Importance sampling for a robust and efficient multilevel Monte Carlo estimator for stochastic reaction networks

Chiheb Ben Hammouda1 · Nadhir Ben Rached2 · Raúl Tempone1,3

Received: 19 November 2019 / Accepted: 11 July 2020 / Published online: 20 August 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract
The multilevel Monte Carlo (MLMC) method for continuous-time Markov chains, first introduced by Anderson and Higham (SIAM Multiscal Model Simul 10(1):146–179, 2012), is a highly efficient simulation technique that can be used to estimate various statistical quantities for stochastic reaction networks, in particular for stochastic biological systems. Unfortunately, the robustness and performance of the multilevel method can be affected by the high kurtosis, a phenomenon observed at the deep levels of MLMC, which leads to inaccurate estimates of the sample variance. In this work, we address cases where the high-kurtosis phenomenon is due to catastrophic coupling (characteristic of pure jump processes where coupled consecutive paths are identical in most of the simulations, while differences only appear in a tiny proportion) and introduce a pathwise-dependent importance sampling (IS) technique that improves the robustness and efficiency of the multilevel method. Our theoretical results, along with the conducted numerical experiments, demonstrate that our proposed method significantly reduces the kurtosis of the deep levels of MLMC, and also improves the strong convergence rate from $\beta = 1$ for the standard case (without IS), to $\beta = 1 + \delta$, where $0 < \delta < 1$ is a user-selected parameter in our IS algorithm. Due to the complexity theorem of MLMC, and given a pre-selected tolerance, TOL, this results in an improvement of the complexity from $O(TOL^{-2} \log(TOL)^2)$ in the standard case to $O(TOL^{-2})$, which is the optimal complexity of the MLMC estimator. We achieve all these improvements with a negligible additional cost since our IS algorithm is only applied a few times across each simulated path.

Keywords Multilevel Monte Carlo · Continuous-time Markov chains · Stochastic reaction networks · Stochastic biological systems · Importance sampling

Mathematics Subject Classification 60H35 · 60J27 · 60J75 · 92C40

1 Introduction
In this work, we propose a novel importance sampling (IS) algorithm that can be combined with the multilevel Monte Carlo (MLMC) estimator to numerically solve stochastic differential equations (SDEs) driven by Poisson random measures (Li 2007; Çınlar 2011).

We focus on a particular class of continuous-time Markov chains known as stochastic reaction networks (SRNs) (see Sect. 1.1 for a short introduction). SRNs describe the time evolution of biochemical reactions, epidemic processes (Brauer and Castillo-Chavez 2001; Anderson and Kurtz 2015), and transcription and translation in genomics and virus kinetics (Srivastava et al. 2002; Hensel et al. 2009), among other important applications.

Let $X$ be an SRN taking values in $\mathbb{N}^d$ and defined in the time-interval $[0, T]$, where $T > 0$ is a user-selected final
time. We aim to provide accurate MLMC estimations of the expected value, \( E[g(X(T))] \), where \( g : \mathbb{R}^d \to \mathbb{R} \) is a given scalar observable of \( X \).

The main goal of our new proposed method is to improve the robustness and performance of the MLMC estimator by (i) solving the high-kurtosis phenomenon encountered when using the multilevel method in the context of continuous-time Markov chains (see Sect. 1.4), and (ii) improving the complexity of the MLMC estimator by increasing the strong convergence rate.

Many methods have been developed to simulate exact sample paths of SRNs; for instance, the stochastic simulation algorithm (SSA) was introduced by Gillespie in Gillespie (1976) and the modified next reaction method (MNRM) was proposed by Anderson in Anderson (2007). Pathwise exact realizations of SRNs may be computationally very costly when some reaction channels have high reaction rates. To overcome this issue, Gillespie (2001) and Aparicio and Solari (2001) independently proposed the explicit tau-leap (TL) method (see Sect. 1.2) to simulate approximate paths of \( X \) by evolving the process with fixed time steps, keeping the reaction rates fixed within each time step. Furthermore, other simulation schemes have been proposed to deal with situations with well-separated fast and slow time scales (Cao and Petzold 2005; Rathinam and El Samad 2007; Abdulle et al. 2010; Ahn et al. 2013; Moraes et al. 2016a; Hammouda et al. 2017).

To reduce the computational work needed to estimate \( E[g(X(T))] \), Anderson and Higham (2012) introduced the MLMC method (Giles 2008, 2015) based on the explicit TL scheme in the context of SRNs. Many extensions of the MLMC method have since been introduced to address other challenges. For instance, adaptive multilevel estimators (Lester et al. 2015; Moraes et al. 2016b, a) were proposed to improve the performance of non-adaptive estimators (Anderson and Higham 2012) to simulate SRNs with markedly different time scales. Hammouda et al. (2017) extended Anderson and Higham (2012) to systems with slow and fast time scales, and introduced a hybrid multilevel estimator that uses an implicit scheme for levels where explicit TL cannot be used due to numerical instability.

One important challenge encountered when using MLMC in the context of SRNs is the high-kurtosis phenomenon (see Sect. 1.4 for more details), which may occur due to either catastrophic coupling (characteristic of pure jump processes where coupled consecutive paths are identical in most of the simulations, while differences only appear in a tiny proportion; see Sect. 1.4.1 for more details) or catastrophic decoupling (observed for general stochastic processes where terminal values of the sample paths of both coarse and fine levels become very different from each other; see Sect. 1.4.2 for more details). This poor behavior of the kurtosis affects the accurate estimation of the sample variance needed for the MLMC algorithm. Consequently, it affects the robustness and performance of the multilevel estimator in many cases (see Sect. 1.4 for the illustration of this issue). As of today, few works have addressed this issue; for instance, the authors in Moraes et al. (2016b) mentioned this issue and developed a more accurate estimator for the multilevel variance based on dual-weighted residual expansion techniques. In Lester et al. (2018), a new method has been proposed to address the high-kurtosis phenomenon when it is due to catastrophic decoupling, and introduced a new approach of coupling consecutive levels of MLMC called the common process method (CPM), instead of using the split propensity method (SPM) proposed in Anderson and Higham (2012). The CPM is based on the use of common inhomogeneous Poisson processes for both coarse and fine sample paths. Although the CPM improves the robustness and reliability of the multilevel estimator by dramatically decreasing the kurtosis, it nonetheless incurs remarkable additional computational and memory costs because for each level it requires (i) running the TL algorithm twice, and (ii) storing the total number of times each Poisson process has fired over each time step.

In the work presented here, compared to Lester et al. (2018), we address cases of high kurtosis observed in the MLMC estimator due to catastrophic coupling and propose a novel method that provides a more robust multilevel estimator. We introduce a pathwise-dependent IS technique to dramatically decrease the high kurtosis caused by the SPM strategy for coupling the paths of two consecutive levels. We should note that other IS methods were proposed, in the context of biochemical systems and SRNs, but for the efficient estimation of rare events (Kuwahara and Mura 2008; Daigle Jr et al. 2011; Cao and Liang 2013). Furthermore, these IS methods were combined with the MC method instead of the MLMC method that we present here.

We show that our proposed method not only improves the robustness of the multilevel estimator by significantly reducing the kurtosis, but also improves the strong convergence rate from \( \beta = 1 \) for the standard case (without IS), to \( \beta = 1 + \delta \), where \( 0 < \delta < 1 \) is a user-selected parameter in our IS algorithm. Due to the complexity theorem of MLMC (Cliffe et al. 2011), and given a pre-selected tolerance, TOL, this results in an improvement of the complexity of MLMC from \( \mathcal{O}(TOL^{-2}\log(TOL)^2) \) to the optimal complexity, i.e., \( \mathcal{O}(TOL^{-2}) \). We achieve all these improvements with a negligible additional cost since our IS algorithm is only applied a few times across each simulated path.

Alternatively, the optimal MLMC complexity of order \( \mathcal{O}(TOL^{-2}) \) can be achieved by using (i) MC with an exact scheme (for instance SSA), or (ii) an unbiased MLMC estimator (Anderson and Higham 2012), where the deepest level is simulated with an exact scheme, or (iii) a biased hybrid MLMC estimator (Moraes et al. 2016b), where the paths are
simulated in a hybrid fashion that switches adaptively, based on the relative computational cost, between the TL and an exact method. Both approaches (i) and (ii) incur a substantial additional cost by introducing an exact scheme. This significant additional cost is not manifested in the rate exponent but in a large constant that deteriorates the actual complexity. Although our method is based on a biased MLMC estimator, without steps simulated with an exact scheme as in Moraes et al. (2016b), it still achieves a complexity of order $O(TOL^{-2})$ with a smaller constant than those produced by the methods (i), (ii) and (iii) mentioned above. Compared to (Anderson and Higham 2012), we suggest an orthogonal approach of lowering the complexity rate by improving the strong convergence rate, instead of removing the bias (weak error). Similarly to our work, the authors in Moraes et al. (2016b) improve the strong convergence rate to reach the complexity of order $O(TOL^{-2})$. However, compared to (Moraes et al. 2016b), we use a different strategy based on a pathwise-dependent IS coupled with the TL scheme, instead of using a hybrid approach that switches between an exact and the TL scheme.

We also propose a new approach to overcome the high-kurtosis phenomenon, which affects the robustness and reliability of the MLMC estimator introduced in Anderson and Higham (2012). Although this issue can be addressed differently, using the dual-weighted residual expansion techniques developed in Moraes et al. (2016b) in order to estimate more accurately the sample variance and bias on the deepest levels of MLMC, we believe that our approach has two main advantages over the approach in Moraes et al. (2016b): first, our method is much simpler and easier to generalize to other schemes, such as the split-step implicit TL scheme (Hammouda et al. 2017) where it is difficult to get estimates using the dual-weighted residual expansion techniques. Furthermore, although the approach in Moraes et al. (2016b) provides a more accurate estimate of the variance than the sample variance estimate, there is still no clear analysis of how accurate (biased) those estimates are. The difficulty of establishing such analysis is mainly due to the lack of sharp concentration inequalities for linear combinations of independent Poisson random variables (rdvs), as stated in Remark 4 in Moraes et al. (2016b). Finally, we should emphasize that the hybrid scheme in Moraes et al. (2016b) is an efficient algorithm that avoids the simulated paths to take negative values, which is an undesirable consequence of the TL approximation. In this case, for problems where we are close to the boundary, combining the two approaches [our approach and the approach in Moraes et al. (2016b)] may lead to more efficient results.

This work is structured as follows: we start by giving an overview of concepts used in this work such as SRNs (Sect. 1.1), explicit TL approximation (Sect. 1.2), and the MLMC method (Sect. 1.3). Then, in Sect. 1.4, we explain the high-kurtosis phenomenon along with its leading causes in the context of SRNs. In Sects. 2, 3 and 4, we present the details of our IS algorithm that we combine with the MLMC method. We start by presenting in Sect. 2 the motivation of our idea by the sampling under an optimal measure for simulating SRNs. Then, in Sect. 3, we present a summary of the main results of this work, and in Sect. 4, we analyze our proposed IS algorithm and state the main convergence theorems related to the kurtosis and the variance estimates of our approach. Furthermore, we present, in the same section, a cost analysis of the MLMC methods presented in this work, with and without IS. Before concluding, we show, in Sect. 5, the results obtained through the numerical experiments conducted across different examples of SRNs.

1.1 Stochastic reaction networks (SRNs)

We are interested in the time evolution of a homogeneously mixed chemical reacting system described by the Markovian pure jump process, $X: [0, T] \times \Omega \to \mathbb{N}^d$, where $(\Omega, \mathcal{F}, P)$ is a probability space. In this framework, we assume that $d$ different species interact through $J$ reaction channels. The $i$-th component, $X^{(i)}(t)$, describes the abundance of the $i$-th species present in the chemical system at time $t$. This work aims to study the time evolution of the state vector,

$$X(t) = (X^{(1)}(t), \ldots, X^{(d)}(t)) \in \mathbb{N}^d.$$  

Each reaction channel, $R_j$, is a pair $(a_j, \nu_j)$ defined by its propensity function, $a_j : \mathbb{R}^d \to \mathbb{R}^+$, and its state change vector, $\nu_j = (\nu_{j,1}, \nu_{j,2}, \ldots, \nu_{j,d})$, satisfying

$$\text{Prob} \left( X(t + \Delta t) = x + \nu_j ; X(t) = x \right) = a_j(x)\Delta t + o(\Delta t), \quad j = 1, 2, \ldots, J. \quad (1.1)$$

Formula (1.1) states that the probability of observing a jump in the process, $X$, from state $x$ to state $x + \nu_j$, a consequence of the firing of reaction $R_j$ during a small time interval, $(t, t + \Delta t)$, is proportional to the length of the time interval, $\Delta t$, with $a_j(x)$ as the constant of proportionality.

