Revisiting the Effects of Stochasticity for Hamiltonian Samplers

Giulio Franzese [1] **Dimitrios Milios** [1] **Maurizio Filippone** [1] **Pietro Michiardi** [1]

Abstract

We revisit the theoretical properties of Hamiltonian stochastic differential equations (SDEs) for Bayesian posterior sampling, and we study the two types of errors that arise from numerical SDE simulation: the discretization error and the error due to noisy gradient estimates in the context of data subsampling. Our main result is a novel analysis for the effect of mini-batches through the lens of differential operator splitting, revising previous literature results. The stochastic component of a Hamiltonian SDE is decoupled from the gradient noise, for which we make no normality assumptions. This leads to the identification of a convergence bottleneck: when considering mini-batches, the best achievable error rate is $\mathcal{O}(\eta^2)$, with η being the integrator step size. Our theoretical results are supported by an empirical study on a variety of regression and classification tasks for Bayesian neural networks.

1. Introduction

Hamiltonian Monte Carlo (HMC) is a popular approach to obtain samples from intractable distributions (Neal, 1996; 2011; Hoffman & Gelman, 2014). It presents, however, significant computational challenges for large datasets, as it requires access to the full gradient of the associated Hamiltonian system. Stochastic-Gradient HMC (SGHMC) (Chen et al., 2014) was proposed as a scalable alternative to HMC, by admitting noisy estimates of the gradient using mini-batching. A complementary family of techniques relies on pseudo-marginal techniques to achieve scalability (Alenlöv et al., 2021). In SGHMC, the Hamiltonian dynamics is modified so as to include a friction term that counteracts the effects of the gradient noise. This approach has proven effective in dealing with the difficulties in sampling from the posterior distribution over model parameters of Bayesian Neural/Convolutional Networks (BNNs) (Wenzel et al., 2020; Tran et al., 2022).

Stochastic gradient (SG) methods have been extensively studied as a means for Markov chain Monte Carlo (MCMC)-based algorithms to scale to large data. Variants of SG-MCMC algorithms have been studied through the lenses of first (Welling & Teh, 2011; Ahn et al., 2012; Patterson & Teh, 2013) or second-order (Chen et al., 2014; Ma et al., 2015) Langevin dynamics; these are mathematically convenient continuous-time processes which correspond to discrete-time gradient methods with and without momentum, respectively. Langevin dynamics are formally captured by an appropriate set of stochastic differential equations (SDEs), whose theoretical properties have been extensively studied (Kloeden & Platen, 1995; Debussche & Faou, 2012) with a particular emphasis on the stationary property of these processes (Abdulle et al., 2014; Milstein & Tretyakov, 2007). As in Abdulle et al. (2014; 2015), we are interested in the asymptotic (in time) performance of such sampling schemes. The reader is referred to Vollmer et al. (2016); Gao et al. (2018; 2020); Futami et al. (2020); Xu et al. (2018) for additional insights into the non-asymptotic behavior of these methods.

In this work, we seek to re-evaluate the connections between SG and stochastic Hamiltonian dynamics (known as second order or underdamped Langevin dynamics) with the aim of improving our current understanding of the goodness of sampling from intractable distributions when considering mini-batching. We consider a system with potential $U(\boldsymbol{\theta})$ which is the negative of the logarithm of the density function associated with the distribution we aim to sample from. We introduce position variables $\boldsymbol{\theta} \in \mathbb{R}^d$ (i.e., parameters) and momentum variables $\mathbf{r} \in \mathbb{R}^d$ obeying the following SDE:

$$\begin{aligned} d\mathbf{r}(t) &= -\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}(t))dt - C\mathbf{M}^{-1}\mathbf{r}(t)dt + \sqrt{2C}d\mathbf{w}(t) \\ d\boldsymbol{\theta}(t) &= \mathbf{M}^{-1}\mathbf{r}(t)dt. \end{aligned} \tag{1}$$

This is an extension of an Hamiltonian system with a friction term and an appropriately scaled Brownian motion $\mathbf{w}(t)$, where $C > 0$[1] and $\mathbf{M}$ is a symmetric, positive definite matrix (a.k.a. mass matrix). A common assumption in the

[1]Data Science Department, EURECOM, France. Correspondence to: Giulio Franzese <giulio.franzese@eurecom.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

[1]Here for simplicity we consider $C \in \mathbb{R}$, but in general C can be a matrix.

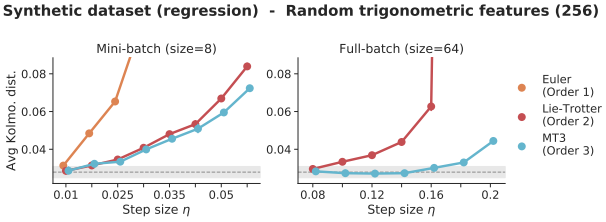

Figure 1. We evaluate integrators of different order based on the distance from the true posterior distribution for a Gaussian-linear system. Left: for mini-batch size equal to 8, there is an area for which the 3-rd order integrator does not perform any better than an integrator of order two, while these both perform better than an integrator of order one. Right: There is no such bottleneck for the full-batch version. The orange line (order one) extends beyond the limits of this plot.

literature (Chen et al., 2014; Ahn et al., 2012; Ma et al., 2015), is that the noise associated with stochastic estimates of $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}(t))$ is normally distributed. This noise is linked with the additive Brownian motion term in Equation (1).

In this work, we challenge this assumption and we advocate that the SG noise should be decoupled from the SDE dynamics. More specifically, we show that the Brownian motion term is not a good model for the stochasticity of the gradient introduced by the mini-batching scheme (see also Appendix A for an extended discussion). In this sense, our framework is similar to Chen et al. (2015); however we propose an interpretation of the effect of mini-batching through the lenses of differential operator splitting, by leveraging the huge literature concerning the simulation of high dimensional Hamiltonian systems (Childs et al., 2021; Suzuki, 1977; Hatano & Suzuki, 2005; Childs & Su, 2019; Low et al., 2019). Earlier attempts (Betancourt, 2015; Shahbaba et al., 2014) to use operator splitting as a tool to describe mini-batching focus on HMC; Cobb & Jalaian (2021) implement a symmetric splitting scheme w.r.t. the data, but again only for classical HMC, and with no theoretical analysis. Recently Zou & Gu (2021) have studied non-asymptotic convergence for HMC when considering mini-batches, but these results are limited to strongly convex potentials. Regarding Hamiltonian SDE dynamics, operator splitting scheme have been used before (Leimkuhler & Shang, 2016a; Monmarché, 2021), but not with the purpose of modeling the effect of mini-batching, to the best of our knowledge.

Our main contribution can be summarized as the derivation of new convergence results considering the two types of error induced by the simulation of the proposed SDE scheme: the discretization and the SG errors. The discretization of SDEs has been extensively studied in the literature (Kloeden & Platen, 1995; Debussche & Faou, 2012; Abdulle et al., 2014; Milstein & Tretyakov, 2007). In Section 2, we leverage these results to present quantitatively precise claims about sampling schemes in the context of Bayesian inference for modern machine learning problems. In Section 3, our treatment of mini-batches in terms of differential operator splitting does not rely on Gaussian assumptions regarding the form of the noise. Given a numerical integrator of order p and step size η, we show that although the discretization error vanishes at a rate $O(\eta^p)$, the SG error vanishes at rate $O(\eta^2)$. We thus identify a convergence bottleneck introduced by mini-batching, as demonstrated on an example in Figure 1 (details in Appendix F.1). This result is in contrast with the current understanding of convergence properties of sampling algorithms (Chen et al., 2015). Our work then updates the previous result about convergence rates for a broad range of integrators and with minimal assumptions about the mini-batching process. Although we focus on Hamiltonian dynamics our results can be adapted to other dynamics (Welling & Teh, 2011; Leimkuhler & Shang, 2016b; Shang et al., 2015).

Our second contribution (Section 4) is a reinterpretation of the HMC algorithm with partial momentum refreshment (Neal, 2011; Horowitz, 1991) as an integrator for Hamiltonian SDEs (Abdulle et al., 2015). This connection provides insights into the behavior of classical HMC schemes and extends the results of our theoretical analysis.

In Section 5, we conduct an extensive experimental campaign that corroborates our theory on the convergence rate of various Hamiltonian-based SDE schemes, by exploring the behavior of step size and mini-batch size for a large number of models and datasets.

2. Hamiltonian SDEs For Sampling

Given a dataset of observations $\mathrm{D} = \{\mathbf{x}_i\}_{i=1}^N$, the Bayesian treatment of machine learning models can be summarized as the combination of a prior belief $p(\boldsymbol{\theta})$ and a likelihood function $p(\mathrm{D}|\boldsymbol{\theta})$, into a posterior distribution $p(\boldsymbol{\theta}|\mathrm{D}) \propto p(\mathrm{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Since the posterior is analytically intractable for nonlinear models such as BNNs (Bishop, 2006), our goal is to draw samples from the density $p(\boldsymbol{\theta}|\mathrm{D})$, which is expressed implicitly in exponential form as $p(\boldsymbol{\theta}|\mathrm{D}) \propto \exp(-U(\boldsymbol{\theta}))$, where:

$$-U(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(\mathbf{x}_i|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}),$$

$$\text{and} \qquad p(\mathrm{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta}). \tag{2}$$

In Hamiltonian systems, $U(\boldsymbol{\theta})$ is known as the *potential*, which, together with a *kinetic energy* term, yields the Hamiltonian function $H(\mathbf{z}) = U(\boldsymbol{\theta}) + 1/2||\mathbf{M}^{-1}\mathbf{r}||^2$. The vector $\mathbf{z} = [\mathbf{r}^\top, \boldsymbol{\theta}^\top]^\top$ denotes the overall system state, which includes the position $\boldsymbol{\theta}$ (i.e., parameters) and the conjugate

momentum $\mathbf{r}$, while $\mathbf{M}$ is a symmetric, positive definite mass matrix. Then, we consider the following SDE, which is a compact form of Equation (1):

$$d\mathbf{z}(t) = \begin{bmatrix} -C\boldsymbol{I} & -\boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0} \end{bmatrix} \nabla H(\mathbf{z}(t))dt + \begin{bmatrix} \sqrt{2C}d\mathbf{w}(t) \\ \mathbf{0} \end{bmatrix}. \quad (3)$$

