantifying information loss in transient regimes. In this last section, we consider the case where we wish to study the performance of the operator splitting scheme in a transient regime, before convergence to the stationary distribution takes place. Note that in the proofs of Theorems 5.2 and 8.6, we derived the asymptotic expressions of the various quantities without referring to the stationary measure $\mu_Q$. Therefore those results do not depend on the choice of the sampling measure. That is, with the assumptions of Theorem 8.6 and $\nu$ a probability distribution on the state space $S^M$ such that $\nu(\sigma) > 0$ for all states $\sigma$, then

$$H_{\nu}(Q_{\Delta t}|P_{\Delta t}) = \sum_{\sigma \in S^M} \nu(\sigma)Q_{\Delta t}(\sigma_0, \sigma_1)\frac{Q_{\Delta t}(\sigma_0, \sigma_1)}{P_{\Delta t}(\sigma_0, \sigma_1)} = O(\Delta t^{2p-k}).$$

Therefore, the order of the RER is independent of the sampling measure. As a result, we gain Theorem 9.1, an extension of Theorem 8.6 to transient time regimes.

Theorem 9.1. With the assumptions of Theorem 8.6 for the RER, we have that for any $T > 0$

$$R(Q_0:T|P_{0:T}) = \frac{R(\mu_0|\nu_0)}{T} + O\left(\Delta t^{2p-k}\right).$$

Theorem 9.1 is implied by the decomposition of the relative entropy in terms of rates that depend on $\nu_t$ (first discussed in section 3). If $M$ is a positive integer, $\Delta t$ is the scheme’s time step, and $T = M\Delta t$, then

$$R(Q_0:T|P_{0:T}) = R(\mu_0|\nu_0) + \sum_{i=1}^{M} H_{\nu_i}(Q_{\Delta t}|P_{\Delta t}).$$
Proof of Theorem 9.1. From (9.1) we have that the order of the RER does not depend on the sampling measure \( \nu \), as long as \( \nu(\sigma) > 0 \) for all \( \sigma \). Therefore, \( H_{\nu_i}(Q_{\Delta t}|P_{\Delta t}) = O(\Delta t^{3p-\hat{k}}) \) for \( i = 1, \ldots, M \). This, combined with (9.3), implies the result.

Therefore, our results about the RER are applicable for parallel KMC even for practitioners that are interested in simulating the dynamics in the transient regime.

Remark 9.2 (RER versus pathwise relative entropy). In section 3, we saw that, in the stationary regime, we can relate the pathwise relative entropy with the RER via

\[
R(Q_0:T|P_0:T) = TH(Q_{\Delta t}|P_{\Delta t}) + R(Q_{\Delta t}|P_{\Delta t}).
\]

In this section, we connected the RER with the relative entropy for transient regimes by using relation (9.3). Ultimately, those relations motivate the use of the RER as an information criterion in place of the pathwise relative entropy, but there are other advantages too:

1. The RER does not depend on the length of the simulated path. Additionally, it can be estimated from a single path, while the pathwise relative entropy requires several.
2. For large \( T \), the relative entropy and RER encapsulate the same amount of information about the similarity of \( Q_{\Delta t} \) and \( P_{\Delta t} \).

10. Conclusions. We introduced the RER, i.e., path-space relative entropy per unit time, as a means to quantify the long-time accuracy of splitting schemes for stochastic dynamics and in particular parallel KMC algorithms. We demonstrated, using posteriori error expansions, the dependence of RER on the following elements: the local error analysis of the splitting schemes captured by the operator commutators; the local error order \( p \) and the splitting time step \( \Delta t \), which in the case of Parallel KMC controls the asynchrony between processors; and the diameter of the graph associated with the approximated Markov jump process.

Based on this analysis, we showed that RER defines a computable path-space information criterion that allows us to compare, select, and design different splitting schemes, taking into account both error tolerance (e.g., accuracy of the scheme) and practical concerns such as asynchrony and processor communication cost. It is also appropriate to think of the RER as a diagnostic quantity that can be estimated on systems of smaller size and consequently be used to compare schemes and tune parameters without slowing down the target simulation.

Finally we note that numerical analysis of stochastic systems is typically concerned with controlling the weak error for observable functions \( \phi \),

\[
\sup_{0 \leq n \leq N} \left| E_{P_{0:T}}[\phi(X(n\Delta t))] - E_{Q_{0:T}}[\phi(X_n)] \right|,
\]

where \( X_n \) represents the approximate chain and \( X(n\Delta t) \) the \( \Delta t \)-skeleton chain of the exact process, \( T = M \cdot \Delta t \). However, our results measure the information loss on path space between the approximate chain and the \( \Delta t \)-skeleton chain of the exact process, using RER. Controlling RER also implies upper bounds for observables at long times, using uncertainty quantification information inequalities developed in [15, 21]. We also showed how those results can be extended to finite-time regimes.
Appendix A. Coefficients of the relative entropy rate for Lie and Strang.

For the adsorption-desorption example considered in section 6 of the main text we need to estimate the highest-order coefficients $A, B$ for Lie and Strang, respectively. To accomplish this, we have to collect all the coefficients of $\Delta t$ and $\Delta t^2$ that appear in the expansion of RER in the proof of Theorem 5.2. The result is a summable series for each coefficient. For Lie, we have

(A.1) \[ A = \mathbb{E}_{\mu_L}(\sigma) \left[ \sum_{x,y \in \Lambda} F_L(\sigma, \sigma^{x,y}) \right] = \sum_{\sigma} \mu_L(\sigma) \sum_{x,y \in \Lambda} F_L(\sigma, \sigma^{x,y}), \]

(A.2) \[ F_L(\sigma, \sigma') := C_L(\sigma, \sigma') M_L(\sigma, \sigma') - 2L_L^2[\delta_{\sigma'}](\arctanh(M_L(\sigma, \sigma')) - M_L(\sigma, \sigma')), \]

\[ M_L(\sigma, \sigma') := C_L(\sigma, \sigma')/(L_L^2[\delta_{\sigma'}](\sigma) + C_L(\sigma, \sigma')), \]

where we remind the reader that $L_L^2$ stands for all the coefficients of $\Delta t^2/2$ in the expansion of the Lie splitting and $C_L(\sigma, \sigma') = 1/2[L_1, L_2][\delta_{\sigma'}](\sigma)$ is the Lie commutator term. Similarly, for the Strang case,

(A.3) \[ B = \mathbb{E}_{\mu_S}(\sigma) \left[ \sum_{x,y,z \in \Lambda} F_S(\sigma, \sigma^{x,y,z}) \right] = \sum_{\sigma} \mu_S(\sigma) \sum_{x,y,z \in \Lambda} F_S(\sigma, \sigma^{x,y,z}), \]

(A.4) \[ F_S(\sigma, \sigma') := C_S(\sigma, \sigma') M_S(\sigma, \sigma') - 2L_S^3[\delta_{\sigma'}](\arctanh(M_S(\sigma, \sigma')) - M_S(\sigma, \sigma')), \]

(A.5) \[ M_S(\sigma, \sigma') := C_S(\sigma, \sigma')/(L_S^3[\delta_{\sigma'}](\sigma) + C_S(\sigma, \sigma')). \]

Since both (A.1) and (A.3) are expected values, we can estimate them as ergodic averages.

REFERENCES