We set $a_j(x) = 0$ for $x$ such that $x + \nu_j \notin \mathbb{N}^d$ (the non-negativity assumption: the system can never produce negative population values).

As a consequence of relation (1.1), the process $X$ is a continuous-time, discrete-space Markov chain that can be characterized by the random time change representation of Kurtz Ethier and Kurtz (1986)

$$X(t) = x_0 + \sum_{j=1}^J Y_j \left( \int_0^t a_j(X(s)) \, ds \right) \nu_j, \quad (1.2)$$

Hereafter, we use $\text{Prob}(A; B)$ and $\text{E}[A; B]$ to denote the conditional probability and conditional expectation of $A$ given $B$, respectively.
where \( Y_j : \mathbb{R}_+ \times \Omega \rightarrow \mathbb{N} \) are independent unit-rate Poisson processes. Conditions on the reaction channels can be imposed to ensure uniqueness (Anderson and Kurtz 2015) and to avoid explosions in finite time (Engblom 2012; Rathinam 2013; Gupta et al. 2014).

We emphasize that, by using the stochastic mass-action kinetics principle, we assume that the propensity function, \( a_j(\cdot) \), for a reaction channel \( R_j \), represented by the following diagram\(^2\):

\[
\alpha_j,1 S_1 + \cdots + \alpha_j,d S_d \xrightarrow{\theta_j} \beta_j,1 S_1 + \cdots + \beta_j,d S_d,
\]

obeys the following relation

\[
a_j(x) := \theta_j \prod_{i=1}^{d} \frac{x_i!}{(x_i - \alpha_{j,i})!} 1_{\{x_i \geq \alpha_{j,i}\}}, \quad (1.3)
\]

where \( \{\theta_j\}_{j=1}^{J} \) are positive constant reaction rates, \( x_i \) is the counting number of the species \( S_i \), and \( 1_{A} \) is the indicator function of the set \( A \).

### 1.2 The explicit Tau-leap (Explicit-TL) approximation

The explicit-TL scheme is a pathwise-approximate method independently introduced in Gillespie (2001) and Aparicio and Solari (2001) to overcome the computational drawback of exact methods, i.e., when many reactions fire during a short time interval. This scheme can be derived from the random time change representation of Kurtz (1.2) by approximating the integral \( \int_{t_i}^{t_i+\Delta t} a_j(X(t)) \, dt \) by \( a_j(X(t_i))(t_i+\Delta t - t_i) \), i.e., using the forward-Euler method with a time mesh \( \{t_0 = 0, t_1, \ldots, t_N = T\} \). In this way, the explicit-TL approximation of \( X \) should satisfy for \( k \in \{1, 2, \ldots, N\} \)

\[
Z(t_k) = x_0 + \sum_{j=1}^{J} Y_j \left( \sum_{i=0}^{k-1} a_{j}(Z(t_i))(t_{i+1} - t_i) \right) v_j.
\]

Given a uniform time mesh of size \( \Delta t \) and \( Z(0) := x_0 \), we simulate a path of \( Z \) as follows

\[
Z(t_k) := z + \sum_{j=1}^{J} \mathcal{P}_j(a_j(Z(t_k))\Delta t)v_j, \quad 1 \leq k \leq N,
\]

iteratively, where \( z = Z(t_{k-1}) \) and \( \{\mathcal{P}_j(r_j)\}_{j=1}^{J} \) are independent Poisson r.v.s with respective rates, \( r_j \). Note that the explicit-TL path, \( Z \), is defined only at the points of the time mesh, but it can be naturally extended to \([0, T]\) as a piecewise constant path.

### 1.3 The multilevel Monte Carlo (MLMC) method

Let \( X \) be a stochastic process and \( g : \mathbb{R}^d \rightarrow \mathbb{R} \) a scalar observable. Let us assume that we want to approximate \( E[g(X(T))] \), but instead of sampling directly from \( X(T) \), we sample from \( Z_{\Delta t}(T) \), which are r.v.s generated by an approximate method with step size \( \Delta t \). Let us also assume that the variates \( Z_{\Delta t}(T) \) are generated with an algorithm with weak order, \( \mathcal{O}(\Delta t) \), i.e., \( E[g(X(T)) - g(Z_{\Delta t}(T))] = \mathcal{O}(\Delta t)^{1.3} \).

Let \( \mu_M \) be the standard Monte Carlo estimator of \( E[g(Z_{\Delta t}(T))] \) defined by

\[
\mu_M := \frac{1}{M} \sum_{m=1}^{M} g(Z_{\Delta t,m}(T)),
\]

where \( \{Z_{\Delta t,m}(T)\}_{m=1}^{M} \) are independent and distributed as \( Z_{\Delta t}(T) \).

We define the global error of the MC estimator as \( (E[g(X(T)) - \mu_M])^2 \). Then, we write the following error decomposition

\[
E \left[ (E[g(X(T)) - \mu_M])^2 \right] = (E[g(X(T)) - g(Z_{\Delta t}(T))]^2 + (E[g(Z_{\Delta t}(T))] - \mu_M)^2.
\]

Variance

To achieve the desired accuracy, TOL, it is sufficient to take \( \Delta t = \mathcal{O}(\text{TOL}) \) so that the bias is \( \mathcal{O}(\text{TOL}) \) and impose \( M = \mathcal{O}(\text{TOL}^{-3}) \) so that the variance is \( \mathcal{O}(\text{TOL}) \) (Duffie and Glynn 1995). As a consequence, the expected total computational work is \( \mathcal{O}(\text{TOL}^{-3}) \).

The MLMC estimator, introduced by Giles (2008) (see also Kebaier 2005 for the two-level construction), allows us to reduce the total computational work up to \( \mathcal{O} \left( \text{TOL}^{-2} \log(\text{TOL})^{2} \times 1_{[0,1]} \right) \), where (\( \alpha, \beta, \gamma \) are weak, strong, and work rates, respectively (see Theorem 1.1 for more details). The basic idea of MLMC is to generate, and couple in a clever manner, paths with different step sizes. We can construct the MLMC estimator as follows: consider a hierarchy of nested meshes of the time interval \([0, T]\), indexed by \( \ell = 0, 1, \ldots, L \). We denote by \( \Delta t_\ell \) the step size used at level \( \ell = 0 \). The size of the subsequent time steps for levels \( \ell \geq 1 \) is given by \( \Delta t_\ell = K^{-\ell} \Delta t_0 \), where \( K > 1 \)

\(^2\) \( \alpha_{j,i} \) molecules of the species \( S_i \) are consumed and \( \beta_{j,j} \) are produced. Thus, \( \alpha_{j,i}, \beta_{j,j} \in \mathbb{N}^2 \) but \( \beta_{j,i} - \alpha_{j,i} \) can be a negative integer, constituting the vector \( v_j = (\beta_{j,1} - \alpha_{j,1}, \ldots, \beta_{j,d} - \alpha_{j,d}) \in \mathbb{Z}^d \).

\(^3\) We refer to Li (2007) for the underlying assumptions and proofs of this statement, in the context of the TL scheme.

\( \copyright \) Springer
is a given integer constant. In this work, we take \( K = 2 \). Furthermore, we denote by \( M_\ell \) the number of samples per level in the MLMC estimator. To simplify the notation, hereafter \( Z_\ell := Z_{\Delta \ell} \) denotes the approximate process generated using a step size of \( \Delta \ell \).

Consider now the following telescoping decomposition of \( \mathbb{E} [g(Z_L(T))] \)

\[
\mathbb{E} [g(Z_L(T))] = \mathbb{E} [g(Z_0(T))] + \sum_{\ell=1}^L \mathbb{E} [g(Z_\ell(T)) - g(Z_{\ell-1}(T))]
\]

Then, by defining

\[
\hat{Q}_0 := \frac{1}{M_0} \sum_{m_0=1}^{M_0} g(Z_0, [m_0](T)) \\
\hat{Q}_\ell := \frac{1}{M_\ell} \sum_{m_\ell=1}^{M_\ell} \left( g(Z_\ell, [m_\ell](T)) - g(Z_{\ell-1}, [m_{\ell-1}](T)) \right),
\]

we arrive at the unbiased MLMC estimator, \( \hat{Q} \), of \( \mathbb{E} [g(Z_L(T))] \)

\[
\hat{Q} := \sum_{\ell=0}^L \hat{Q}_\ell.
\]

We note that the key point here is that both \( Z_{\ell, [m_\ell]}(T) \) and \( Z_{\ell-1, [m_{\ell-1}]}(T) \) are sampled using different time discretizations but with the same generated randomness.

Theorem 1.1 from Cliffe et al. (2011) states the computational complexity of the MLMC estimator for different scenarios:

**Theorem 1.1 (MLMC complexity)** Let \( g := g(X(T)) \) denote a rdv, and let \( g_\ell := g(Z_\ell(T)) \) denote the corresponding level \( \ell \) numerical approximation. If there exist independent estimators \( \hat{Q}_\ell \) based on \( M_\ell \) Monte Carlo samples, each with expected cost \( W_\ell \) and variance \( V_\ell \), and positive constants \( \alpha \) (weak convergence rate), \( \beta \) (strong convergence rate), \( \gamma \) (work rate), \( c_1, c_2, c_3 \) such that \( \alpha \geq \min(\beta, \gamma) \) and

(i) \( |\mathbb{E} [g_\ell - g]| \leq c_1 2^{-\alpha \ell} \)

(ii) \( \mathbb{E} [\hat{Q}_\ell] = \mathbb{E} [g_0], \ \ell = 0 \)

(iii) \( V_\ell := \text{Var} [g_\ell - g_{\ell-1}], \ \ell > 0 \)

(iv) \( W_\ell \leq c_3 2^{\gamma \ell} \),

then there exists a positive constant \( c_4 \) such that for any \( TOL < e^{-1} \), there are values \( L \) and \( M_\ell \) for which the multi-level estimator

\[
\hat{Q} = \sum_{\ell=0}^L \hat{Q}_\ell,
\]

has a mean-square-error with bound

\[
\mathbb{E} \left[ (\hat{Q} - \mathbb{E} [g])^2 \right] < TOL^2,
\]

with a computational complexity \( W \) with bound

\[
\mathbb{E} [W] = \begin{cases} c_4 TOL^{-2}, & \beta > \gamma, \\ c_4 TOL^{-2} (\log(TOL))^2, & \beta = \gamma, \\ c_4 TOL^{-2 - \frac{\beta}{\gamma}}, & \beta < \gamma. \end{cases}
\]

We emphasize that Theorem 1.1 still applies to our approach, proposed in Sect. 4, since we only modify the way we sample coupled paths in this context, by combining the standard way of coupling two tau-leap paths with our IS algorithm. Our proposed IS technique does not change the weak rate but improves the strong convergence rate, \( \beta \), thus leading to an improvement of the MLMC complexity rate, to reach the optimal rate.

**1.4 The high-Kurtosis phenomenon**

Let \( g \) denote a rdv, and let \( g_\ell \) denote the corresponding level \( \ell \) numerical approximation. We also define \( Y_\ell := g_\ell - g_{\ell-1} \).

The standard deviation of the sample variance for the rdv \( Y_\ell \) is given by

\[
\sigma^2_{\mathbb{S}^2(Y_\ell)} = \frac{\text{Var} [Y_\ell]}{M_\ell} \sqrt{\kappa_\ell - 1 + \frac{2}{M_\ell - 1}},
\]

where the kurtosis \( \kappa_\ell = \frac{\mathbb{E} [(Y_\ell - \mathbb{E} [Y_\ell])^4]}{(\text{Var} [Y_\ell])^2} \).