We study the stationary behavior of the process of Equation (3), defining $\nabla_{\mathbf{z}} = [\nabla_{\mathbf{r}}^\top, \nabla_{\boldsymbol{\theta}}^\top]^\top$, through the following differential operator:

$$\begin{aligned} \mathcal{L} = &- \left(\nabla_{\boldsymbol{\theta}}^\top U(\boldsymbol{\theta})\right) \nabla_{\mathbf{r}} + \left(\left(\mathbf{M}^{-1}\mathbf{r}\right)^\top\right) \nabla_{\boldsymbol{\theta}} \\ &- C \left(\left(\mathbf{M}^{-1}\mathbf{r}\right)^\top\right) \nabla_{\mathbf{r}} + C\nabla_{\mathbf{r}}^\top \nabla_{\mathbf{r}}, \end{aligned} \quad (4)$$

which is known as the *infinitesimal generator*. The Fokker Planck equation, that can be used to obtain the stationary distribution $\rho_{ss}(\mathbf{z})$ of the stochastic process, writes as $\mathcal{L}^\dagger \rho_{ss} = 0$, where † indicates the adjoint of the operator. Some sample paths for a simple two-parameter logistic regression can be seen in Figure 2, where we vary the constant C. Although the stationary distribution is independent of C, the transient dynamics change, resulting in paths of different form, as we see in the figure. Essentially, C is a user-defined parameter whose effect we explore in Appendix F.4 for a wide range of machine learning models. Then we have the following theorem:

Theorem 2.1. *For an ergodic stochastic process described by the* SDE *of Equation* (3) *with stationary distribution* $\rho_{ss}(\mathbf{z})$ *we have:*

$$\rho_{ss}(\mathbf{z}) \propto \exp(-H(\mathbf{z})) = \exp(-U(\boldsymbol{\theta}) - 1/2||\mathbf{M}^{-1}\mathbf{r}||^2) \quad (5)$$

The proof can be found in Appendix B.2. A direct implication is that simulating the stochastic process allows us to compute ergodic averages of functions of interest $\phi(\mathbf{z})$, of the form $\int \phi(\mathbf{z})\rho_{ss}(\mathbf{z})d\mathbf{z}$ (details in Appendix B). Since $\rho_{ss}(\mathbf{z}) = \rho_{ss}(\boldsymbol{\theta})\rho_{ss}(\mathbf{r})$, the procedure can be used to perform Bayesian averages of functions of $\boldsymbol{\theta}$ only, $\phi(\boldsymbol{\theta})$, as the following holds $\int \phi(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathrm{D})d\boldsymbol{\theta} = \int \phi(\boldsymbol{\theta})\rho_{ss}(\boldsymbol{\theta})d\boldsymbol{\theta}$. In this paper the theoretical derivations are carried out for generic functions of $\mathbf{z}$.

We shall refer to our scheme as SDE-based HMC (SHMC). This is different from SGHMC (Chen et al., 2014), for which Equation (3) is modified so that the Brownian motion term has covariance $2(\mathbf{C} - \tilde{\mathbf{V}})$, where $\tilde{\mathbf{V}}$ is an estimate of the covariance of the gradient. This was done to counterbalance the effect of the stochastic gradient, which we believe is not necessary, as we discuss in Section 5.

Remark: In recent literature, Equation (3) has been associated with normally-distributed estimates of the gradient (Chen et al., 2014; Mandt et al., 2017). In our view, however, this connection is not well-justified, as for the Brownian motion term of Equation (3) we have that $d\mathbf{w}(t) = \sqrt{dt}\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, while a (Gaussian) stochastic gradient term in the continuous limit becomes $dt\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. A more detailed exposition can be found in Appendix A. Our alternative treatment for the study of the effect of mini-batches follows in Section 3.

2.1. Ergodic errors of SDEs

Except from a handful of cases, it is not possible to draw exact sample paths from arbitrary SDEs. We consider a generic numerical integrator ψ, with step size η, whose purpose is to simulate the stochastic evolution of the SDE of interest. Formally, we look for a stochastic mapping that, from a given initial condition $\mathbf{z}_0$, generates a new random variable $\mathbf{z}_1 = \psi(\mathbf{z}_0; \eta)$ by faithfully simulating the true continuous time stochastic dynamics $\mathbf{z}(\eta)|_{\mathbf{z}(0)=\mathbf{z}_0}$. Several quantitative metrics measuring the degree of accuracy of the simulation are available (Kloeden & Platen, 1995). The simplest is the strong error, which is the expected difference between true paths and simulated ones. Relaxing to the expected difference between functions of paths corresponds to quantifying the weak error:

Definition 2.2. (Debussche & Faou, 2012) Consider Equation (3) with initial condition $\mathbf{z}_0$. The numerical integrator ψ has weak order of convergence p if $|\mathbb{E}\left[\phi(\psi(\mathbf{z}_0; \eta))\right] - \mathbb{E}\left[\phi(\mathbf{z}(\eta))\right]| = \mathcal{O}(\eta^{p+1})$, where $\mathbf{z}(\eta)$ is the value taken by Equation (3) after a time η, the chosen step size.

The transformation of the functions $\phi(\mathbf{z})$ of the true stochastic dynamics $\mathbf{z}(t)$, starting from initial conditions $\mathbf{z}_0$ is described by the Kolmogorov differential equation $\mathbb{E}\left[\phi(\mathbf{z}(t))\right] = \exp\left(t\mathcal{L}\right)\phi(\mathbf{z})|_{\mathbf{z}=\mathbf{z}_0}$, that throughout the paper we indicate with abuse of notation as $\mathbb{E}\left[\phi(\mathbf{z}(t))\right] = \exp\left(t\mathcal{L}\right)\phi(\mathbf{z}_0)$. The operator $\mathcal{U}$, defined as $\mathcal{U}\phi(\mathbf{z}_0) = \mathbb{E}\left[\phi(\psi(\mathbf{z}_0; \eta))\right]$, represents a order-$p$ weak integrator if $\mathcal{U} = \exp\left(\eta\mathcal{L}\right) + \mathcal{O}(\eta^{p+1})$.

Since we are only interested in samples from $p(\boldsymbol{\theta}|D)$, it is sufficient to consider the even weaker *ergodic average error*, defined hereafter. The numerical integrator iteratively induces a stochastic process $\mathbf{z}_i = \psi(\mathbf{z}_{i-1}; \eta), i = 1, \dots N$. The ergodic average of a given function $\phi(\cdot)$ converges to the integral $\int \phi(\mathbf{z})\rho_{ss}^{\psi}(\mathbf{z})d\mathbf{z}$, where ρ_{ss}^{ψ} is the stationary distribution of the stochastic process induced by the numerical integrator ψ. The ergodic average error is the difference between the ergodic average of the numerical integrator and the true average obtained with the stationary distribution of Equation (3):

$$e(\psi, \phi) = \int \phi(\mathbf{z})\rho_{ss}^{\psi}(\mathbf{z})d\mathbf{z} - \int \phi(\mathbf{z})\rho_{ss}(\mathbf{z})d\mathbf{z}. \quad (6)$$

Albeit outside the scope of this work, the results presented here can be extended to non-asymptotic (in time) settings.

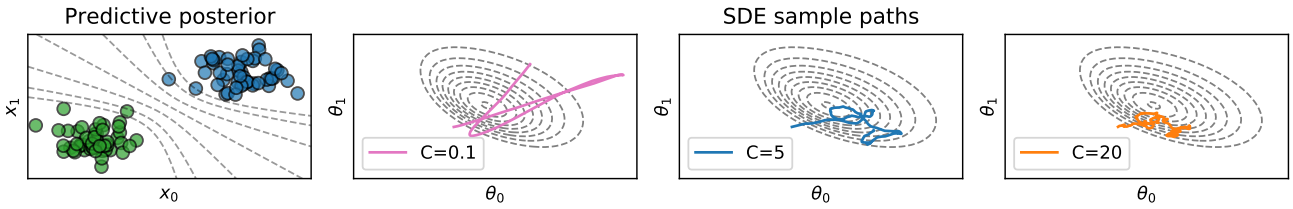

Figure 2. Sample paths for simple a two-parameter logistic regression problem (shown in the leftmost panel). We show SDE paths for different choices of friction coefficient C.

Indeed, the error with respect to a finite time empirical average can be expressed as $\frac{1}{N}\sum_{n=0}^{N-1}\phi(\mathbf{z}_n) - \int \phi(\mathbf{z})\rho_{ss}(\mathbf{z})d\mathbf{z}$. By the triangle inequality, its absolute value can be bounded by: $\left|\frac{1}{N}\sum_{n=0}^{N-1}\phi(\mathbf{z}_n) - \int \phi(\mathbf{z})\rho_{ss}^{\psi}(\mathbf{z})d\mathbf{z}\right| + |e(\psi,\phi)|$. The first term can be bounded in expectation by considering the convergence speed (as a function of time) to the stationary distribution of the numerical integration scheme, either by classical arguments or attempting a weak backward error analysis (Debussche & Faou, 2012). In this work, we consider only the quantity in Equation (6).

Our goal is to characterize the rate of convergence to zero for $e(\psi,\phi)$ as a function of the step size η, the most important free parameter. A sufficient condition for an integrator to be of a given ergodic order p, i.e. $e(\psi,\phi) = \mathcal{O}(\eta^p)$, is to have weak order p (Abdulle et al., 2014). See Appendix B.3 for a more detailed exposition.

2.2. Numerical integrators

We explore the effect of a number of SDE integrators of weak order two and three; these schemes are outlined in detail in Appendix B.5. Hamiltonian systems evolve on manifolds with peculiar geometrical properties whose study is the subject of *symplectic geometry*. In Appendix B.4 we present the class of quasi-symplectic integrators (Milstein & Tretyakov, 2003). To understand the main text it is sufficient to know that such integrators are empirically known to outperform their non-symplectic counterpart.

The LEAPFROG scheme is probably the simplest form of quasi-symplectic integrator, which involves updating the position and momentum at interleaved time steps. Importantly, while it has theoretically order-one convergence both in weak and ergodic sense, we observe empirically that its performance is very close to other quasi-symplectic order-two schemes. A similar phenomenon has been observed by Milstein & Tretyakov (2003).

Finally we consider a LIE-TROTTER splitting scheme (Abdulle et al., 2015) that is of weak order-two, while it is also quasi-symplectic. In Section 4 this scheme will be used to draw a theoretical connection between HMC (Neal, 2011) and the proposed SDE framework in Equation (3).

In order to demonstrate the convergence bottleneck more clearly, we also include an integrator of third order in our comparisons. We employ the 3-stage quasi-symplectic integrator of Milstein & Tretyakov (2003). This scheme might not be always practical, as it involves three stages and one computation of the Hessian. However, it is useful to demonstrate the convergence bottleneck effect due to mini-batches.

3. Mini-batches as Operator Splitting

We explore the concept of operator splitting and its applicability to the case of Hamiltonian SDE simulation when considering subsets of the full dataset D for computing the driving force. Without loss of generality, suppose that the dataset D is split into two mini-batches $\mathrm{D}_1, \mathrm{D}_2$. Following Equation (4) we can define the infinitesimal generators $\mathcal{L}_1, \mathcal{L}_2$, for which we have $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$. Intuitively, given an operator in exponential form $\exp(\eta(\mathcal{L}_1+\mathcal{L}_2))$, we would like to determine under which conditions the following holds

$$\exp(\eta(\mathcal{L}_1 + \mathcal{L}_2)) \simeq \exp(\eta\mathcal{L}_1)\exp(\eta\mathcal{L}_2), \tag{7}$$

and to rigorously quantify the discrepancy error. The barrier to an exact identity is the non commutativity of the differential operators $\mathcal{A}, \mathcal{B}$: in general the commutator $[\mathcal{A},\mathcal{B}] \coloneqq \mathcal{A}\mathcal{B} - \mathcal{B}\mathcal{A} \neq 0$. More broadly, we split $\mathcal{L}$ into K mini-batches of the form $\mathcal{L} = \sum_{i=1}^{K}\mathcal{L}_i$ where

$$\begin{aligned}\mathcal{L}_i = &-\left(\nabla_{\boldsymbol{\theta}}^{\top}U_i(\boldsymbol{\theta})\right)\nabla_{\mathbf{r}} + K^{-1}\left(\left(\mathbf{M}^{-1}\mathbf{r}\right)^{\top}\right)\nabla_{\boldsymbol{\theta}} \\ &- K^{-1}C\left(\left(\mathbf{M}^{-1}\mathbf{r}\right)^{\top}\right)\nabla_{\mathbf{r}} + K^{-1}C\nabla_{\mathbf{r}}^{\top}\nabla_{\mathbf{r}},\end{aligned} \tag{8}$$

with $U_i(\boldsymbol{\theta})$ the potential computed using only the i_{th} mini-batch. The theorems presented in this section clarify how concatenations of the form $\prod_i \exp(\eta K\mathcal{L}_i)$ induce errors and clarify their relevance for the considered problem.

Average of back and forth operators. As an intermediate result of our analysis, we derive the following theorem, to reason about the local error of the average of chaining a series of operators in certain order (forward) and chaining

the same operators in the reverse order (backward). More formally, we have:

Theorem 3.1. *Let $\mathcal{L} = \sum_{i=1}^{K} \mathcal{L}_i$ be a differential operator, and $\mathcal{P}$ be the operator obtained by the average:*

$$\mathcal{P} = \frac{\prod_{i=K}^{1} \exp(\eta K \mathcal{L}_i) + \prod_{i=1}^{K} \exp(\eta K \mathcal{L}_i)}{2}. \tag{9}$$

Then $\mathcal{P}$ has error $\mathcal{O}(K\eta^3)$ for the operator $\exp(\eta K \mathcal{L})$, i.e.

$$\mathcal{P} = \exp(\eta K \mathcal{L}) + \mathcal{O}(K\eta^3). \tag{10}$$

The proof is based on the Taylor expansion of the operators and *completing the square*. Intuitively, given differential operator $\mathcal{C} = \mathcal{A} + \mathcal{B}$, we have $\exp(\eta\mathcal{A})\exp(\eta\mathcal{B}) = \mathcal{I} + \eta(\mathcal{A}+\mathcal{B}) + \frac{\eta^2}{2}(\mathcal{A}^2 + 2\mathcal{A}\mathcal{B} + \mathcal{B}^2) + \mathcal{O}(\eta^3)$. Since $\mathcal{C}^2 = \mathcal{A}^2 + \mathcal{A}\mathcal{B} + \mathcal{B}\mathcal{A} + \mathcal{B}^2$, simple algebra shows $\exp(\eta\mathcal{C}) = 1/2\,(\exp(\eta\mathcal{A})\exp(\eta\mathcal{B}) + \exp(\eta\mathcal{B})\exp(\eta\mathcal{A})) + \mathcal{O}(\eta^3)$. Details are presented for $K > 2$ in Appendix C.2.

Randomized sequence of numerical integration steps. Theorem 3.1 allows us to reason about the local error of a randomized sequence of numerical integrators and the induced differential operators at the level of functions transformation. In most machine learning contexts, mini-batches are created by sampling a dataset without replacement (Shamir, 2016; Haochen & Sra, 2019). In our setting, we consider K numerical integrators $\{\psi_i\}_{i=1}^{K}$ of order p obtained using K mini-batches, and the corresponding differential operators $\mathcal{U}_i \coloneqq \mathbb{E}\left[\phi(\psi_i(\mathbf{z}_0;\eta))\right] = \exp(\eta K \mathcal{L}_i)\phi(\mathbf{z}_0) + \mathcal{O}(\eta^{p+1})$.

A pass of the entire dataset is often referred to as an epoch, which corresponds to one of $K!$ possible permutations: $\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^{K!}$. Assuming a uniform distribution over all permutations, the expected differential operator $\mathcal{U}$ is obtained by the law of total expectation considering the average over (i) the numerical integration randomness and (ii) the random permutation $\boldsymbol{\pi}$:

$$\begin{aligned}
\mathcal{U}\phi(\mathbf{z}_0) &= \mathbb{E}[\phi(\psi_{\pi_K}(\ldots\psi_{\pi_1}(\mathbf{z}_0)))] = \\
&= \sum_{j=1}^{K!} \frac{1}{K!}\left(\mathbb{E}\left[\phi(\psi_{\pi_K^j}(\ldots\psi_{\pi_1^j}(\mathbf{z}_0)))\right]\right) = \\
&= \left(\sum_{j=1}^{K!} \frac{1}{K!} \mathcal{U}_{\pi_K^j}\mathcal{U}_{\pi_{K-1}^j}\ldots\mathcal{U}_{\pi_1^j}\right)\phi(\mathbf{z}_0),
\end{aligned}$$

It is possible to rewrite the set of all possible permutations as the union of two sets of equal cardinality, such that each element of one of the two set has a (unique) element of the other set that is its lexicographic reverse:

$$\mathcal{U} = \frac{1}{K!}\sum_{j=1}^{K!/2}\left(\mathcal{U}_{\pi_K^j}\mathcal{U}_{\pi_{K-1}^j}\ldots\mathcal{U}_{\pi_1^j} + \mathcal{U}_{\pi_1^j}\mathcal{U}_{\pi_2^j}\ldots\mathcal{U}_{\pi_K^j}\right)$$

This observation, in conjunction with Theorem 3.1, implies that the expected operator $\mathcal{U}$ has locally an error that depends on p, and, crucially, on the back and forth splitting. Interestingly, randomizing the sequence of the integrator achieves the same accuracy of a symmetric splitting, based on the Baker–Campbell–Hausdorff formula (Dynkin, 1947) where the commutator terms cancel out. In addition, the randomized scheme is simpler to implement (Zhang, 2012). The following theorem formalizes our discussion.

Theorem 3.2. *Consider numerical integrators $\{\psi_i\}_{i=1}^{K}$ of order p obtained using K mini-batches. Extract uniformly $\boldsymbol{\pi} = \left[\pi_1, \ldots, \pi_K\right] \in \mathbb{P}$, the set of all the possible permutations of the indices $\{1, K\}$. The scheme in which initial condition has stochastic evolution through the chain of integrators with the following ordering:*

$$\mathbf{z}_{fin} = \psi_{\pi_K}(\psi_{\pi_{K-1}}(\ldots\psi_{\pi_1}(\mathbf{z}_0))), \tag{11}$$

transforms the functions ϕ with an operator $\mathcal{U}$ that has the following expression:

$$\mathcal{U} = \exp(\eta K \mathcal{L}) + \mathcal{O}(K\eta^{p+1}) + \mathcal{O}(K\eta^3) \tag{12}$$

In other words, the randomized mini-batch splitting introduces a bottleneck, as the term $\mathcal{O}(K\eta^3)$ dominates over $\mathcal{O}(K\eta^{p+1})$ whenever $p > 2$. This is reflected into the behavior of ergodic error. The complete proof can be found in Appendix C.4.

Ergodic error bottleneck. Regarding the ergodic error, we combine Theorem 3.2 with the result of Abdulle et al. (2014) (Appendix B.3), which states that for an integrator to be of ergodic order p, it is sufficient to have weak order p. Then, the following theorem holds.

Theorem 3.3. *Consider the settings described by Theorem 3.2. Repeatedly apply the numerical integration scheme. Then the ergodic error has expansion*

$$e(\psi, \phi) = \mathcal{O}(\eta^{\min(p,2)}) \tag{13}$$

The effect of mini-batches adds an extra error of order two in the convergence rate: this error can become a bottleneck whenever $p > 2$, and we thus cannot guarantee any higher ergodic convergence in the general case. Note that the extra error is not due to an equivalent noise injection into the SDE dynamics. See Appendix C.5 for the proof.

We include a toy example taken from Vollmer et al. (2016), where an analytical solution is available. The complete specification can be found in Appendix E. The results, reported in Figure 3, are obtained by comparing an order-two integrator against an analytical solution, when considering or not mini-batches. Qualitatively, the presence of a bottleneck is clear: when considering mini-batches, the stationary distribution is not the desired one, even with a perfect integrator.

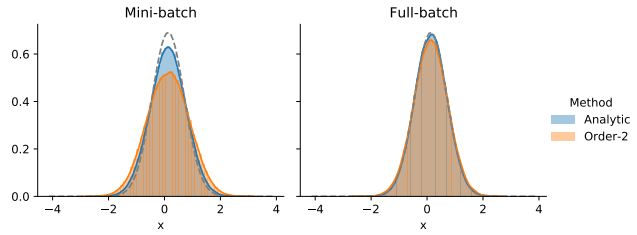

Figure 3. Estimated densities of stationary distributions. The grey dotted line denotes the true posterior density. Irrespectively of order, mini-batching prevents from convergence to true posterior.

Our main result, the presence of a bottleneck, revises previous results presented in Chen et al. (2015), where the mini-batching does not affect the ergodic average error. Instead, our theoretical and empirical discussions clearly show that the introduction of mini-batches impedes convergence to the desired posterior.

Details on the constants in the convergence results. Some clarifications are in order. First, the constants in the $\mathcal{O}$ notation of weak and ergodic errors, could be refined considering the geometry of the potentials and the norms inequality of differential operators. We leave such a possibility for future works and refer the interested reader to Childs et al. (2021; 2019) and in particular also to Zhang (2012), that proposes randomized schemes strictly related to the one discussed here. Second, we stress that when performing sampling for Bayesian inference problems, the full potential is divided into mini-batches, and each subset rescaled by the constant K. This rescaling allows one to "cover" the same amount of distance per steps independently from K, given that with a pure splitting one would need K steps to simulate $\exp(\eta\mathcal{L})$. For the case of HMC, this connection is acknowledged in Betancourt (2015), drawing an equivalence with the work of Shahbaba et al. (2014).