For the setting of the MLMC method, accurate estimates of \( V_\ell = \text{Var} [Y_\ell] \) are required since the optimal number of samples per level, \( M_\ell^* \), for the MLMC estimator is given by (see Giles 2015 for more details)

\[
M_\ell^* = \left[ 2 TOL^{-2} \sqrt{V_\ell W_\ell^{-1}} \sum_{\ell=L_0}^L \sqrt{V_\ell W_\ell} \right],
\]

where \( [x] := \text{ceil}(\times) \), \( W_\ell \) is the cost per sample path, TOL is the accuracy of the MLMC estimator, and \( L_0 \) \(^4\) and \( L \) are the coarsest and the deepest levels of the MLMC estimator, respectively.

\(^4\) We set \( L_0 = 0 \) unless otherwise stated. In our numerical experiments, we select \( L_0 \) such that \( \text{Var} [g_{L_0+1} - g_{L_0}] \ll \text{Var} [g_{L_0}] \), in order to ensure the stability of the variance of the coupled paths of our MLMC estimator.
The high kurtosis makes it challenging to estimate \( V_\ell \) accurately, since \( O(\kappa) \) samples are required to obtain a reasonable estimate of the variance [see (1.7)]. Two possible consequences of the high kurtosis may occur, and deteriorate the robustness and the performance of the MLMC estimator:

- The sample variance, \( V_\ell \), is an under-estimate. The effect is that the required confidence interval semi-length is not faithfully attained, due to \( \sigma^2 (V_\ell) \) given by (1.7).
- The sample variance, \( V_\ell \), is an over-estimate. In this case, too many sample paths are generated, and the algorithm takes substantially more time to run.

Several studies (Giles 2015; Gou 2016; Moraes et al. 2016b; Hammouda et al. 2017; Lester et al. 2018; Bayer et al. 2020) discussed the issue of high kurtosis when using MLMC, for different applications. In the context of SRNs, there are mainly two causes of the high-kurtosis phenomenon: (i) Catastrophic coupling or (ii) Catastrophic decoupling. In the following subsections, we explain these two causes.

### 1.4.1 Catastrophic coupling

The high-kurtosis phenomenon, in this case, is caused by catastrophic coupling (see Section 1.7 of Moraes et al. 2016b), which is a characteristic of pure jump processes that motivates this work. When using the MLMC estimator in this context, the following issue is usually encountered: When \( \ell \) (MLMC level) becomes large, due to the used coupling strategy (see Sect. 2), \( Y_\ell \) is different from zero only in a tiny proportion of the simulated coupled paths (see Figs. 8, 9, 10). This behavior is one of the leading causes of the high-kurtosis phenomenon (see Figs. 5, 6 and 7), resulting in inaccurate estimates of the sample variance [see (1.7)].

As an illustration of catastrophic coupling, consider an example when \( g \) takes values in \([0, 1]\), and let \( g_\ell \) denote the corresponding level \( \ell \) numerical approximation in the MLMC estimator. In this case, we have

\[
Y_\ell = g_\ell - g_{\ell - 1} = \begin{cases} 
1, & \text{with probability } p_\ell \\
-1, & \text{with probability } q_\ell \\
0, & \text{with probability } 1 - p_\ell - q_\ell. 
\end{cases} \tag{1.9}
\]

Observe that this example is a true illustration of the SRNs that we consider in this work. For instance, by observing the histograms in Figs. 8, 9 and 10, we can check that we usually encounter the situation manifested by (1.9), with \( p_\ell, q_\ell \ll 1 \), and \( \operatorname{Prob}(g_\ell - g_{\ell - 1} = 0) \to 1 \) as \( \ell \) increases.

If \( p_\ell, q_\ell \ll 1 \), then \( \mathbb{E}[Y_\ell] \approx 0 \) and \( \kappa_\ell \approx (p_\ell + q_\ell)^{-1} \gg 1 \). Therefore, many samples are required for an accurate estimate of \( V_\ell = \text{Var}[Y_\ell] \), since using (1.7), we need \( M_\ell \gg \kappa_\ell \ell \to \infty \); otherwise, we may get all samples \( Y_\ell = 0 \), which gives an estimated variance of zero. Furthermore, the kurtosis becomes worse as \( \ell \to \infty \) since \( p_\ell, q_\ell \to 0 \) due to weak convergence.

### 1.4.2 Catastrophic decoupling

The high-kurtosis phenomenon can also occur because of catastrophic decoupling, as explained in Lester et al. (2016) and observed in Lester et al. (2018). Catastrophic decoupling occurs when the terminal values of the sample paths of both coarse and fine levels become very different from each other. In fact, due to the SPM coupling strategy (see Sect. 2), all reactions start immediately in the fine level and not in the coarse level, since reactions cannot happen until the reaction propensities are updated. We note that this issue becomes more severe when dealing with large scales of species count.

We emphasize that we do not treat catastrophic decoupling with our novel proposed method, but rather we address the case of catastrophic coupling. Nonetheless, catastrophic decoupling can be addressed by using a different coupling, such as CPM coupling (Lester et al. 2018). In a future work, to address the issue of catastrophic decoupling, we intend to explore the possibility of introducing a new IS scheme for MLMC based on SPM coupling.

**Remark 1.1**

As proposed in Gillespie (2000), SRNs paths can be approximated using the chemical Langevin equation (CLE), which is only valid when the expected number of occurrences of each reaction channel \( R_j \) in \([t, t + \Delta t]\) is much larger than 1, i.e.,

\[
a_j(x_t) \Delta t \gg 1, \quad \forall j \in \{0, 1, \ldots, J\}. \tag{1.10}
\]

Assumption (1.10), implicitly implies that the system has large molecular population numbers. In this work, we do not impose this restriction on the examples we consider. Moreover, such an assumption does not hold in our setting and more precisely in the deepest level of MLMC estimator (\( \Delta t \) very small).

## 2 Motivation

### 2.1 Characterization of the original coupling measure

Let us use the notations of Sect. 1.3, and denote \( g_\ell := g(Z_\ell(T)) \). Then, we can rewrite (1.4) as

\[
E[g_\ell] = \sum_{\ell=1}^{L} E[g_\ell - g_{\ell-1}] + E[g_0]. \tag{2.1}
\]
where each term in (2.1) can be written as

$$E [g_0] = \int g_0 d\mathbb{P}_0, \quad E [g_\ell - g_{\ell-1}] = \int (g_\ell - g_{\ell-1}) d\mathbb{P}_\ell,$$

(2.2)

where $\mathbb{P}_\ell$ is the coupling measure and $\mathbb{P}_0$ is the single level measure.

To characterize the original coupling measure $\mathbb{P}_\ell$ in the context of SRNs, we define the pure jump process $X$ by the Kurtz representation, as in (1.2). For the sake of simplicity, let us consider $X$ to be one-dimensional (only one species), only one reaction ($J = 1$) (in this case we denote the state change scalar by $v_1$; see (1.1)), and $g(x) = x, \ x \in \mathbb{R}$. We denote $\bar{X}_{\ell-1}, \bar{X}_\ell$ the two TL approximations of the true process $X$ based on two consecutive grid levels $(\ell - 1, \ell)$ and recall that $\Delta t_{\ell-1} = 2\Delta t$ (equivalently, we denote by $N_{\ell-1}$ and $N_\ell$ the number of times steps used at levels $\ell - 1$ and $\ell$, respectively). Let $0 \leq n \leq N_{\ell-1} - 1$. If we consider two consecutive time-mesh points for $\bar{X}_{\ell-1}, (n, n+1)$, and three consecutive time-mesh points for $\bar{X}_\ell, (n, n + \Delta t, n+1)$, then we have

$$\bar{X}_{\ell-1}(n+1) = \bar{X}_{\ell-1}(n) + v_1 Y_1,n (a (\bar{X}_{\ell-1}(n)) \Delta t_{\ell-1})$$

$$\bar{X}_\ell(t_{n+1}) = \bar{X}_\ell(t_n) + v_1 Q_1,n (a (\bar{X}_\ell(t_n)) \Delta t_\ell) + v_1 R_1,n (a (\bar{X}_\ell(t_n + \Delta t_\ell)) \Delta t_\ell),$$

(2.3)

where $Y_1,n, Q_1,n, R_1,n$ are conditionally independent Poisson rdvs.

To couple the $\bar{X}_{\ell-1}$ and $\bar{X}_\ell$ processes, we first decompose $Y_1,n (a (\bar{X}_{\ell-1}(n)) \Delta t_{\ell-1})$ as the sum of two conditionally independent Poisson rdvs, $P_{1,n} (a (\bar{X}_{\ell-1}(n)) \Delta t_{\ell}) + P_{2,n} (a (\bar{X}_{\ell-1}(n)) \Delta t_{\ell})$. Then, by applying this decomposition in (2.3), we obtain

$$\bar{X}_{\ell-1}(n+1) = \bar{X}_{\ell-1}(n) + v_1 P_{1,n} (a (\bar{X}_{\ell-1}(n)) \Delta t_{\ell})$$

$$\bar{X}_\ell(t_{n+1}) = \bar{X}_\ell(t_n) + v_1 Q_1,n (a (\bar{X}_\ell(t_n)) \Delta t_\ell) + v_1 R_1,n (a (\bar{X}_\ell(t_n + \Delta t_\ell)) \Delta t_\ell).$$

Furthermore, by using the same reasoning of coupling strategy as in Anderson and Higham (2012), we can show that for the first time interval $[n, n + \Delta t]$, we have

$$\bar{X}_{\ell-1}(n+1) = \bar{X}_{\ell-1}(n) + \left( P_{1,n} (m_{1,n}^1 \Delta t_{\ell}) + \mathbb{P}_n^\prime \left( (a (\bar{X}_{\ell-1}(n)) - m_{1,n}^1 \Delta t_{\ell}) \right) \right) v_1$$

$$\bar{X}_\ell(t_{n+1}) = \bar{X}_\ell(t_n) + \left( P_{1,n} (m_{1,n}^1 \Delta t_\ell) \right) v_1,$$

(2.4)

where $m_{1,n}^1 = \min \left( a (\bar{X}_{\ell-1}(n)), a (\bar{X}_{\ell-1}(n)) \right)$, and $\mathbb{P}_n, \mathbb{P}_n^\prime$, $\mathbb{P}_n^\prime$ are conditionally independent Poisson rdvs.

For the time interval $[t_n + \Delta t, t_{n+1}]$, we have

$$\bar{X}_{\ell-1}(t_{n+1}) = \bar{X}_{\ell-1}(n + \Delta t_\ell) + \left( Q_n^\prime (m_{2,n}^2 \Delta t_\ell) + \mathbb{P}_n^\prime \left( (a (\bar{X}_{\ell-1}(n)) - m_{2,n}^2 \Delta t_\ell) \right) \right) v_1$$

$$\bar{X}_\ell(t_{n+1}) = \bar{X}_\ell(t_n + \Delta t_\ell) + \left( Q_n^\prime (m_{2,n}^2 \Delta t_\ell) + \mathbb{P}_n^\prime \left( (a (\bar{X}_\ell(t_n + \Delta t_\ell)) - m_{2,n}^2 \Delta t_\ell) \right) \right) v_1,$$

(2.5)

where $m_{2,n}^2 = \min \left( a (\bar{X}_\ell(t_n + \Delta t_\ell)), a (\bar{X}_{\ell-1}(n)) \right)$, and $Q_n^\prime, Q_n^\prime, Q_n^\prime$ are conditionally independent Poisson rdvs.

Equations (2.4) and (2.5) imply that

$$\bar{X}_{\ell-1}(t_{n+1}) - \bar{X}_{\ell-1}(t_n) = \mathbb{P}_n (a (\bar{X}_{\ell-1}(n)) \Delta t_{\ell-1})$$

$$+ v_1 \left( P_{1,n} (\Delta a_{1,n}^1 \Delta t_{\ell}) 1_{\Delta a_{1,n}^1 > 0} - P_{2,n} (\Delta a_{1,n}^1 \Delta t_{\ell}) 1_{\Delta a_{1,n}^1 < 0} \right),$$

(2.6)

where $\Delta a_{1,n}^1 = a (\bar{X}_\ell(t_n)) - a (\bar{X}_{\ell-1}(n))$ and $\Delta a_{1,n}^1 = a (\bar{X}_\ell(t_n + \Delta t_\ell)) - a (\bar{X}_{\ell-1}(n))$.