4. A Lie-Trotter Integration Scheme and Connections with HMC

We discuss a SDE integration scheme that relies on a LIE-TROTTER splitting of the infinitesimal generator $\mathcal{L}$. That allows us to draw a connection between Hamiltonian SDEs and the original HMC family of algorithms, by showing that the latter can be interpreted as an integration scheme of the SDE dynamics.

The infinitesimal generator in Equation (4) can be expressed as the sum $\mathcal{L} = \mathcal{H} + \mathcal{D}$, where:

$$\mathcal{H} = -\left(\nabla_{\boldsymbol{\theta}}^{\top} U(\boldsymbol{\theta})\right) \nabla_{\mathbf{r}} + \left(\left(\mathbf{M}^{-1}\mathbf{r}\right)^{\top}\right) \nabla_{\boldsymbol{\theta}},$$
$$\mathcal{D} = -C\left(\left(\mathbf{M}^{-1}\mathbf{r}\right)^{\top}\right) \nabla_{\mathbf{r}} + C\nabla_{\mathbf{r}}^{\top}\nabla_{\mathbf{r}}.$$

In its general form (Abdulle et al., 2015), the LIE-TROTTER scheme is derived as an application of the Baker–Campbell–Hausdorff formula (Dynkin, 1947), where the dynamics are approximated as follows:

$$\exp(\eta\mathcal{L}) = \exp(\eta/2\mathcal{D}) \exp(\eta\mathcal{H}) \exp(\eta/2\mathcal{D}) + \mathcal{O}(\eta^3)$$

In practice, the scheme consists of alternating steps that solve the $\mathcal{H}$ and $\mathcal{D}$ parts of $\mathcal{L}$. The design space is the one of deterministic integrators for the Hamiltonian part $\mathcal{H}$, as the term $\mathcal{D}$ can be solved exactly. Having chosen a deterministic integrator ψ, a single step of the SDE simulation has the following form:

$$\mathbf{z}^* = \left[\mathbf{r}^{*\top}, \boldsymbol{\theta}^{*\top}\right]^{\top} = \psi(\mathbf{z}_0), \tag{14}$$
$$\mathbf{z}_1 = \left[\exp(-\eta C)\mathbf{r}^{*\top} + \sqrt{1-\exp(-2\eta C)}\mathbf{w}^{\top}, \boldsymbol{\theta}^{*\top}\right] \tag{15}$$

It can be shown (Abdulle et al., 2015) that whenever the numerical integrator ψ for the Hamiltonian part $\mathcal{H}$ has order p, i.e. $\mathcal{U}\phi = \exp(\eta\mathcal{H}) + \mathcal{O}(\eta^{p+1})$, the LIE-TROTTER scheme has convergence order $\mathcal{O}(\eta^p)$.

In the experiments of Section 5, we consider ψ to be the *deterministic version* of LEAPFROG, for which we have $p = 2$. Therefore, our practical application of LIE-TROTTER is of order $\mathcal{O}(\eta^2)$.

HMC with partial momentum refreshment. We examine a generalized HMC variant featuring partial momentum refreshment (Horowitz, 1991; Neal, 2011), which consists of two (repeated) steps. First, we have a numerical approximation of a Hamiltonian system by means of a (deterministic) integrator ψ:

$$\mathbf{z}^* = \left[\mathbf{r}^{*\top}, \boldsymbol{\theta}^{*\top}\right]^{\top} = \underbrace{\psi(\ldots\psi(\psi(}_{N_l \text{times}}\mathbf{z}_0))), \tag{16}$$

where N_l denotes the number of integration steps and it is assumed to be finite. Then, we have a a partial momentum update as follows:

$$\mathbf{z}_1 = \left[\alpha\mathbf{r}^{*\top} + \sqrt{1-\alpha^2}\mathbf{w}^{\top}, \boldsymbol{\theta}^{*\top}\right], \tag{17}$$

where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\alpha > 0$.

The connection between HMC and the LIE-TROTTER is drawn by noticing that mechanically, the two schemes simulate SDEs (and thus transform functions) in a similar fashion. LIE-TROTTER transforms functions after a step as $\mathbb{E}\left[\phi(\mathbf{z}_1)\right] = \mathcal{U}\exp(\eta\mathcal{D})$, while HMC as $\mathbb{E}\left[\phi(\mathbf{z}_1)\right] = (\mathcal{U})^{N_l}\exp(\eta N_l\mathcal{D})$. The claim is true when we consider the α of HMC to be $\alpha = \exp(-\eta N_l C)$ (the full momentum resampling corresponds to $C \to \infty$).

By exploiting this connection to classical SDE integrators, we can thus study the ergodic error of HMC scheme with and without mini-batches. The ergodic error for the two cases is $\mathcal{O}(\eta^p)$ and $\mathcal{O}(\eta^{\min(p,2)})$ respectively, showing again the bottleneck introduced by mini-batches. See Appendix D for the derivation details.

In both cases, the orders of convergence are *independent* on N_l. This is reflected in the experimental results in Section 5. We note that, in practice, the approximation quality degrades by increasing N_l for higher learning rates.

5. Experiments

Comparison framework. The main objective of the experiments is to investigate convergence to the true posterior distribution for a wide range of BNN models and datasets. Metrics that reflect regression and classification accuracy are not of main interest; nevertheless, some of these are reported in Appendix F.1. Instead, we turn our attention to the quality of the predictive distribution.

We consider the true predictive posterior as the ground truth, which is approximated by a careful application of SGHMC (Springenberg et al., 2016) featuring a very small step size and full-batch gradient calculation; this is referred to as the *oracle*. Any comparison between high-dimensional empirical distributions gives rise to significant challenges. Therefore, we resort to comparing one-dimensional predictive distributions by means of the Kolmogorov distance. For a given test dataset, we explore the average Kolmogorov distance from the true posterior predictive distributions for different methods, step sizes and mini-batch sizes. These should be compared with average Kolmogorov *self-distance* for the oracle, which is marked as grey dotted lines in the figures that follow[2]. In all cases we compare empirical distributions of 200 samples; the example of Figure 1 is an exception, where we compare distributions of 2000 samples. See Appendix F.2 for a complete account.

For our SHMC method, we examine integrators of different orders: LEAPFROG, LIE-TROTTER and MT3, and we investigate whether they differ from SGHMC (Chen et al., 2014). The LEAPFROG scheme, which for SDEs is provably of order one, is also used in SGHMC. The LIE-TROTTER integrator introduced in Section 4, based on a deterministic variant of LEAPFROG, is of order two, while MT3 denotes the order-three quasi-symplectic integrator of Milstein & Tretyakov (Milstein & Tretyakov, 2003). In all cases we set $C = 5$; we find this to be a reasonable choice, as can be seen in the exploration of Appendix F.4.

We consider four regression datasets (BOSTON, CONCRETE, ENERGY, YACHT) and two classification datasets (IONOSPHERE, VEHICLE) from the UCI repository, as well as a 1-D synthetic dataset, for which the regression result is shown in Appendix F.1. Due to space limitations, here we only present a summary of the results in Figures 4 and 5. A more detailed exposition, including a more fine-grained exploration of the batch size, can be found in Appendix F.3 and Appendix F.4. In what follows, we summarize some findings that apply to all the datasets and models we have considered.

Comparing with SGHMC. In Figure 4 we focus our attention on the comparison between our SHMC method (with LEAPFROG integrator) and SGHMC (Chen et al., 2014). We note that SHMC (LEAPFROG) is different from SGHMC in the sense that no counterbalancing of the noise is performed. Nevertheless, we do not observe significant difference between the two approaches. We argue that this result is compatible with our position that counterbalancing the gradient noise is not necessary to sample from the posterior.

On the convergence bottleneck. As a general remark, we notice that for larger mini-batch sizes (or equivalently, for smaller gradient noise) the model can admit larger step sizes, while as the gradient noise becomes larger, then a smaller value for η is required to guarantee a good approximation.

When comparing integrators of different order in Figures 1 and 4, the most interesting finding is that all of our results appear to confirm the existence of the convergence bottleneck identified in Section 3. This becomes particularly obvious in Figure 1, where we use SHMC to draw 2000 samples. In the cases where mini-batches are used, there is a zone in which higher-order integrators do not deliver significant improvements. Note that this zone is absent from the full-batch comparison, as higher order schemes are consistently better everywhere in the range of step size η.

Comparison of integrators. In Figures 1 and 4, although all of the integrators considered converge to the true posterior regardless of the mini-batch size given a sufficiently small step size, we observe that the methods behave differently depending on their theoretical properties.

Although the LEAPFROG scheme is provably of order one, it has been mostly competitive to higher order schemes (LIE-TROTTER, MT3) in our experiments; its performance deteriorates for larger values for η only. Lastly, the third order MT3 scheme consistently outperforms the rest of the integrators in the full batch case, especially for larger step sizes; this has little effect from a practical point of view however, as MT3 requires extra calculations per step (i.e. for 3 gradients and the Hessian).

Exploring the connection with HMC. As a last experiment, we consider a generalized LIE-TROTTER scheme where we explore different values for the integration length N_l. For

[2] This emulates the famous Kolmogorov-Smirnov test, but with no Gaussianity assumptions.

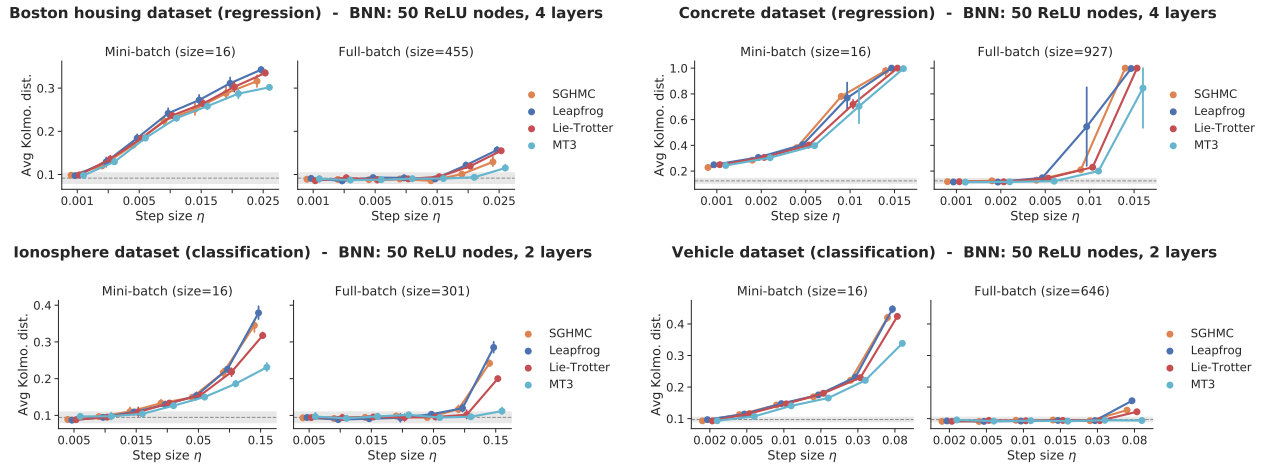

Figure 4. Exploration of step size and batch size for different Hamiltonian-based methods; the grey dotted line denotes the self-distance for the distribution of the oracle. When using mini-batches, there is an area for which MT3 (order-three) does not deliver any significant improvement over lower order integrators. The curves are slightly shifted to improve readability.

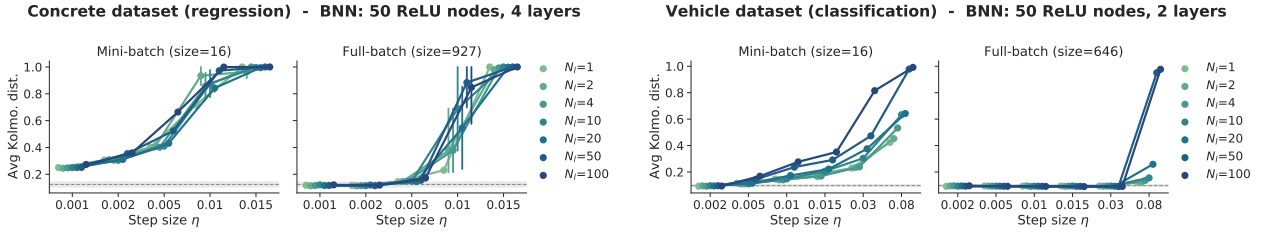

Figure 5. Generalized LIE-TROTTER scheme: Exploration of step size and batch size for different values of the (deterministic) integration length N_l. The grey dotted line denotes the self-distance for the distribution of the oracle.

$N_l = 1$ we obtain the standard LIE-TROTTER of Section 4, while as N_l grows we recover a HMC algorithm with partial momentum refreshments. We have claimed that regardless of the N_l value, the scheme enjoys ergodic convergence of order $\mathcal{O}(\eta^2)$. The results of the exploration can be seen in Figure 5. For smaller step sizes η, the order of convergence is shown to be similar for all N_l, especially in the full-batch case. Nevertheless, we observe larger errors as η grows and N_l approaches 100. This effect seems to be further pronounced by the introduction of mini-batches. Note that the divergence we observe for larger step sizes does not contradict our claim for the same ergodic convergence, as this is not the region of interest in terms of $\mathcal{O}(\cdot)$ notation. This finding is not surprising nevertheless: it is well-established in the literature that HMC is sensitive to the choice of integration length (Neal, 2011; Hoffman & Gelman, 2014). We remark that a sensible choice is to use $N_l = 1$, which is the most direct simulation of the Hamiltonian SDE system.

6. Conclusions

In this work, we revisited the connections between SG and stochastic Hamiltonian dynamics to improve our current understanding of the role played by mini-batching on the goodness of sampling from intractable distributions, by means of simulating a Hamiltonian SDE. We challenged the common assumption of associating stochastic gradient estimates that arise due to data subsampling to the stochastic component of the SDE, arguing that the Brownian motion is a poor model of this kind of gradient noise.

Our main contribution was to produce new convergence results for Hamiltonian SDE-based sampling methods, by studying their properties through the lenses of differential operator splitting. We found that, for an integrator of weak order p and step size η, the discretization error may vanish at rate $\mathcal{O}(\eta^p)$, but the mini-batch error vanishes at rate $\mathcal{O}(\eta^2)$, which is a bottleneck that has been overlooked in the literature.

Using our theory, we also showed that HMC with partial momentum refreshments can be interpreted as an integration

scheme for the same class of SDEs. Then, we showed that convergence rates are independent on the HMC inner loop size N_l, as we confirmed experimentally.

We have demonstrated the validity of our theory on a wide range of experiments, where we have meticulously documented the deviation from the true posterior distribution. As a practical implication of our work, we recommended a straightforward simulation of the SDE in Equation (1) by means of a quasi-symplectic second order integrator. Indeed, the bottleneck introduced by mini-batching would neutralize the benefits of higher order integrators.

Acknowledgements

GF was partially supported by Huawei Technologies, Paris. MF gratefully acknowledges support from the AXA Research Fund and the Agence Nationale de la Recherche (grant ANR-18-CE46-0002 and ANR-19-P3IA-0002).

References

Abdulle, A., Vilmart, G., and Zygalakis, K. C. High order numerical approximation of the invariant measure of ergodic sdes. *SIAM Journal on Numerical Analysis*, 52(4): 1600–1622, 2014.

Abdulle, A., Vilmart, G., and Zygalakis, K. C. Long time accuracy of Lie-Trotter splitting methods for Langevin dynamics. *SIAM Journal on Numerical Analysis*, 53(1): 1–16, 2015.

Ahn, S., Korattikara, A., and Welling, M. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *International Conference on Machine Learning*, pp. 1771–1778, 2012.

Alenlöv, J., Doucet, A., and Lindsten, F. Pseudo-marginal Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 22(141):1–45, 2021.

Betancourt, M. The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling. In *International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 533–540. PMLR, 2015.

Betancourt, M., Jordan, M. I., and Wilson, A. C. On symplectic optimization. *arXiv preprint arXiv:1802.03653*, 2018.

Bishop, C. M. *Pattern recognition and machine learning*. Springer, 1st ed. 2006. corr. 2nd printing 2011 edition, 2006. ISBN 0387310738.

Bou-Rabee, N. and Owhadi, H. Long-run accuracy of variational integrators in the stochastic context. *SIAM Journal on Numerical Analysis*, 48(1):278–297, 2010.

Castella, F., Chartier, P., Descombes, S., and Vilmart, G. Splitting methods with complex times for parabolic equations. *BIT Numerical Mathematics*, 49(3):487–508, 2009.

Chaudhari, P. and Soatto, S. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10. IEEE, 2018.

Chen, C., Ding, N., and Carin, L. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2278–2286. Curran Associates, Inc., 2015.

Chen, T., Fox, E., and Guestrin, C. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1683–1691. PMLR, 2014.

Childs, A. M. and Su, Y. Nearly optimal lattice simulation by product formulas. *Physical review letters*, 123(5): 050503, 2019.

Childs, A. M., Ostrander, A., and Su, Y. Faster quantum simulation by randomization. *Quantum*, 3:182, 2019.

Childs, A. M., Su, Y., Tran, M. C., Wiebe, N., and Zhu, S. Theory of Trotter error with commutator scaling. *Physical Review X*, 11(1):011020, 2021.

Cobb, A. D. and Jalaian, B. Scaling Hamiltonian Monte Carlo inference for Bayesian neural networks with symmetric splitting. In *Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 675–685. PMLR, 2021.

Debussche, A. and Faou, E. Weak backward error analysis for sdes. *SIAM Journal on Numerical Analysis*, 50(3): 1735–1752, 2012.

Dynkin, E. B. Calculation of the coefficients in the Campbell-Hausdorff formula. *Dokl. Akad. Nauk. SSSR (NS)*, 57:323–326, 1947.

França, G., Sulam, J., Robinson, D. P., and Vidal, R. Conformal symplectic and relativistic optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2020 (12):124008, 2020.

França, G., Jordan, M. I., and Vidal, R. On dissipative symplectic integration with applications to gradient-based optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(4):043402, 2021.

Franzese, G., Milios, D., Filippone, M., and Michiardi, P. A scalable bayesian sampling method based on stochastic gradient descent isotropization. *Entropy*, 23(11):1426, 2021.

Futami, F., Sato, I., and Sugiyama, M. Accelerating the diffusion-based ensemble sampling by non-reversible dynamics. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3337–3347. PMLR, 2020.

Gao, X., Gürbüzbalaban, M., and Zhu, L. Global convergence of stochastic gradient Hamiltonian Monte Carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration. *arXiv preprint arXiv:1809.04618*, 2018.

Gao, X., Gurbuzbalaban, M., and Zhu, L. Breaking reversibility accelerates Langevin dynamics for non-convex optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17850–17862. Curran Associates, Inc., 2020.

Gardiner, C. W. *Handbook of stochastic methods for physics, chemistry and the natural sciences*, volume 13 of *Springer Series in Synergetics*. Springer-Verlag, third edition, 2004. ISBN 3-540-20882-8.

Gurbuzbalaban, M., Simsekli, U., and Zhu, L. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3964–3975. PMLR, 2021.

Haochen, J. and Sra, S. Random shuffling beats SGD after finite epochs. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2624–2633. PMLR, 2019.

Hatano, N. and Suzuki, M. Finding exponential product formulas of higher orders. In *Quantum annealing and other optimization methods*, pp. 37–68. Springer, 2005.

Hoffman, M. D. and Gelman, A. The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

Hörmander, L. Hypoelliptic second order differential equations. *Acta Mathematica*, 119(1):147–171, 1967.

Horowitz, A. M. A generalized guided Monte Carlo algorithm. *Physics Letters B*, 268(2):247–252, 1991.

Jacot, A., Gabriel, F., and Hongler, C. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8571–8580. Curran Associates, Inc., 2018.

Kloeden, P. E. and Platen, E. *Numerical solution of Stochastic Differential Equations*. Springer-Verlag, 1995.

Lee, J., Schoenholz, S. S., Pennington, J., Adlam, B., Xiao, L., Novak, R., and Sohl-Dickstein, J. Finite versus infinite neural networks: an empirical study. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15156–15172. Curran Associates, Inc., 2020.

Leen, T. and Moody, J. Weight space probability densities in stochastic learning: I. Dynamics and equilibria. In *Advances in Neural Information Processing Systems*, volume 5, pp. 451–458. Morgan-Kaufmann, 1992.

Leimkuhler, B. and Shang, X. Adaptive thermostats for noisy gradient systems. *SIAM Journal on Scientific Computing*, 38(2):A712–A736, 2016a.

Leimkuhler, B. and Shang, X. Adaptive thermostats for noisy gradient systems. *SIAM Journal on Scientific Computing*, 38(2):A712–A736, 2016b.