In the following, we denote, for $0 \leq n \leq N_{\ell-1} - 1$ (note that $N_\ell = 2N_{\ell-1}$),

$$\Delta a_{\ell,2n} = |\Delta a_{1,n}^1|, \quad \Delta a_{\ell,2n+1} = |\Delta a_{2,n}^1|, \quad \Delta a_{\ell,2n+2} = |\Delta a_{1,n}^1|, \quad \Delta a_{\ell,2n+3} = |\Delta a_{2,n}^1|,$$

(2.7)

Note that in (2.6), not only are $\mathbb{P}_n^\prime, \mathbb{P}_n^\prime, \mathbb{P}_n^\prime$ rdvs, but $\Delta a_{\ell,2n}$ and $\Delta a_{\ell,2n+1}$ [defined in (2.7)] are also rdvs, because of their dependence on $\bar{X}_{\ell-1}(t_n), \bar{X}_\ell(t_n)$, and $\bar{X}_\ell(t_n + \Delta t_\ell)$. Therefore, to derive some of the following formulas for analyzing our IS algorithm, we need to consider a sigma-algebra, $\mathbb{F}_n$, with $0 \leq n_\ell \leq N_\ell - 1$, such that $\Delta a_{\ell,n_\ell}$, conditioned on $\mathbb{F}_n$, is deterministic, i.e., $\Delta a_{\ell,n_\ell}$ is measurable with respect to $\mathbb{F}_n$. This way, the only randomness being considered comes from the Poisson rdvs used for updating the states of $\bar{X}_\ell(t_{n+1})$ and $\bar{X}_{\ell-1}(t_{n+1})$. For this purpose, we consider for a fixed $n_\ell, \mathbb{F}_n$ as the sigma algebra

$$\mathbb{F}_n := \sigma \left( (\Delta a_{\ell,k})_{k=0,\ldots,n_\ell} \right), \quad n_\ell = 0, \ldots, N_\ell - 1.$$
In what follows, the terms \( \Delta a_{\ell,n}^{j} \), defined in (2.7), will be denoted for the multi-channel case, by \( \{ \Delta a_{\ell,n}^{j} \}_{n=0}^{N_{\ell} - 1} \), where \( j \in \{1, \ldots, J \} \) corresponds to the index of the reaction channel.

### 2.2 Characterization of the optimal change of measure

It is known that, the optimal change of measure, \( \pi_{\ell} \), the one that achieves the minimum variance, satisfies

\[
d\pi_{0} \propto |g_0|d\mathbb{P}_0, \quad d\pi_{\ell} \propto |g_{\ell} - g_{\ell-1}|d\mathbb{P}_\ell.
\]

Observe that the optimal measure, \( \pi_{\ell} \), removes the probability mass at zero, where most of \( \mathbb{P}_\ell \) is concentrated due to catastrophic coupling (explained in Sect. 1.4.1). We emphasize that, in this work, we aim to perform a change of measure with respect to \( \mathbb{P}_\ell \), while keeping the single level measure \( \mathbb{P}_0 \) unchanged.

The minimum variance is given by

\[
\text{Var}_{\pi_{\ell}} \left[ (g_{\ell} - g_{\ell-1}) \frac{d\mathbb{P}_{\ell}}{d\pi_{\ell}} \right] = \left( E_{\mathbb{P}_{\ell}} \left[ |g_{\ell} - g_{\ell-1}| \right] \right)^2 \left( 1 - \left( E_{\pi_{\ell}} \left[ \text{sgn}(g_{\ell} - g_{\ell-1}) \right] \right)^2 \right)
\]

\[
= \left( E_{\mathbb{P}_{\ell}} \left[ |g_{\ell} - g_{\ell-1}| \right] \right)^2 \left( 1 - (\pi_{\ell}(g_{\ell} - g_{\ell-1} > 0) - \pi_{\ell}(g_{\ell} - g_{\ell-1} < 0)) \right)^2,
\]

where \( \text{sgn}(.) \) is the sign function.

Interestingly, using Theorem 3.2 in Anderson et al. (2011) in the context of the explicit TL scheme for pure jump processes, we conclude that \( E_{\mathbb{P}_{\ell}} \left[ |g_{\ell} - g_{\ell-1}| \right] = O(\Delta t_{\ell}) \) for any Lipschitz function \( g \). Therefore, we clearly observe that the optimal IS improves the strong convergence rate, and hence leads to the optimal complexity rate of the MLMC estimator (see Theorem 1.1).

Unfortunately, it is unfeasible to sample from \( \pi_{\ell} \); therefore, our goal in the following sections is to propose a practical IS algorithm with a sub-optimal change of measure, \( \pi_{\ell} \).

### 3 Main results

Our analysis and theoretical estimates in Sect. 4, and numerical experiments in Sect. 5 show that

1. For \( \ell = 1, \ldots, L \), and \( g_{\ell} := g \left( \bar{X}_{\ell} \right) \): the change of measure is performed at each time step, for \( 0 \leq n \leq N_{\ell} - 1 \), by going forward in time, and is only applied when

   (i) \( j \in J_{1} := \{1 \leq j \leq J; \ g(\bar{X} + v_{j}) \neq g(\bar{X}) \} \) &

   (ii) \( \Delta a_{\ell,n}^{j} \neq 0 \) & (iii) \( \Delta g_{\ell}(t_{n}) = 0 \)

2. If (3.1) is fulfilled: instead of using \( \Delta a_{\ell,n}^{j} \Delta t_{\ell} \) in (2.7), we propose to use \( \lambda_{\ell,n}^{j} \Delta t_{\ell} \), with \( \lambda_{\ell,n}^{j} \) is given by

\[
\lambda_{\ell,n}^{j} = c_{\ell} \Delta a_{\ell,n}^{j} = \Delta t_{\ell}^{-3} \Delta a_{\ell,n}^{j}, \quad 0 < \delta < 1,
\]

where \( \delta \) is a scale parameter in our IS algorithm.

3. We show that our proposed method (MLMC with IS) significantly reduces the kurtosis at the deep levels of MLMC (small \( \Delta t_{\ell} \)) (see Theorem 4.1 and the numerical experiments in Sect. 5), and also improves the strong convergence rate from \( \beta = 1 \), for the standard case (without IS), to \( \beta = 1 + \delta \), where \( 0 < \delta < 1 \) is a user-selected parameter in our IS algorithm (see Theorem 4.2, and the numerical experiments in Sect. 5). Due to Theorem 1.1, and given a pre-selected tolerance, TOL, this results in an improvement of the complexity from \( O \left( \text{TOL}^{-2} \log(\text{TOL}) \right) \), in the standard case, to the optimal complexity, i.e., \( O \left( \text{TOL}^{-3} \right) \). These improvements come with a negligible additional cost since we show in Sect. 4.4 that \( W_{\text{without IS}} \approx W_{\text{IS with sample}} \), \( W_{\text{sample}} \) denotes the average cost of simulating coupled MLMC paths at level \( \ell \). We show a summary of these results in Table 1; see Sects. 4 and 5 for more details.

### 4 Construction of the IS measure and convergence estimates

#### 4.1 Construction of the IS measure: the one-dimensional case

We start with the one-dimensional case (only one species), where the number of reactions is \( J = 1 \). Instead of using \( \Delta a_{\ell,n} \Delta t_{\ell} \) as the rate parameter of the Poisson rdvs used in each time step to update the states of the coupled paths (2.4), (2.5) where \( \Delta a_{\ell,n} \) is given by (2.7), we suggest using \( \lambda_{\ell,n} \Delta t_{\ell} \), with the parameter \( \lambda_{\ell,n} \), which will be determined given some constraints that we impose to ensure that our change of measure is (i) reducing the kurtosis of the MLMC estimator at the deep levels, (ii) reducing the variance of the MLMC levels and increasing the strong convergence rate. In the following, we denote \( g_{\ell} := g \left( \bar{X}_{\ell} \right) \), and \( \Delta g_{\ell} := g_{\ell} - g_{\ell-1} \), for \( \ell = 1, \ldots, L \).

The change of measure is performed at each time step by going forward in time, and is only applied when it is needed, i.e., we impose the following condition for applying the change of measure

\[
\Delta a_{\ell,n} \neq 0 \quad \& \quad \Delta g_{\ell}(t_{n}) = 0, \quad 0 \leq n \leq N_{\ell} - 1, \quad \ell = 1, \ldots, L.
\]
Condition (4.1) is motivated by the fact that (i) we need to change the measure only in cases where the coupled paths at the nth time step are equal and (ii) for cases where the rates of the Poisson rdvs are non zero, so we do not have the issue of the likelihood being equal to zero.

Whenever (4.1) holds, the change of measure is given by changing the rate of the Poisson rdvs [see (2.6) and (2.7)] in the tau-leap approximation from $\Delta a_{\ell,n} \Delta t_{\ell}$ to $\lambda_{\ell,n} \Delta t_{\ell}$. Hence, the conditional likelihood is then given by the ratio between the probability mass functions of two Poisson rdvs with rates $\Delta a_{\ell,n} \Delta t_{\ell}$ and $\lambda_{\ell,n} \Delta t_{\ell}$. Through a simple computation, this leads to

$$L_{\ell,n} = \frac{e^{-\Delta a_{\ell,n} \Delta t_{\ell}}}{e^{-\lambda_{\ell,n} \Delta t_{\ell}}} \left( \frac{\Delta a_{\ell,n}}{\lambda_{\ell,n}} \right)^{k_n}, \quad n \in I_{\ell}^n,$$

where $\lambda_{\ell,n}$ is the number of jumps that occurs at the nth time step where we apply the change of measure, and $I_{\ell}^n$ is the random set including the time steps at level $\ell$ where we simulate the Poisson rdvs under the new measure.

Thus, across one path, the likelihood ratio is given by

$$L_{\ell} = e^{-\Delta t_{\ell} \sum_{n \in I_{\ell}} (\Delta a_{\ell,n} - \lambda_{\ell,n})} \prod_{n \in I_{\ell}} \left( \frac{\Delta a_{\ell,n}}{\lambda_{\ell,n}} \right)^{k_n}.$$

Furthermore, if we impose that $\lambda_{\ell,n} = c_{\ell} \Delta a_{\ell,n}$ then we obtain

$$L_{\ell} = \left( e^{(c_{\ell} - 1) \Delta t_{\ell} \sum_{n \in I_{\ell}} \Delta a_{\ell,n}} \right)^{-\frac{1}{c_{\ell}}} \left( \sum_{n \in I_{\ell}} k_n \right)^{c_{\ell}}.$$

We note that imposing $\lambda_{\ell,n} = c_{\ell} \Delta a_{\ell,n}$ can be motivated by the fact that we want to keep the same physical structure of the rate of the Poisson process driving the state changes, i.e., depending on $\Delta a_{\ell,n}$. However, we try to introduce a scaled constant $c_{\ell}$ that depends on $\Delta t_{\ell}$ so that we reduce the probability of having $\Delta g_{\ell}(T) = 0$, under the new measure. A reasonable choice of $c_{\ell}$ is given by

$$c_{\ell} = \Delta t_{\ell}^{-\delta},$$

where $\delta > 0$ is the scale parameter to be determined. Note that the case $\delta = 0$ is similar to the case of using the old measure in all time steps.

Remark 4.1 Observe that for $\delta \in (0, 1)$ and $g(x) = x, x \in \mathbb{R}$, we have $\pi_{\ell} (|\Delta \bar{X}_{\ell}(T)| = 0)$ is still approaching 1 as $\Delta t_{\ell}$ decreases, but compared to the initial situation (without IS), we decrease the rate of convergence with respect to $\Delta t_{\ell}$ (compare Fig. 8, for the case without IS, and Fig. 17 for the case with IS with $\delta = \frac{3}{4}$).