Low, G. H., Kliuchnikov, V., and Wiebe, N. Well-conditioned multiproduct Hamiltonian simulation. *arXiv preprint arXiv:1907.11679*, 2019.

Ma, Y.-A., Chen, T., and Fox, E. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2917–2925. Curran Associates, Inc., 2015.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Mattingly, J. C., Stuart, A. M., and Tretyakov, M. V. Convergence of numerical time-averaging and stationary measures via poisson equations. *SIAM Journal on Numerical Analysis*, 48(2):552–577, 2010.

Milstein, G. and Tretyakov, M. Computing ergodic limits for Langevin equations. *Physica D: Nonlinear Phenomena*, 229(1):81–95, 2007. ISSN 0167-2789.

Milstein, G. and Tretyakov, M. V. Quasi-symplectic methods for Langevin-type equations. *IMA journal of Numerical Analysis*, 23(4):593–626, 2003.

Milstein, G. N., Repin, Y. M., and Tretyakov, M. V. Symplectic integration of Hamiltonian systems with additive noise. *SIAM Journal on Numerical Analysis*, 39(6):2066–2088, 2002.

Monmarché, P. High-dimensional MCMC with a standard splitting scheme for the underdamped Langevin diffusion. *Electronic Journal of Statistics*, 15:4117–4166, 2021.

Mukhoti, J., Stenetorp, P., and Gal, Y. On the importance of strong baselines in Bayesian deep learning. In *NeurIPS 2018: Workshop on Bayesian Deep Learning*, 2018.

Neal, R. M. Bayesian Learning for Neural Networks. In *Lecture Notes in Statistics*, volume 118. Springer, 1996.

Neal, R. M. MCMC using Hamiltonian dynamics. *Handbook of Markov chain Monte Carlo*, 54:113–162, 2011.

Patterson, S. and Teh, Y. W. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, volume 26, pp. 3102–3110. Curran Associates, Inc., 2013.

Shahbaba, B., Lan, S., Johnson, W. O., and Neal, R. M. Split Hamiltonian Monte Carlo. *Statistics and Computing*, 24 (3):339–349, 2014.

Shamir, O. Without-replacement sampling for stochastic gradient methods. In *Advances in Neural Information Processing Systems*, volume 29, pp. 46–54. Curran Associates, Inc., 2016.

Shang, X., Zhu, Z., Leimkuhler, B., and Storkey, A. J. Covariance-controlled adaptive Langevin thermostat for large-scale Bayesian sampling. In *Advances in Neural Information Processing Systems*, volume 28, pp. 37–45. Curran Associates, Inc., 2015.

Shi, B., Su, W. J., and Jordan, M. I. On learning rates and Schrödinger operators. *arXiv preprint arXiv:2004.06977*, 2020.

Springenberg, J. T., Klein, A., Falkner, S., and Hutter, F. Bayesian optimization with robust Bayesian neural networks. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4134–4142. Curran Associates, Inc., 2016.

Suzuki, M. On the convergence of exponential operators—the Zassenhaus formula, BCH formula and systematic approximants. *Communications in Mathematical Physics*, 57(3):193–200, 1977.

Talay, D. Stochastic Hamiltonian systems: exponential convergence to the invariant measure, and discretization by the implicit Euler scheme. *Markov Process. Related Fields*, 8(2):163–198, 2002.

Talay, D. and Tubaro, L. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and Applications*, 8(4):483–509, 1990.

Tran, B.-H., Rossi, S., Milios, D., and Filippone, M. All you need is a good functional prior for Bayesian deep learning. *Journal of Machine Learning Research*, 23(74): 1–56, 2022.

Vollmer, S. J., Zygalakis, K. C., and Teh, Y. W. Exploration of the (non-) asymptotic bias and variance of stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17(159):1–48, 2016.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, pp. 681–688, 2011.

Wenzel, F., Roth, K., Veeling, B., Swiatkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. How good is the Bayes posterior in deep neural networks really? In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10248–10259. PMLR, 2020.

Xu, P., Chen, J., Zou, D., and Gu, Q. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31, pp. 3122–3133. Curran Associates, Inc., 2018.

Yaida, S. Fluctuation-dissipation relations for stochastic gradient descent. In *International Conference on Learning Representations*, 2019.

Zhang, C. Randomized algorithms for Hamiltonian simulation. In *Monte Carlo and Quasi-Monte Carlo Methods 2010*, pp. 709–719. Springer, 2012.

Zou, D. and Gu, Q. On the convergence of hamiltonian monte carlo with stochastic gradients. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 13012–13022. PMLR, 2021.

A. Limitations of Current SDE Approaches to Analyze SG-based Algorithms

In this Section, we expose a limitation of the literature on SDE modeling of stochastic gradient-based sampling and optimization algorithms, which in our view is contributing to foster an imprecise understanding and use of such algorithms, as we discuss in the main paper. A common approach to analyze the errors introduced by the mini-batching is to take the limit for a step size going to zero (Chen et al., 2014). The derivations in the literature are based mainly on two assumptions: (i) the learning rate (or step size) is small enough such that transforming the discrete time update equations into continuous time ones is a valid approximation, and (ii) the mini-batch subsampling operation can be modelled, with the help of the central limit theorem, as an additional source of Gaussian noise that interferes with the dynamics. These assumptions, combined together, are used to justify the use of SDEs as models to study and analyze the dynamics of stochastic gradient-based algorithms. We mention for completeness, that some work (e.g. (Gurbuzbalaban et al., 2021)) model SG noise using *alpha-stable* distributions.

According to the standard practice (Chen et al., 2014; Mandt et al., 2017), we consider the potential $U(\cdot)$, a set of parameters $\boldsymbol{\theta}$, and a minibatch gradient $\nabla\tilde{U}(\boldsymbol{\theta})$; then, by means of central limit theorem we assume:

$$\nabla\tilde{U}(\boldsymbol{\theta}) = \nabla U(\boldsymbol{\theta}) + \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{18}$$

such that the discrete time dynamics of SGD becomes:

$$\begin{aligned} \delta\boldsymbol{\theta} &= -\eta\nabla U(\boldsymbol{\theta}) - \eta\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \\ &= -\eta\nabla U(\boldsymbol{\theta}) - \boldsymbol{\Sigma}^{1/2}\sqrt{\eta}\mathcal{N}(0, \eta\mathbf{I}). \end{aligned} \tag{19}$$

In the limit of vanishing step size, $\eta \to dt$, we can translate the discrete time dynamics into the continuous time ones as follows:

$$\boldsymbol{d\theta} = -\nabla U(\boldsymbol{\theta})dt + \boldsymbol{\Sigma}^{1/2}\sqrt{\eta}\mathbf{dw}(t), \tag{20}$$

where $\mathbf{w_t}$ is a standard Brownian motion. Loosely speaking, Brownian motion is a stochastic process whose differential satisfies $\mathbf{dw}(t) \sim \mathcal{N}(0, dt\mathbf{I})$, and whose increments are independent with respect to t. Notice that the stochastic term of Equation (19) involves two copies of η, in the covariance $\eta\boldsymbol{I}$ and in the external multiplication factor $\sqrt{\eta}$, but we consider $\eta \to dt$ only for the copy that scales the covariance of the Gaussian component.

The issue with this derivation becomes apparent by considering a proper stochastic calculus perspective, where both η terms are considered as infinitesimals, yielding $\sqrt{dt}\mathbf{dw}(t) = \mathbf{0}$. This is also explicitly noted by Yaida (2019), where the approximation $\eta^2 \simeq \eta dt$ was criticized. Considering the true limit, it is not possible to arbitrarily decide that the same quantity can be considered both as an infinitesimal and as an arbitrarily small but finite value. The matter is not only a mathematical subtlety. If one considers the correct formulation based on Fokker-Planck equations without going into continuous time but remaining in the discrete time (Leen & Moody, 1992), noncentral moments of the stochastic gradients need to be considered. Crucially, these induce quantitatively and qualitatively different results for the dynamics.

Note that our work is not in contrast with the literature whose purpose is the analysis of the stochastic gradient noise (Chaudhari & Soatto, 2018; Shi et al., 2020) or its approximate validity as a sampling method (Mandt et al., 2017; Franzese et al., 2021), where the formalism of SDEs is justified as a modelling tool. Our argument can be summarized as follows: if we are interested in the formal characterization of the convergence of these sampling schemes, such an assumption is invalid. When considering mini-batching in the context of sampling via SDE simulations, it is not correct to consider the stochastic gradients as an additional source of noise.

B. Additional Derivations for Section 2

B.1. Technical assumptions

In this Section we present the technical assumptions needed for the analysis in the paper.

We need to ensure that the Kolmogorov equation associated with the SDE Equation (3) has regular smooth solutions and unique invariant stationary distribution independent of initial conditions. To ensure smoothness and uniqueness , it is sufficient to prove that the infinitesimal generator satisfies certain regularity conditions (Hörmander, 1967; Mattingly et al., 2010), whereas to ensure ergodicity we require some growth conditions on the potential (Talay, 2002; Abdulle et al., 2015). Then, we need to assume that the functions $\phi(\cdot)$ satisfy certain conditions, to ensure that Taylor expansions of the form

Equation (22) are valid. Finally, we need to assume that the expected value of the considered numerical integration schemes can be validly expanded into series of the form Equation (23).

In the following we present sufficient conditions that guarantee the validity of the aforementioned assumptions. The reader interested in relaxing such technical conditions is referred to the original sources, such as Abdulle et al. (2014; 2015) and the works cited therein.

The first assumption guarantees that the Kolmogorov equation has regular unique smooth solutions and that the considered SDE is ergodic.

Assumption B.1. The potential $U(\boldsymbol{\theta})$ has continuous smooth bounded derivatives of arbitrary order, and it satisfies the growth condition

$$\boldsymbol{\theta}^\top \nabla U(\boldsymbol{\theta}) \geq R_1 ||\boldsymbol{\theta}||^2 - R_2, \tag{21}$$

for some $R_1, R_2 > 0$

It is important to notice that these requirements are independent from considerations on the choice of integrators, and without the assumption of ergodicity none of the SDE based sampling methods are conceptually meaningful.

Next, we restrict the class of considered $\phi(\cdot)$ functions to ensure that Taylor expansions of the transformation of equation is valid.

Assumption B.2. The function $\phi(\cdot)$ belongs to the set of functions l times continuously differentiable with all the partial derivatives with polynomial growth, for some positive $l > 0$. Furthermore there exists an $s \geq 0$ such that $|\phi(\boldsymbol{\theta})| \leq R_3(1 + |\boldsymbol{\theta}|^s)$ for some $R_3 > 0$.

This ensures that we can write

$$\exp(\eta\mathcal{L})\phi = \sum_{j=0}^{l} \frac{(\eta\mathcal{L})^j}{j!}\phi + \mathcal{O}(\eta^{l+1}), \tag{22}$$

for generic l, (Talay & Tubaro, 1990).

To be able to perform the same kind of expansions for numerical integrators, we need the following assumption.

Assumption B.3. Given an initial condition $\mathbf{z}_0$, any considered numerical integrator ψ is such that

$$|\mathbb{E}(|\mathbf{z}_1 - \mathbf{z}_0|)| \leq R_4(1 + |\mathbf{z}_0|)\eta, \quad |\mathbf{z}_1 - \mathbf{z}_0| \leq M_n(1 + |\mathbf{z}_0|)\sqrt{\eta},$$

with $\mathbf{z}_1 = \psi(\mathbf{z}_0)$, $R_4 > 0$ is independent of η, assumed small enough, and M_n has bounded moments of all orders independent of η.

This allows us to write the expansion of $\mathbb{E}[\phi(\mathbf{z_1})|\mathbf{z_0}]$ up to generic order z as

$$\mathbb{E}[\phi(\mathbf{z_1})|\mathbf{z_0}] = \mathcal{U}\phi = \left(\mathcal{I} + \eta\mathcal{A}_1 + \frac{\eta^2}{2}\mathcal{A}_2 + \ldots \frac{\eta^z}{z!}\mathcal{A}_z + \ldots\right)\phi + \mathcal{O}(\eta^{z+1}) \tag{23}$$

B.2. Proof of Theorem 2.1

We prove that the stationary distribution of Equation (3) is the posterior of interest (Gardiner, 2004). Start from the infinitesimal generator

$$\mathcal{L} = \underbrace{-\left(\nabla_{\boldsymbol{\theta}}^\top U(\boldsymbol{\theta})\right)\nabla_{\mathbf{r}} + \left(\mathbf{M}^{-1}\mathbf{r}\right)^\top \nabla_{\boldsymbol{\theta}}}_{\text{pure Hamiltonian evolution} \quad =\mathcal{H}} \underbrace{-C\left(\mathbf{M}^{-1}\mathbf{r}\right)^\top \nabla_{\mathbf{r}} + C\nabla_{\mathbf{r}}^\top \nabla_{\mathbf{r}}}_{\text{friction and Noise} \quad =\mathcal{D}}.$$

The adjoint of the first term, the pure Hamiltonian, can be rewritten as

$$\mathcal{H}^\dagger = \left(\nabla_{\boldsymbol{\theta}}^\top U(\boldsymbol{\theta})\right)\nabla_{\mathbf{r}} - \left(\mathbf{M}^{-1}\mathbf{r}\right)^\top \nabla_{\boldsymbol{\theta}} = \nabla_{\mathbf{z}}^\top H(\mathbf{z}) \begin{bmatrix} \mathbf{0} & -\boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0} \end{bmatrix} \nabla_{\mathbf{z}} \tag{24}$$

Assuming $\rho_{ss}(\mathbf{z}) \propto \exp\left(-H(\mathbf{z})\right)$ we have

$$\mathcal{H}^\dagger \rho_{ss}(\mathbf{z}) \propto \mathcal{H}^\dagger \exp\left(-H(\mathbf{z})\right) = \nabla_{\mathbf{z}}^\top H(\mathbf{z}) \begin{bmatrix} \mathbf{0} & -\boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0} \end{bmatrix} \left(\nabla_{\mathbf{z}} \exp\left(-H(\mathbf{z})\right)\right) =$$
$$- \nabla_{\mathbf{z}}^\top H(\mathbf{z}) \begin{bmatrix} \mathbf{0} & -\boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0} \end{bmatrix} \nabla_{\mathbf{z}} H(\mathbf{z}) \exp\left(-H(\mathbf{z})\right) = 0.$$

Similarly, considering the adjoint of $\mathcal{D}$

$$\mathcal{D}^\dagger = \nabla_{\mathbf{r}}^\top \left(C \left(\mathbf{M}^{-1}\mathbf{r}\right) \cdot\right) + C \nabla_{\mathbf{r}}^\top \nabla_{\mathbf{r}} = \nabla_{\mathbf{r}}^\top \left(C \left(\mathbf{M}^{-1}\mathbf{r}\right) \cdot + C \nabla_{\mathbf{r}}\right). \tag{25}$$

We study the term

$$\begin{aligned}
&\left(C \left(\mathbf{M}^{-1}\mathbf{r}\right) \cdot + C\nabla_{\mathbf{r}}\right) \rho_{ss}(\mathbf{z}) \propto \\
&\left(C \left(\mathbf{M}^{-1}\mathbf{r}\right) \cdot + C\nabla_{\mathbf{r}}\right) \exp(-U(\boldsymbol{\theta}) - 1/2 \left|\left|\mathbf{M}^{-1}\mathbf{r}\right|\right|^2) = \\
&\exp(-U(\boldsymbol{\theta})) \left(C \left(\mathbf{M}^{-1}\mathbf{r}\right) \exp(-1/2 \left|\left|\mathbf{M}^{-1}\mathbf{r}\right|\right|^2) + C\nabla_{\mathbf{r}} \left(\exp(-1/2 \left|\left|\mathbf{M}^{-1}\mathbf{r}\right|\right|^2)\right)\right) = \\
&\exp(-U(\boldsymbol{\theta})) \left(C \left(\mathbf{M}^{-1}\mathbf{r}\right) \exp(-1/2 \left|\left|\mathbf{M}^{-1}\mathbf{r}\right|\right|^2) - C \left(\mathbf{M}^{-1}\mathbf{r}\right) \exp(-1/2 \left|\left|\mathbf{M}^{-1}\mathbf{r}\right|\right|^2)\right) = 0.
\end{aligned}$$

Consequently $\mathcal{L}^\dagger \rho_{ss} = (\mathcal{H}^\dagger + \mathcal{D}^\dagger)\rho_{ss} = 0$, concluding the proof.

B.3. Ergodic order of convergence

In this section we present the result of Abdulle et al. (2015), that we use to state the order of generic schemes considered in this work.

Proposition B.4. *A sufficient condition for an integrator to be of a given ergodic error p, i.e. $e(\psi, \phi) = \mathcal{O}(\eta^p)$, is to have weak order p. This is not a necessary condition, as carefully described in Theorem B.5*

The following Theorem describes the generic conditions to achieve a given ergodic error.

Theorem B.5. *(Abdulle et al., 2015). Consider the assumptions of Section B.1. Suppose that the elements of the expansion of eq. Equation (23) satisfy*

$$\mathcal{A}_1 = \mathcal{L}, \quad \mathcal{A}_j^\dagger \rho = 0, \quad j = 2, \dots r. \tag{26}$$

Then the ergodic average has order of convergence r, i.e.

$$e(\psi, \phi) = \int \phi(\mathbf{z}) \rho_{ss}^{\psi}(\mathbf{z}) d\mathbf{z} - \int \phi(\mathbf{z}) \rho_{ss}(\mathbf{z}) d\mathbf{z} = \mathcal{O}(\eta^r). \tag{27}$$

The result can also be refined as follows:

$$e(\psi, \phi) = -\eta^r \int_0^\infty \int \left(\mathcal{A}_{r+1} \exp(t\mathcal{L})\phi(\mathbf{z})\right) \rho(\mathbf{z}) d\mathbf{z} + \mathcal{O}(\eta^{r+1}) \tag{28}$$

The proof is presented in Abdulle et al. (2015). A similar characterization has been attempted in Chen et al. (2015), where the requirement to achieve a given convergence rate is that the integrator has weak local error of some given order. We suggest to rely instead on the weaker requirements $\mathcal{A}_j^\dagger \rho = 0$ for $j = 1, \dots r$. In fact, if a method has local error of order k, then, by standard backward error analysis arguments, we have that: $\mathcal{U} = \mathcal{I} + \eta\mathcal{L} + \frac{\eta^2}{2}\mathcal{L}^2 + \dots \frac{\eta^k}{k!}\mathcal{L}^k + \mathcal{O}(\eta^{k+1})$ where we easily notice that $\mathcal{A}_j = \mathcal{L}^j$, with $j = 1, \dots k$. Consequently $\mathcal{A}_j^\dagger \rho = 0$, for $j = 1, \dots r$, since $\mathcal{L}^\dagger \rho = 0$ due to the Fokker Planck equation. In summary, the ergodic error order is at least the weak order of the integrator.

B.4. Quasi-symplectic integrators

In the context of numerical simulations, it is a known fact that integrators that preserve the underlying geometry (a.k.a. symplectic structure) perform better than generic ones (Milstein et al., 2002; Milstein & Tretyakov, 2003). Symplectic (Milstein et al., 2002) and quasi symplectic (Milstein & Tretyakov, 2003; Abdulle et al., 2014; Bou-Rabee & Owhadi, 2010)

SDEs have been studied in the past, albeit in a different context than Bayesian sampling. Recently, symplectic geometry has been revisited, to characterize generic optimization problems (Betancourt et al., 2018; França et al., 2020; 2021), whereas in this work we are interested only on sampling properties.

The stochastic evolution induced by the SDE in Equation (3), starting from generic initial conditions $\mathbf{z}(0)$, $\mathbf{z}(t) = \tau(\mathbf{z}(0))$, dissipates the volume of regions exponentially fast. In precise mathematical terms, the evolution has Jacobian $\boldsymbol{\Omega}_{m,n} = \frac{\partial \tau_m(z)}{\partial z_n}$ that satisfies $\det(\boldsymbol{\Omega}) = \exp\left(-C\mathrm{Tr}\left(\mathbf{M}^{-1}\right)t\right)$. The underlying geometry is in many aspects related to the one of symplectic manifolds, where volume is preserved exactly. It is natural to expect that numerical integrators that almost preserve the symplectic structure, defined as quasi-symplectic, when applied to systems as Equation (3), will perform better than generic integrators. An excellent exposition of the details of such statement is presented in (Milstein et al., 2002; Milstein & Tretyakov, 2003; 2007), while hereafter we present a self-contained, shorter, discussion.

We start our discussion by considering the definition of a generic *symplectic* mapping. This is an important condition, since it can be shown that the evolution of deterministic Hamiltonian ODE naturally respects this condition (França et al., 2021).