### 4.2 Construction of the IS measure: the multi-channels and high dimensional states case

Extending our method to a higher dimension in the number of reaction channels, $J$, and in the state vector $\mathbf{X}$ is straightforward with slight modifications. We first define the set $J_1$ as

$$J_1 = \{1 \leq j \leq J; \quad g(x + v_j) \neq g(x)\}.$$

In the multi-channel case, we are only interested in changing the measure for reactions whose stoichiometric vector, $v_j$, changes the state of the quantity of interest, i.e., for reactions with index $j \in J_1$. In Algorithm 4.1, we summarize our methodology for simulating two coupled explicit TL paths with IS.

We consider a number of reactions $J > 1, \mathbf{X} \in \mathbb{N}^d$, $d \geq 1$ and $g : \mathbb{R}^d \to \mathbb{R}$. Hereafter, we denote by $\{v_{j,i}\}_{j=1}^d$ the coordinates in the stoichiometric vectors, $\{v_{j,i}\}_{j=1}^d$, corresponding to the state change of the ith species. For a fixed $0 \leq n \leq N_{\ell} - 1$, we define $F_n$ to be the sigma algebra given by

$$F_n := \sigma \left( (\Delta a_{\ell,k})_{j=1,\ldots,J;\ k=0,\ldots,n} \right), \quad n = 0, \ldots, N_{\ell} - 1.$$

---

**Table 1** Main results for the comparison of MLMC combined with our IS algorithm, and standard MLMC

| Quantity of interest | MLMC without IS (standard case) | MLMC with IS ($0 < \delta < 1$) |
|----------------------|---------------------------------|---------------------------------|
| $\kappa_{\ell}$      | $O \left( \Delta t_{\ell}^{-1} \right)$ | $O \left( \Delta t_{\ell}^{\delta - 1} \right)$ |
| $V_{\ell}$           | $O \left( \Delta t_{\ell} \right)$ | $O \left( \Delta t_{\ell}^{1+\delta} \right)$ |
| $W_{\ell, \text{sample}}$ | $\approx 2 \times J \times C_p \times \Delta t_{\ell}^{-1}$ | $\approx 2 \times J \times C_p \times \Delta t_{\ell}^{-1}$ |
| WorkMLMC             | $O \left( \text{TOL}^{-2} \log (\text{TOL})^2 \right)$ | $O \left( \text{TOL}^{-2} \right)$ |

$\kappa_{\ell}$ denotes the kurtosis of the coupled MLMC paths at level $\ell$. $V_{\ell}$ denotes the variance of the coupled MLMC paths at level $\ell$. $C_p$ is the cost of generating one Poisson rdv.
Algorithm 4.1 Simulates two coupled explicit TL paths with IS, and computes the likelihood ratio.

1: Fix $\Delta t_\ell > 0$ and set $\Delta t_{\ell - 1} = 2 \times \Delta t_\ell$.
2: Set $Z_\ell(0) = Z_{\ell - 1}(0) = \mathbf{0}, t_\ell = t_{\ell - 1} = 0, n = 0$.
3: Set $c_\ell = \Delta t_\ell^{-\delta}, \delta \in (0, 1)$
4: while $t_\ell < T$ do
5: $n = n + 1$
6: for $j = 1$ to $J_1$ do
7: if then $a_j(Z_{\ell}(t_\ell)) \neq a_j(Z_{\ell - 1}(t_{\ell - 1}))$ & $g(Z_{\ell}(t_\ell)) = g(Z_{\ell - 1}(t_{\ell - 1}))$
8: $A_{3}(j-1)+1 = \min (a_j(Z_{\ell}(t_\ell)), a_j(Z_{\ell - 1}(t_{\ell - 1})))$
9: $A_{3}(j-1)+2 = c_\ell a_j(Z_{\ell}(t_\ell)) - A_{3}(j-1)+1$
10: $A_{3}(j-1)+3 = c_\ell a_j(Z_{\ell - 1}(t_{\ell - 1})) - A_{3}(j-1)+1$
11: Compute $L_{\ell,n}$ using (4.7) and (4.8).
12: $L_{\ell} = L_{\ell} + L_{\ell}$ using (4.7) and (4.8).
13: else
14: $A_{3}(j-1)+1 = \min (a_j(Z_{\ell}(t_\ell)), a_j(Z_{\ell - 1}(t_{\ell - 1})))$
15: $A_{3}(j-1)+2 = a_j(Z_{\ell}(t_\ell)) - A_{3}(j-1)+1$
16: $A_{3}(j-1)+3 = a_j(Z_{\ell - 1}(t_{\ell - 1})) - A_{3}(j-1)+1$
17: $A_{3}(j-1)+4 = \text{Poisson} (A_{3}(j-1)+4)$
18: $A_{3}(j-1)+5 = \text{Poisson} (A_{3}(j-1)+5)$
19: State updating
20: if $(n \mod 2) = 0$ then $t_{\ell - 1} = t_{\ell - 1} + \Delta t_{\ell - 1}$
21: $t_\ell = t_\ell + \Delta t_\ell$

The likelihood ratio for each reaction channel $j \in J_1$ has a similar expression to (4.4), and is given by

$$L_{\ell}^{j} = \left( e^{(c_\ell - 1)\Delta t_\ell} \sum_{a \in T_\ell^{j}} k_{a}^{j} \right) \left( -\frac{\sum_{a \in T_\ell^{j}} k_{a}^{j}}{c_\ell} \right), \quad j \in J_1,$$

(4.7)

where $k_{a}^{j}$ is the number of jumps associated with the $j$th reaction channel that occurs at the nth step where we apply the change of measure, and $T_\ell^{j}$ is the random set including the time steps at level $\ell$, where we simulate the Poisson r.v.s under the new measure for the $j$th reaction channel.

Thus, across one path, the likelihood ratio is given by

$$L_{\ell} = \prod_{j \in J_1} L_{\ell}^{j} = \left( e^{(c_\ell - 1)\Delta t_\ell} \sum_{a \in T_\ell^{j}} k_{a}^{j} \right) \left( -\frac{\sum_{a \in T_\ell^{j}} k_{a}^{j}}{c_\ell} \right),$$

(4.8)

Similarly to Sect. 4.1, we choose $c_\ell$ to be given by (4.5), with $\delta > 0$. Remark 4.1 holds for the high dimensional case. In particular, compare Figs. 9 and 10, for the case without IS, and Figs. 18 and 19 for the case using IS with $\delta = \frac{1}{2}$, for $g(X) = X(i)$, i.e., the projection on the $i$th coordinate of the state vector $X$.

4.3 Convergence estimates of MLMC combined with IS

In this section, we aim to derive convergence estimates of the kurtosis and the variance. We start by stating the main two assumptions (Assumptions 4.1 and 4.2), needed to derive the main results in this section. For the ease of presentation, we consider $g(X) = X(i)$, the projection on the $i$th coordinate of the state vector $X$.

Assumption 4.1. For a small $\Delta t_\ell$, and conditioning on $F_{N_{\ell-1}}$ and $(\mathcal{T}_{\ell}^{j} = S_{j})_{j \in J_1}$, we denote, for $j \in J_1$, $K_j = \sum_{n \in S_{j}} k_{n}^{j}$, and we assume that, for $0 \leq \delta < 1$,

(a) for all $K \in \mathbb{N}^{J_1}$ such that $\sum_{j \in J_1} v_{j} K_{j} \neq 0$, we have

$$\pi_{\ell} \left( |\Delta g_{\ell}(T)| = |v_{j_{\ell}}|, \{K_{j} = 0\}_{j \in J_1} \right) \leq \xi_{\ell} \Delta t_{\ell}^{(1-\delta)q},$$

with $\xi_{\ell}$ independent of $\ell$ and $q$.

(b) for all $q \geq 1$, there exists $\eta_{q, \ell} > 0$ such that,

$$\pi_{\ell} \left( |\Delta g_{\ell}(T)| = |v_{j_{\ell}}|, \{K_{j} = 1\}_{j \in J_1} \right) \leq \eta_{q, \ell} \Delta t_{\ell}^{(1-\delta)q},$$

(c) for all $j \in J_1$, there exist a single $n_{j}^{*} \in S_{j}$ such that

$$\pi_{\ell} \left( |\Delta g_{\ell}(T)| = |v_{j_{\ell}}|, \{K_{j} = 1\}_{j \in J_1} \right) \leq \eta_{q, \ell} \Delta t_{\ell}^{(1-\delta)q} \left( \Delta t_{\ell}^{1-\delta} \Delta a_{n_{j}^{*}}^{j} \right) \right) \right) \right).$$

Assumption 4.2. There exists $\ell_{0} \geq 0$ such that for all $\ell \geq \ell_{0}$, we have

$$0 < C_{1} \leq \sum_{j \in J_1} E_{\pi_{\ell}} \left[ e^{-\Delta a_{n_{j}^{*}}^{j} \Delta a_{n_{j}^{*}}^{j}} \right] \leq C_{2} < \infty,$$

where $C_{1}, C_{2}$ are independent of $\ell$, and $n_{j}^{*}$ are defined in Assumption 4.1(c).

Assumption 4.1(a) can be interpreted by the fact that the probability of events where we observe $\Delta g_{\ell}(T) \neq 0$, due
to jumps simulated with the IS measure, are dominated by the ones where only one jump occurs under the new measure and due to the firing of only one reaction channel in $\mathcal{J}_1$. Moreover, the motivation behind Assumption 4.1(b) is that jumps observed in $\Delta g_1(T)$, are mainly caused by the IS measure. Furthermore, Assumption 4.1(c) is motivated by our numerical observations, which suggest that for small values of $\Delta t_\ell$, we sample at most one single step using our IS algorithm, which separates the two paths (see Figs. 20, 21 and 22 in “Appendix B”). Finally, by observing that $\Delta a_{\ell,n} = O(1)$, \( \forall j \in \mathcal{J}_1 \), Assumption 4.2 is motivated by our numerical observations (see Fig. 1a, b), which show that $E\left[ \sum_{j \in \mathcal{J}_1} I^2_{\ell,j} \right] = O(1)$.

Now, we state the main results of this section through Theorems 4.1 and 4.2. The proof of these theorems are identical to the one dimensional proofs [one species and one reaction channel] with slight differences. Consequently, for ease of presentation, we present in “Appendix A” the one dimensional proofs. The key result for these proofs is Lemma 4.1 which is proven in “Appendix A”. In the following, without loss of generality, we also assume that $|\nu| = 1$.

**Lemma 4.1** (Conditional $L^p$ moments estimates) Let $J = 1$, $p \geq 1$ and $0 \leq \delta < 1$, and suppose that Assumptions 4.1 and 4.2 hold, then, for $\Delta t_\ell \to 0$, we have

$$E_{\pi_\ell} \left[ |\Delta g_1|^{p} \left( T \right) L_\ell^{n} \right]:= \mathcal{F}_{\mathcal{N}_{t-1}, I_\ell = S} \right] = \Delta t^{(p-1)\delta+1}(\Delta t_{\ell,n}) e^{p(\Delta t^{1-\delta}-\Delta t_{\ell})} \sum_{n \in \mathcal{S}} \Delta a_{\ell,n} \left( 1 + h_{p,\ell} \right),$$

such that $h_{p,\ell} \to 0$.

**Remark 4.2** Note that for $J \geq 1$, Lemma 4.1 is extended to the multi-channel case by expressing the right-hand side of 4.9 as a summation over the set $\mathcal{J}_1$ of similar terms but involving $\Delta a_{\ell,n}^{j}$ instead of $\Delta a_{\ell,n}$. These terms correspond to only one jump occurring under the new measure and due to the firing of only one reaction channel $j \in \mathcal{J}_1$.

Finally a further assumption (Assumption 4.3) is needed to prove Theorems 4.1 and 4.2.

**Assumption 4.3** For a sufficiently large $\ell$, we assume that there exists a constant $C_p$, independent of $\ell$, such that $h_{p,\ell}$ in Lemma 4.1 fulfills $E_{\pi_\ell} \left[ h_{p,\ell}^2 \right] \leq C_p < \infty$.

Theorem 4.1 shows that the kurtosis at level $\ell$ of the MLMC estimator combined with our IS algorithm, $\kappa_\ell$, is $O \left( \Delta t^{\delta-1}_{\ell} \right)$.

**Theorem 4.1** Let $J \geq 1$ and let us denote $Y_\ell := \Delta g_\ell(T) L_\ell$. Suppose that Assumptions 4.1, 4.2 and 4.3 hold. Then, for $0 \leq \delta < 1$ and $\Delta t_\ell \to 0$, we have

$$\kappa_\ell := \frac{E_{\pi_\ell} \left[ (Y_\ell - E_{\pi_\ell} [Y_\ell])^4 \right]}{(\text{Var}_{\pi_\ell} [Y_\ell])^2} = O \left( \Delta t^{\delta-1}_{\ell} \right).$$

**Theorem 4.2** Let $0 \leq \delta < 1$ and $J \geq 1$ and let us denote $Y_\ell := \Delta g_\ell(T) L_\ell$. Suppose that Assumptions 4.1, 4.2 and 4.3 hold. Then, for $\Delta t_\ell \to 0$, we have

$$(\text{Var}_{\pi_\ell} [Y_\ell]) = O \left( \Delta t^{1+\delta}_{\ell} \right).$$

The result in Theorem 4.2 is confirmed by the numerical experiments in Sect. 5, which demonstrate that our IS algorithm improves the strong convergence rate from $\beta = 1$ (see Figs. 5, 6, and 7) to $\beta = 1 + \delta$ with $\delta > 0$ (see Figs. 11 and 12 for Example 5.1, Fig. 13 and 14 for Example 5.2, and Figs. 15 and 16 for Example 5.3) for the case with IS. Due to Theorem 1.1, and given that $\gamma = 1$ (work rate) for both cases, with and without IS, we improve the complexity of the MLMC method from $O \left( \text{TOL}^{-2} \log(\text{TOL})^2 \right)$ for the case without IS to the optimal complexity, i.e., $O \left( \text{TOL}^{-\delta} \right)$, for the case with IS, where TOL is a pre-selected tolerance.

**Remark 4.3** [More general observable $g$] For ease of presentation, we formulate our assumptions and show our proofs for an observable $g$ in the class of projections. However, our results can be easily extended to include linear maps, and linear combination of indicator functions.

### 4.4 Cost analysis

In this section, we analyze briefly the computational costs when using MLMC with our IS technique compared to standard MLMC, in the context of SRNs. Let $M_{\ell}$ denote the number of samples at level $\ell$, and $W_{\ell}$ the expected cost per sample path at level $\ell$. Observe that the expected computational cost of the MLMC estimator is given by

$$W_{\text{MLMC}} := \sum_{\ell=0}^{L} M_{\ell} W_{\ell, \text{sample}},$$

If we denote by $W_{\ell, \text{sample}}^{\text{with IS}}$ and $W_{\ell, \text{sample}}^{\text{without IS}}$ the expected costs of simulating one sample path at level $\ell$ with and without IS, respectively, then we have

$$W_{\ell, \text{sample}}^{\text{without IS}} \approx 2 \times J \times C_p \times \Delta t^{-1}_{\ell}.$$
With IS,$ \ell$, sample $\approx 2 \times J \times C_p \times \Delta t_{\ell}^{-1}$

$$+ \frac{C_{\text{lik}} \times \sum_{j \in J_1} \# I_{\ell,j}}{\ll C_p \times \Delta t_{\ell}^{-1}} \ll C_{\text{lik}}$$

where $C_p$ is the cost of generating one Poisson rdv, $C_{\text{lik}}$ is the cost of computing the likelihood ratio, and $\sum_{j \in J_1} \# I_{\ell,j}$ is the average number of time steps at level $\ell$, where we simulate under the new measure the $j$th reaction channel. We note that the inequality $\sum_{j \in J_1} \# I_{\ell,j} \ll \Delta t_{\ell}^{-1}$ is motivated and justified by the construction of our IS algorithm, where IS is only applied a few times across each simulated path. This is also confirmed by Fig. 1. Furthermore, we refer to Fig. 2 for evidence of the observation made by (4.11).

Furthermore, if we denote by $V_{\ell} = \text{Var}\left[g_{\ell} - g_{\ell-1}\right]$, then from our analysis in Sect. 4.2, and our numerical experiments in Sect. 5, it is shown that $V_{\ell}^{\text{with IS}} \ll V_{\ell}^{\text{without IS}}$ implying that $M_{\ell}^{\text{with IS}} \ll M_{\ell}^{\text{without IS}}$ (see Fig. 3).

Finally, from the previous analysis, we conclude that combining our pathwise IS with the MLMC estimator reduces the cost significantly. This gain is obtained by having the same cost per sample path per level as the standard MLMC estimator, and using less samples to achieve a certain error tolerance, TOL.
5 Numerical experiments

In the following, we illustrate the main benefits of the MLMC-based method, when used in combination with our IS algorithm explained in Sect. 4, compared to the standard MLMC used in Anderson and Higham (2012). We consider three different examples of SRNs, given by Examples 5.1, 5.2, and 5.3, where we use the MLMC method to estimate \( E[g(X(T))] \), where \( X \) is the state vector representing the counting number of each species in the system, \( g : \mathbb{R}^d \rightarrow \mathbb{R} \) is a given scalar observable of \( X \), and \( T > 0 \) is a user-selected final time. We note that our numerical results were obtained using an Intel(R) Xeon(R) CPU E5-2680 architecture. Furthermore, the computer code is written in the MATLAB programming language (version R2019a), and it can be downloaded from https://github.com/hammouc/MLMC_IS_SRNs.

Example 5.1 (Decay example) This model has one reaction,
\[
X \xrightarrow{\theta_1} 0,
\]
with \( \theta_1 = 1 \), \( T = 1 \), and \( X_0 = 10 \). The stoichiometric scalar \( \nu = -1 \) and the propensity function \( a(x) = \theta_1 x \). The quantity of interest in this example is \( E[X(T)] \).

Example 5.2 [Gene transcription and translation (Anderson and Higham 2012)] This model has five reactions,
\[
\begin{align*}
\emptyset & \xrightarrow{\theta_1} R, \quad R \xrightarrow{\theta_2} R + P \\
2P & \xrightarrow{\theta_3} D, \quad R \xrightarrow{\theta_4} \emptyset \\
P & \xrightarrow{\theta_5} \emptyset
\end{align*}
\]
with \( \theta = (25, 10^3, 0.001, 0.1, 1) \), \( T = 1 \), \( X(t) = (R(t), P(t), D(t)) \) and \( X_0 = (0, 0, 0) \). The stoichiometric matrix and the propensity functions are given by
\[
\nu = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & -2 & 1 \\
-1 & 0 & 0 \\
0 & -1 & 0
\end{pmatrix}, \quad a(X) = \begin{pmatrix}
\theta_1 \\
\theta_2 R \\
\theta_3 P(P - 1) \\
\theta_4 R \\
\theta_5 P
\end{pmatrix}
\]
The quantity of interest is \( E[X^{(1)}(T)] \). We note that the choice of \( X^{(1)} \) as the target species was determined by selecting the \( r \)th species with the highest probability of having \( \mathbf{X}_r^{(1)}(T) = \mathbf{X}_{r-1}^{(1)}(T) = 0 \) on the deep levels, resulting in the most severe catastrophic coupling explained in Sect. 1.4.1. In this example, the coarsest level of the MLMC estimator is \( L_0 = 2 \).

Example 5.3 (Michaelis-Menten enzyme kinetics Rao and Arkin 2003) The catalytic conversion of a substrate, \( S \), into a product, \( P \), via an enzymatic reaction involving enzyme, \( E \). This is described by Michaelis-Menten enzyme kinetics with three reactions,
\[
E + S \xrightarrow{\theta_1} C, \quad C \xrightarrow{\theta_2} E + S
\]
\[
C \xrightarrow{\theta_3} E + P,
\]
with \( \theta = (0.001, 0.005, 0.01), T = 1 \), \( X(t) = (E(t), S(t), C(t), P(t)) \) and \( X_0 = (100, 100, 0, 0) \). The stoichiometric matrix and the propensity functions are given by
\[
\nu = \begin{pmatrix}
-1 & -1 & 1 & 0 \\
1 & 1 & -1 & 0 \\
1 & 0 & -1 & 1
\end{pmatrix}, \quad a(X) = \begin{pmatrix}
\theta_1 ES \\
\theta_2 C \\
\theta_3 C
\end{pmatrix}
\]
The quantity of interest in this example is \( E[X^{(3)}(T)] \). We show in Table 2 the summarized results, related to the convergence rates, for the different scenarios with and without IS, and for the different examples that we consider in our numerical experiments. We also show several cases depending on the parameter, \( \delta \), used in the IS algorithm. From this table, we can see that our IS algorithm, besides dramatically reducing the kurtosis, improves the strong convergence rates from 1 to \( 1 + \delta \), which then improves the total complexity of the MLMC estimator from \( \mathcal{O}(\text{TOL}^{-2} \log(\text{TOL})^2) \) to \( \mathcal{O}(\text{TOL}^{-2}) \), where TOL is a pre-selected tolerance. This improvement is confirmed by Fig. 4, which shows that MLMC, when used in combination with our IS algorithm,

| Example | \( \alpha \) | \( \beta \) | \( \gamma \) | \( \kappa_L \) |
|---------|-------------|-------------|-------------|-------------|
| Example 5.1 without IS | 1.04 | 1.03 | 1 | 2191 |
| Example 5.1 with IS (\( \delta = 1/4 \)) | 1.04 | 1.27 | 1 | 275 |
| Example 5.1 with IS (\( \delta = 1/2 \)) | 1.04 | 1.57 | 1 | 34.2 |
| Example 5.1 with IS (\( \delta = 3/4 \)) | 1.04 | 1.93 | 1 | 5.1 |
| Example 5.2 without IS | 1 | 0.99 | 1 | 3290 |
| Example 5.2 with IS (\( \delta = 1/4 \)) | 1 | 1.23 | 1 | 409 |
| Example 5.2 with IS (\( \delta = 1/2 \)) | 1 | 1.47 | 1 | 50 |
| Example 5.2 with IS (\( \delta = 3/4 \)) | 1 | 1.72 | 1 | 5.8 |
| Example 5.3 without IS | 1.02 | 1.03 | 1 | 1130 |
| Example 5.3 with IS (\( \delta = 1/4 \)) | 1.02 | 1.26 | 1 | 208 |
| Example 5.3 with IS (\( \delta = 1/2 \)) | 1.02 | 1.5 | 1 | 36.7 |
| Example 5.3 with IS (\( \delta = 3/4 \)) | 1.03 | 1.75 | 1 | 5.9 |
achieves the same numerical complexity, $O(TOL^{-2})$, as MC with an exact method (SSA), but with a significantly smaller constant. On the other hand, MLMC in combination with IS algorithm significantly outperforms standard MLMC.

5.1 Numerical results of MLMC without IS

In Figs. 5, 6 and 7, we show the convergence plots for the MLMC method without IS for Examples 5.1, 5.2 and 5.3, respectively. In these figures, and specifically from the right plot in the second row, we can see that for deep levels of MLMC, the kurtosis increases dramatically with respect to level $\ell$ of the MLMC method. This poor behavior of the kurtosis is mainly due to the catastrophic coupling issue (explained in Sect. 1.4.1), as illustrated by Figs. 8, 9 and 10.
Fig. 5 MLMC without IS for Example 5.1: Convergence plots with $g_\ell = \bar{X}_\ell(T)$

Fig. 6 Convergence plots of MLMC without IS for Example 5.2
Fig. 7 Convergence plots of MLMC without IS for Example 5.3

Fig. 8 Example 5.1 without IS: Histogram of $g_\ell - g_{\ell-1}$ ($g_\ell = \bar{X}_\ell(T)$), for number of samples $M_\ell = 10^5$. The proportion of samples $\{g_\ell - g_{\ell-1} = 0\}$ is an increasing function of the level, $\ell$, of the MLMC estimator, to reach almost 100% for $\ell = 13$. a $\ell = 5$. b $\ell = 13$

Fig. 9 Example 5.2 without IS: Histogram of $g_\ell - g_{\ell-1}$ ($g_\ell = \bar{X}_\ell^{(1)}(T)$), for number of samples $M_\ell = 10^5$. The proportion of samples $\{g_\ell - g_{\ell-1} = 0\}$ is an increasing function of the level, $\ell$, of the MLMC estimator, to reach almost 100% for $\ell = 11$. a $\ell = 3$. b $\ell = 11$
5.2 Numerical results of MLMC with IS

The MLMC estimator in combination with IS reduces the kurtosis significantly and improves the strong convergence rate from 1 to $1 + \delta$, as illustrated by Figs. 11, and 12 for Example 5.1, Figs. 13, 14 for Example 5.2, and Figs. 15, and 16 for Example 5.3. The notable reduction of the kurtosis is mainly due to the small reduction of the proportion of identical terminal values, $g_\ell$ and $g_{\ell-1}$, after using IS, as can be seen in Figs. 17, 18 and 19.

![Fig. 10](image1)

**Fig. 10** Example 5.3 without IS: Histogram of $g_\ell - g_{\ell-1}$ ($g_\ell = \mathbb{E}_T^{(3)}(T)$), for number of samples $M_\ell = 10^5$. The proportion of samples $\{g_\ell - g_{\ell-1} = 0\}$ is an increasing function of the level, $\ell$, of the MLMC estimator, to reach almost 100% for $\ell = 11$. a $\ell = 3$. b $\ell = 11$.

![Fig. 11](image2)

**Fig. 11** Convergence plots of MLMC with IS ($\delta = 1/2$) for Example 5.1.
Fig. 12 Convergence plots of MLMC with IS ($\delta = 3/4$) for Example 5.1

Fig. 13 Convergence plots of MLMC with IS ($\delta = 1/2$) for Example 5.2

Fig. 14 Convergence plots of MLMC with IS ($\delta = 3/4$) for Example 5.2
Fig. 15 Convergence plots of MLMC with IS ($\delta = 1/2$) for Example 5.3

Fig. 16 Convergence plots of MLMC with IS ($\delta = 3/4$) for Example 5.3

Fig. 17 Example 5.1 with IS ($\delta = 3/4$): Histogram of $g_\ell - g_{\ell-1}$ ($g_\ell = X_\ell(T)$), for number of samples $M_\ell = 10^5$. Our IS reduces the proportion of samples ($g_\ell - g_{\ell-1} = 0$) (compared to the case without IS; see Fig. 8) to reach around 80% for $\ell = 13$. a $\ell = 5$. b $\ell = 13$
Fig. 18 Example 5.2 with IS ($\delta = \frac{3}{4}$): Histogram of $g_\ell - g_{\ell-1}$ ($g_\ell = X^{(1)}_\ell(T)$), for number of samples $M_\ell = 10^5$. Our IS reduces the proportion of samples $\{g_\ell - g_{\ell-1} = 0\}$ (compared to the case without IS; see Fig. 9) to reach around 90% for $\ell = 11$. a $\ell = 3$. b $\ell = 11$

Fig. 19 Example 5.3 with IS ($\delta = \frac{3}{4}$): Histogram of $g_\ell - g_{\ell-1}$ ($g_\ell = X^{(3)}_\ell(T)$), for number of samples $M_\ell = 10^5$. Our IS reduces the proportion of samples $\{g_\ell - g_{\ell-1} = 0\}$ (compared to the case without IS; see Fig. 10) to reach around 90% for $\ell = 11$. a $\ell = 3$. b $\ell = 11$

6 Conclusions and future work

In the work presented here, we address the high-kurtosis phenomenon related to catastrophic coupling, and observed in MLMC estimators when applied in the context of SRNs and pure jumps. We propose a novel path-dependent IS algorithm to be used with MLMC, in order to improve robustness and computational performance.

Our theoretical results and numerical experiments show that our proposed method not only improves the robustness of the multilevel estimator by dramatically reducing the kurtosis, but also improves the strong convergence rate, which results in an improvement of the complexity of the MLMC method, from $O(TOL^{-2}\log(TOL)^2)$ to $O(TOL^{-2})$, with TOL being a pre-selected tolerance. We achieve all these improvements with a negligible additional cost since our IS algorithm is only applied a few times across each simulated path.

Here, we limit ourselves to the use of the IS technique with an explicit TL scheme. In a future study, we intend to investigate the potential of our proposed algorithm when using a split-step implicit TL scheme, as proposed in Hammouda et al. (2017), which is required for systems with the presence of slow and fast timescales (stiff systems). To overcome the catastrophic coupling issue, the authors in Hammouda et al. (2017) used extrapolation to estimate the sample variance when using MLMC. We believe that our new IS technique may help to obtain accurate estimates of the sample variances needed by the MLMC estimator. Another potential research direction may be to investigate a more optimal IS scheme to be used for MLMC; for instance, we may try to use a hierarchy of $\delta_\ell$, where the parameter $\delta$ used in our proposed method would depend on the level of discretization. Furthermore, we may explore the possibility of introducing a new IS scheme for MLMC based on SPM coupling, to address the catastrophic decoupling issue, which is the second cause of the high-kurtosis phenomenon in the context of SRNs when using MLMC. Finally, we can combine the strengths of our method and the hybrid approach in Moraes et al. (2016b) to improve the performance of the MLMC estimator.

Acknowledgements This work was supported by the KAUST Office of Sponsored Research (OSR) under Award No. URF/1/2584-01-01 and the Alexander von Humboldt Foundation. C. Ben Hammouda and R. Tempone are members of the KAUST SRI Center for Uncertainty Quantification in Computational Science and Engineering. The authors would like to thank Dr. Alvaro Moraes and Sophia Franziska Wiechert for their helpful and constructive comments. The authors are also very grateful to the anonymous referees for their valuable comments and suggestions that greatly contributed to shape the final version of the work.
Appendix A: Proofs of Lemma 4.1 and Theorems 4.1 and 4.2

Proof of Lemma 4.1 We denote by \(K = \sum_{n \in S} k_n\), \(L_J(j)\) the likelihood evaluated at \(K = j\). Then, for \(p \geq 1\) and \(0 \leq \delta < 1\), and using relation (4.4), we write

\[
E_{\pi_t}\left[|\Delta g_t|^p (T) L^P_t; (F_{N_t-1}, T^t_t = S)\right] = \sum_{|\Delta g_t(T)| = K = i} i^p L^P_t(i) \pi_t\left(|\Delta g_t(T)| = i, K = i; (F_{N_t-1}, T^t_t = S)\right) + \sum_{|\Delta g_t(T)| = j, K = i} i^p L^P_t(j) \pi_t\left(|\Delta g_t(T)| = j, K = j; (F_{N_t-1}, T^t_t = S)\right)
\]

Then, using Assumptions 4.1(b) and 4.1(c), we obtain

\[
0 \leq \frac{\sum_{|\Delta g_t(T)| = K = i} A_i}{A_1} \leq \sum_{i \in N(0,1)} i^p \Delta t^{(i-1)p\delta_0}.
\]

(4.2)

Now, let us examine the second sum in the right-hand side of (A.1). First, observe that \(B_{0j} = 0, \forall j \geq 1\) and \(B_{00} = 0, \forall i \geq 1\). Although the first observation is clear, we need to explain the second observation, which is mainly due to the fact that \(\pi_t\left(|\Delta g_t(T)| = i, K = 0; (F_{N_t-1}, T^t_t = S)\right) = 0, \forall i \geq 1\). For the purpose of simplification, let us consider \(g_t = X_t\); then considering the first interval in the coarse level, and using the coupling equation (2.6), we have: (i) At \(t = 0:\ X_t(0) = X_{t-1}(0)\) and \(\Delta a_{t-1,0}^0 = 0\). (ii) At \(t = \Delta t:\ X_{t} = X_t - (\Delta t)\) and \(\Delta a_{t-1,0}^0 = a(X_{t-1}(0)) - a(\Delta t)\). (iii) At \(t = t_1 = 2\Delta t\): if \(\Delta a_{t-1,0}^0 = 0\), then we simulate this step under the old measure and consequently we will have \(\bar{X}_t(1) = \bar{X}_{t-1}(1)\) otherwise if \(\Delta a_{t-1,0}^0 \neq 0\), then we simulate this step under the IS measure, but since \(j = 0\), then we will have \(\bar{X}_t(1) = \bar{X}_{t-1}(1)\). Therefore, in both scenarios, we will have the same situation at the start, \(t_0 = 0\). Therefore, we conclude that \(\pi_t\left(|\Delta g_t|^p (T) L^P_t; (F_{N_t-1}, T^t_t = S)\right) = 0, \forall i \geq 1\) and \(B_{0i} = 0, \forall i \geq 1\).

\[
\square
\]

Proof of Theorem 4.1 Let \(0 \leq \delta < 1\). In the first step of the proof, we want to show that

\[
\kappa_t := \frac{E_{\pi_t}\left[(Y_t - E_{\pi_t}[Y_t])^4\right]}{(\text{Var}_{\pi_t}[Y_t])^2} \sim \frac{E_{\pi_t}[Y_t^4]}{(\text{Var}_{\pi_t}[Y_t])^2} \Delta t \to 0
\]

Let us first show that \(\text{Var}_{\pi_t}[Y_t] \sim \frac{E_{\pi_t}[Y_t^2]}{E_{\pi_t}[Y_t]}\). In fact,

\[
\frac{\text{Var}_{\pi_t}[Y_t]}{E_{\pi_t}[Y_t]^2} = \frac{E_{\pi_t}[Y_t^2] - (E_{\pi_t}[Y_t])^2}{E_{\pi_t}[Y_t^2]} = 1 - \frac{(E_{\pi_t}[Y_t])^2}{E_{\pi_t}[Y_t^2]}.
\]
Therefore, we need to show that $I_1 := \frac{(E_{\pi}(Y_1))^2}{E_{\pi}(Y_1^2)} \Delta t \to 0$.

Due to the order one weak error convergence, there exists a constant $d_1 > 0$ such that $(E_{\pi}(Y_1)) \leq d_1 \Delta t$. Therefore, using Lemma 4.1 and Assumption 4.2, we obtain

$$0 \leq I_1 \leq \frac{d_1^2 \Delta t^2}{E_{\pi}(Y_1^2)} \leq \frac{d_1^2 \Delta t^2}{E_{\pi}(Y_1)^2} \leq \frac{d_1^2 \Delta t^2}{E_{\pi}(Y_1)^2} \leq \frac{d_1^2 \Delta t^2}{C_1} \Delta t \to 0.$$

Therefore, we conclude that

$$\text{Var}_{\pi}(Y_1) \sim \text{E}_{\pi}[Y_1^2]. \quad (A.4)$$

$$0 \leq I_2 \leq \frac{d_1 \Delta t E_{\pi}[Y_1^3]}{E_{\pi}(Y_1^2)} \leq \frac{d_1 \Delta t E_{\pi}[Y_1^3]}{E_{\pi}(Y_1)^2} \leq \frac{d_1 \Delta t E_{\pi}[Y_1^3]}{E_{\pi}(Y_1)^2} \leq \frac{d_1 \Delta t E_{\pi}[Y_1^3]}{C_1} \Delta t \to 0.$$

Now, let us show that $E_{\pi}(Y_t - E_{\pi}(Y_t))^4 \sim E_{\pi}(Y_t^4)$. In fact,

$$E_{\pi}(Y_t - E_{\pi}(Y_t))^4 = E_{\pi}(Y_t^4) - 4E_{\pi}(Y_t^3) E_{\pi}(Y_t) + 6E_{\pi}(Y_t^2) (E_{\pi}(Y_t))^2 - 3(E_{\pi}(Y_t))^4$$

Finally, for $I_3$, using Lemma 4.1 and Assumptions 4.2 and 4.3, we obtain

$$0 \leq I_3 \leq \frac{d_1^4 \Delta t^4}{E_{\pi}(Y_1^2)} \leq \frac{d_1^4 \Delta t^4}{E_{\pi}(Y_1)^2} \leq \frac{d_1^4 \Delta t^4}{C_1} \Delta t \to 0.$$
Finally, using (A.4), (A.5), Lemma 4.1 and Assumptions 4.2 and 4.3, we obtain

\[
0 \leq I_3 \leq \frac{d_t^2 \Delta t^2 E_{\pi_t}[|Y_t|^2]}{E_{\pi_t}[Y_t^2]}
\]

\[
= \frac{d_t^2 \Delta t^2 E_{\pi_t}[|Y_{t+1}^2|]}{E_{\pi_t}[Y_t^2]} \left[ \frac{\left( \Delta t_{\ell}^{1+\delta} \right)}{\Delta t_{\ell}} e^{2(\Delta t_{\ell}^{1+\delta} - \Delta t_{\ell})} \sum_{n \in S} \Delta a_{t,n} e^{-\left( \Delta t_{\ell}^{1+\delta} - \Delta a_{t,n} \right)} (\Delta a_{t,n}^*) (1 + h_{2,\ell}) \right]
\]

\[
= \frac{\left( \Delta t_{\ell}^{1+\delta} \right)}{\Delta t_{\ell}} e^{2(\Delta t_{\ell}^{1+\delta} - \Delta t_{\ell})} \sum_{n \in S} \Delta a_{t,n} \Delta a_{t,n}^* (1 + h_{4,\ell})
\]

\[
\leq \tilde{C} \Delta t_{\ell}^{2(1-\delta) \to 0} 0.
\]

Therefore, we conclude that

\[
E_{\pi_t}[\left( Y_{t+1} - E_{\pi_t}[Y_{t+1}] \right)^4] \sim E_{\pi_t}[Y_{t+1}^4]. \tag{A.5}
\]

Finally, using (A.4), (A.5), Lemma 4.1 and Assumptions 4.2 and 4.3, we obtain

\[
\chi_t := E_{\pi_t}[\left( Y_{t+1} - E_{\pi_t}[Y_{t+1}] \right)^4] \sim E_{\pi_t}[Y_{t+1}^4]
\]

\[
= E_{\pi_t}[Y_{t+1}^2(T); (\mathcal{F}_{k_{t+1}}, \mathcal{T}_t = S)]
\]

\[
= \frac{E_{\pi_t}[Y_{t+1}^2(T); (\mathcal{F}_{k_{t+1}}, \mathcal{T}_t = S)]}{E_{\pi_t}[Y_{t+1}^2(T); (\mathcal{F}_{k_{t+1}}, \mathcal{T}_t = S)]}
\]

\[
= \Delta t_{\ell}^{1+\delta} \left[ \frac{e^{2(\Delta t_{\ell}^{1+\delta} - \Delta t_{\ell})} \sum_{n \in S} \Delta a_{t,n} e^{-\left( \Delta t_{\ell}^{1+\delta} - \Delta a_{t,n} \right)} \Delta a_{t,n}^* (1 + h_{2,\ell})}{E_{\pi_t}[Y_{t+1}^2(T); (\mathcal{F}_{k_{t+1}}, \mathcal{T}_t = S)]]} \right]^2
\]

\[
= \mathcal{O} \left( \Delta t_{\ell}^{1+\delta} \right). \tag{A.6}
\]

Proof of Theorem 4.2 Let 0 < \delta < 1. Then, using (A.4), Lemma 4.1 and Assumptions 4.2 and 4.3, we obtain

\[
\text{Var}_{\pi_t}[Y_t] \sim E_{\pi_t}[Y_t^2] = E_{\pi_t}[Y_t^2; (\mathcal{F}_{k_{t+1}}, \mathcal{T}_t = S)]
\]

\[
= \left[ e^{2(\Delta t_{\ell}^{1+\delta} - \Delta t_{\ell})} \sum_{n \in S} \Delta a_{t,n} e^{-\left( \Delta t_{\ell}^{1+\delta} - \Delta a_{t,n} \right)} \Delta a_{t,n}^* (1 + h_{2,\ell}) \right]
\]

\[
= \mathcal{O} \left( \Delta t_{\ell}^{1+\delta} \right). \tag{A.7}
\]

Appendix B: Numerical evidence of Assumption 4.1

In Figs. 20, 21 and 22, we plot the histograms, for Examples 5.1, 5.2 and 5.3, with \delta = 0.5, corresponding to #IS steps : s.t. \mathcal{F}_t := \sum_{j \in \mathcal{J}_t} \sum_{n \in \mathcal{S}_j} k_{h} > 0), that is the number of times where we perform IS and succeeded to separate the two paths. These Figs. show that our assumption 4.1 (c) is valid since for small values of \Delta t_{\ell}, we have at most one jump created by IS such that it separates the two paths.

Fig. 20 Example 5.1 with IS (with \delta = 0.5): Histogram of #\{n \in \mathcal{S} : k > 0\}, for number of samples M_{\ell} = 10^5. a \ell = 6. b \ell = 10
Fig. 21 Example 5.2 with IS (with \( \delta = 0.5 \)): Histogram of \( \# \{ \text{IS steps : s.t. } K := \sum_{j \in J_1} \sum_{n \in S_j} k_n > 0 \} \), for number of samples \( M_\ell = 10^5 \). a \( \ell = 4 \), b \( \ell = 8 \).

Fig. 22 Example 5.3 with IS (with \( \delta = 0.5 \)): Histogram of \( \# \{ \text{IS steps : s.t. } K := \sum_{j \in J_1} \sum_{n \in S_j} k_n > 0 \} \), for number of samples \( M_\ell = 10^3 \). a) \( \ell = 6 \), b) \( \ell = 10 \).

References

Abdulle, A., Yu Cheng, H., Li, T.: Chebyshev methods with discrete noise: the \( \tau \)-rock methods. J. Comput. Math. (2010). https://doi.org/10.4208/jcm.2009.10-m1004

Ahn, T.-H., Sandu, A., Han, X.: Implicit simulation methods for stochastic chemical kinetics (2013). arXiv preprint arXiv:1303.3614

Anderson, D., Higham, D.: Multilevel Monte Carlo for continuous Markov chains, with applications in biochemical kinetics. SIAM Multiscal Model. Simul. 10(1), 146–179 (2012)

Anderson, D.F.: A modified next reaction method for simulating chemical systems with time dependent propensities and delays. J. Chem. Phys. 127(21), 214107 (2007)

Anderson, D.F., Ganguly, A., Kurtz, T.G., et al.: Error analysis of tau-leap simulation methods. Ann. Appl. Probab. 21(6), 2226–2262 (2011)

Anderson, D.F., Kurtz, T.G.: Stochastic Analysis of Biochemical Systems, vol. 1. Springer, Berlin (2015)

Aparicio, J.P., Solari, H.G.: Population dynamics: Poisson approximation and its relation to the langevin process. Phys. Rev. Lett. 86(18), 4183 (2001)

Bayer, C., Hammouda, C.B., Tempone, R.: Numerical smoothing and hierarchical approximations for efficient option pricing and density estimation (2020). arXiv preprint arXiv:2003.05708

Brauer, F., Castillo-Chavez, C.: Mathematical Models in Population Biology and Epidemiology, vol. 40. Springer, Berlin (2001)

Cao, Y., Petzold, L.: Trapezoidal tau-leaping formula for the stochastic simulation of biochemical systems. In: Proceedings of Foundations of Systems Biology in Engineering (FOSBE 2005), pp. 149–152 (2005)

Cao, Y., Liang, J.: Adaptively biased sequential importance sampling for rare events in reaction networks with comparison to exact solutions from finite buffer dCME method. J. Chem. Phys. 139(2), 07B605_1 (2013)

Çınlar, E.: Probability and Stochastics, vol. 261. Springer, Berlin (2011)

Cliffe, K.A., Giles, M.B., Scheichl, R., Teckentrup, A.L.: Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. Comput. Vis. Sci. 14(1), 3 (2011)

Daigle Jr, B.J., Roh, M.K., Gillespie, D.T., Petzold, L.R.: Automated estimation of rare event probabilities in biochemical systems. J. Chem. Phys. 134(4), 01B628 (2011)

Duffie, D., Glynn, P., et al.: Efficient Monte Carlo simulation of security prices. Ann. Appl. Probab. 5(4), 897–905 (1995)

Engblom, S.: On the stability of stochastic jump kinetics (2012). arXiv preprint arXiv:1202.3892

Ethier, S.N., Kurtz, T.G.: Markov Processes: Characterization and Convergence. Wiley Series in Probability and Mathematical Statistics. Wiley, New York (1986)

Giles, M.B.: Multilevel Monte Carlo path simulation. Oper. Res. 56(3), 607–617 (2008)

Giles, M.B.: Multilevel Monte Carlo methods. Acta Numerica 24, 259–328 (2015)

Gillespie, D.T.: A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comput. Phys. 22(4), 403–434 (1976)
Gillespie, D.T.: The chemical Langevin equation. J. Chem. Phys. 113(1), 297–306 (2000)

Gillespie, D.T.: Approximate accelerated stochastic simulation of chemically reacting systems. J. Chem. Phys. 115(4), 1716–1733 (2001). Please provide complete details for reference Gou (2016)

Gou, W.: Estimating value-at-risk using multilevel Monte Carlo maximum entropy method. Master’s thesis, University of Oxford (2016)

Gupta, A., Briat, C., Khammash, M.: A scalable computational framework for establishing long-term behavior of stochastic reaction networks. PLoS Comput. Biol. 10(6), e1003669 (2014)

Hammouda, C.B., Moraes, A., Tempone, R.: Multilevel hybrid split-step implicit tau-leap. Numer. Algorithms 74(2), 527–560 (2017)

Hensel, S.C., Rawlings, J.B., Yin, J.: Stochastic kinetic modeling of vesicular stomatitis virus intracellular growth. Bull. Math. Biol. 71(7), 1671–1692 (2009)

Kebaier, A., et al.: Statistical Romberg extrapolation: a new variance reduction method and applications to option pricing. Ann. Appl. Probab. 15(4), 2681–2705 (2005)

Kuwahara, H., Mura, I.: An efficient and exact stochastic simulation method to analyze rare events in biochemical systems. J. Chem. Phys. 129(16), 10B619 (2008)

Lester, C., Baker, R.E., Giles, M.B., Yates, C.A.: Extending the multi-level method for the simulation of stochastic biological systems. Bull. Math. Biol. 78(8), 1640–1677 (2016)

Lester, C., Yates, C.A., Baker, R.E.: Robustly simulating biochemical reaction kinetics using multi-level Monte Carlo approaches. J. Comput. Phys. 375, 1401–1423 (2018)

Lester, C., Yates, C.A., Giles, M.B., Baker, R.E.: An adaptive multi-level simulation algorithm for stochastic biological systems. J. Chem. Phys. 142(2), 02B612_1 (2015)

Li, T.: Analysis of explicit tau-leaping schemes for simulating chemically reacting systems. Multiscale Model. Simul. 6(2), 417–436 (2007)

Moraes, A., Tempone, R., Vilanova, P.: A multilevel adaptive reaction-splitting simulation method for stochastic reaction networks. SIAM J. Sci. Comput. 38(4), A2091–A2117 (2016)

Moraes, A., Tempone, R., Vilanova, P.: Multilevel hybrid Chernoff tau-leap. BIT Numer. Math. 56(1), 189–239 (2016)

Rao, C.V., Arkin, A.P.: Stochastic chemical kinetics and the quasi-steady-state assumption: application to the Gillespie algorithm. J. Chem. Phys. 118(11), 4999–5010 (2003)

Rathinam, M.: Moment growth bounds on continuous time Markov processes on non-negative integer lattices (2013). arXiv preprint arXiv:1304.5169

Rathinam, M., El Samad, H.: Reversible-equivalent-monomolecular tau: a leaping method for “small number and stiff” stochastic chemical systems. J. Comput. Phys. 224(2), 897–923 (2007)

Srivastava, R., You, L., Summers, J., Yin, J.: Stochastic vs. deterministic modeling of intracellular viral kinetics. J. Theor. Biol. 218(3), 309–321 (2002)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.