Definition B.6. A mapping $\boldsymbol{\kappa} : \mathbf{z}_0 = \left[\mathbf{r}_0^\top, \boldsymbol{\theta}_0^\top\right]^\top \to \mathbf{z}_1 = \left[\mathbf{r}_1^\top, \boldsymbol{\theta}_1^\top\right]^\top$ is said to be symplectic iff

$$\boldsymbol{\Omega}^\top \mathbf{J} \boldsymbol{\Omega} = \mathbf{J}, \tag{29}$$

where $\boldsymbol{\Omega}_{m,n} = \frac{\partial (z_1)_m}{\partial (z_0)_n}$ is the Jacobian. where $\mathbf{J} = \begin{bmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}$.

A direct consequence of the symplectic property is that a symplectic mapping preserves the volume. Precisely, if we consider a region $O \subseteq \mathbb{R}^{2D}$, the volume after the mapping $\boldsymbol{\kappa}$ remains unchanged

$$\mathrm{Vol}\left(\boldsymbol{\kappa}(O)\right) = \int_O |\det(\boldsymbol{\Omega}(\mathbf{z}))| d\mathbf{z} = \int_O d\mathbf{z}, \tag{30}$$

since from Equation (29) we easily derive, considering $\det(\boldsymbol{\Omega}^\top \mathbf{J} \boldsymbol{\Omega}) = \det(\boldsymbol{\Omega})^2$, that $\det(\boldsymbol{\Omega}) = \pm 1$.

In this work we are interested however in stochastic Hamiltonian evolutions in the form of Equation (3). The first simple consideration we can make is that in the limit of vanishing friction, i.e. $C \to 0$, the dynamics are symplectic mappings. This is derived considering that in such a limit the SDE becomes an Hamiltonian ODE and the symplectic property is proved as in (França et al., 2021). When $C \neq 0$, it is possible to show that the stochastic mapping contracts exponentially in time the volume of regions. The precise statement is reported in the following Theorem.

Theorem B.7. *Consider an Hamiltonian* SDE *of the form Equation* (3) *and the stochastic evolution induced by this equation starting from generic initial conditions, i.e.,* $\mathbf{z}(t) = \tau(\mathbf{z}(0))$. *Then* τ *is contracts the volume exponentially in time as* $\det(\boldsymbol{\Omega}) = \exp\left(-C\mathrm{Tr}\left(\mathbf{M}^{-1}\right)t\right)$, *where* $\boldsymbol{\Omega}_{m,n} = \frac{\partial \tau_m(z)}{\partial z_n}$ *is the (random) Jacobian.*

Proof. Starting from Equation (3), we have

$$d\mathbf{z}^\top(t) = \nabla^\top H(\mathbf{z}(t)) \begin{bmatrix} -C\boldsymbol{I} & -\boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0} \end{bmatrix}^\top dt + \begin{bmatrix} \sqrt{2C}\mathbf{dw}^\top(t) & \mathbf{0}^\top \end{bmatrix}. \tag{31}$$

Consequently,

$$\begin{aligned} d\nabla_{\mathbf{z}(0)}\mathbf{z}^\top(t) = \nabla_{\mathbf{z}(0)} d\mathbf{z}^\top(t) &= \nabla_{\mathbf{z}(0)} \nabla^\top H(\mathbf{z}(t)) \begin{bmatrix} -C\boldsymbol{I} & -\boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0} \end{bmatrix}^\top dt + \nabla_{\mathbf{z}(0)} \begin{bmatrix} \sqrt{2C}\mathbf{dw}^\top(t) & \mathbf{0}^\top \end{bmatrix} \\ &= \nabla_{\mathbf{z}(0)} \nabla^\top H(\mathbf{z}(t)) \begin{bmatrix} -C\boldsymbol{I} & -\boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0} \end{bmatrix}^\top dt = \nabla_{\mathbf{z}(0)} \mathbf{z}^\top(t) \Delta H(\mathbf{z}(t)) \begin{bmatrix} -C\boldsymbol{I} & -\boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0} \end{bmatrix}^\top dt, \end{aligned}$$

where we use the shorthand $\Delta = \nabla\nabla^\top$. The last expression can be rewritten as

$$\mathbf{d}\boldsymbol{\Omega}^\top(t) = \boldsymbol{\Omega}^\top(t) \begin{bmatrix} \mathbf{M}^{-1} & \mathbf{0} \\ \mathbf{0} & \Delta_{\boldsymbol{\theta}} U(\boldsymbol{\theta}(t)) \end{bmatrix} \begin{bmatrix} -C\boldsymbol{I} & -\boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0} \end{bmatrix}^\top dt.$$

By the rule of differentiation of determinants, we can write

$$
\begin{aligned}
&d\det(\boldsymbol{\Omega}^\top(t)) = \mathrm{Tr}\left(\mathrm{adj}\left(\boldsymbol{\Omega}^\top(t)\right)\mathbf{d}\boldsymbol{\Omega}^\top(t)\right) = \\
&\mathrm{Tr}\left(\mathrm{adj}\left(\boldsymbol{\Omega}^\top(t)\right)\boldsymbol{\Omega}^\top(t)\begin{bmatrix}\mathbf{M}^{-1} & \mathbf{0}\\ \mathbf{0} & \Delta_{\boldsymbol{\theta}}U(\boldsymbol{\theta}(t))\end{bmatrix}\begin{bmatrix}-C\boldsymbol{I} & -\boldsymbol{I}\\ \boldsymbol{I} & \mathbf{0}\end{bmatrix}^\top dt\right) = \\
&\det(\boldsymbol{\Omega}^\top(t))\mathrm{Tr}\left(\begin{bmatrix}\mathbf{M}^{-1} & \mathbf{0}\\ \mathbf{0} & \Delta_{\boldsymbol{\theta}}U(\boldsymbol{\theta}(t))\end{bmatrix}\begin{bmatrix}-C\boldsymbol{I} & -\boldsymbol{I}\\ \boldsymbol{I} & \mathbf{0}\end{bmatrix}^\top dt\right) = \\
&\det(\boldsymbol{\Omega}^\top(t))\mathrm{Tr}\left(\begin{bmatrix}\mathbf{M}^{-1} & \mathbf{0}\\ \mathbf{0} & \Delta_{\boldsymbol{\theta}}U(\boldsymbol{\theta}(t))\end{bmatrix}\begin{bmatrix}-C\boldsymbol{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}^\top dt\right) = -\det(\boldsymbol{\Omega}^\top(t))C\mathrm{Tr}\left(\mathbf{M}^{-1}\right)dt.
\end{aligned}
$$

This, toghether with the condition $\boldsymbol{\Omega}(0) = \boldsymbol{I}$ proves that

$$\det(\boldsymbol{\Omega}^\top(t)) = \exp(-C\mathrm{Tr}\left(\mathbf{M}^{-1}\right)t)$$

□

Having acknowledged the rich geometrical structure of Hamiltonian ODEs and SDEs, we are finally ready to introduce the concept of quasi-symplectic integrators (Milstein & Tretyakov, 2003), as numerical schemes that aim at preserving the underlying geometry.

Definition B.8. An numerical (stochastic) integrator $\phi : \mathbf{z}_i \to \mathbf{z}_{i+1}$ with step size η is defined to be quasi-symplectic if the following two properties holds:

1. The determinant of the Jacobian $\boldsymbol{\Omega} = (\nabla_{\mathbf{z}_{i-1}}\phi(\mathbf{z}_i)^\top)$ of the numerical integrator does not depend on $\mathbf{z}_i = (\mathbf{r}_i, \boldsymbol{\theta}_i)$
2. In the limit of vanishing friction, i.e. $C \to 0$, the Jacobian $\boldsymbol{\Omega}$ satisfies the symplectic condition

$$\boldsymbol{\Omega}^\top\mathbf{J}\boldsymbol{\Omega} = \mathbf{J}. \tag{32}$$

B.5. Numerical integrators

In this section, we present some examples of integrators, which we explore in the experimental Section 5 in the main paper, with their corresponding order of convergence.

B.5.1. LEAPFROG

The second integrator we consider is the widely used LEAPFROG scheme

$$
\begin{cases}
\boldsymbol{\theta}^* = \boldsymbol{\theta}_{i-1} + \frac{\eta}{2}\mathbf{M}^{-1}\mathbf{r}_{i-1}\\
\mathbf{r}_i = \mathbf{r}_{i-1} - \eta\nabla U(\boldsymbol{\theta}^*) - \eta C\mathbf{M}^{-1}\mathbf{r}_{i-1} + \sqrt{2C\eta}\mathbf{w}, \quad \mathbf{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})\\
\boldsymbol{\theta}_i = \boldsymbol{\theta}^* + \frac{\eta}{2}\mathbf{M}^{-1}\mathbf{r}_i
\end{cases} \tag{33}
$$

This scheme is quasi-symplectic as shown hereafter. We rewrite equivalently the update scheme as a unique step as

$$
\begin{cases}
\mathbf{r}_i = \mathbf{r}_{i-1} - \eta\nabla U(\boldsymbol{\theta}_{i-1} + \frac{\eta}{2}\mathbf{M}^{-1}\mathbf{r}_{i-1}) - \eta C\mathbf{M}^{-1}\mathbf{r}_{i-1} + \sqrt{2C\eta}\mathbf{w}\\
\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} + \frac{\eta}{2}\mathbf{M}^{-1}\mathbf{r}_{i-1} + \frac{\eta}{2}\mathbf{M}^{-1}\left(\mathbf{r}_{i-1} - \eta\nabla U(\boldsymbol{\theta}_{i-1} + \frac{\eta}{2}\mathbf{M}^{-1}\mathbf{r}_{i-1}) - \eta C\mathbf{M}^{-1}\mathbf{r}_{i-1} + \sqrt{2C\eta}\mathbf{w}\right)
\end{cases} \tag{34}
$$

Define $\mathbf{q} = \boldsymbol{\theta}_{i-1} + \frac{\eta}{2}\mathbf{M}^{-1}\mathbf{r}_{i-1}$, $q^{(k)} = \theta_{i-1}^{(k)} + \eta m_{kp}r_{i-1}^{(p)}$, with $\mathbf{M}^{-1}|_{kp} = m_{kp}$.

$$\frac{\partial}{\partial r_{i-1}^{(m)}}\partial_n U(\mathbf{q}) = \partial_{nk}U(\mathbf{q})\frac{q^{(k)}}{\partial r_{i-1}^{(m)}} = \partial_{nk}U(\mathbf{q})\frac{\eta}{2}m_{kp}\delta_{pm} = \frac{\eta}{2}\partial_{nk}U(\mathbf{q})m_{km}. \tag{35}$$

Consequently $\nabla_{\mathbf{r}_{i-1}} \nabla^\top U(\mathbf{q}) = \frac{\eta}{2}\mathbf{M}^{-1}\Delta U(\mathbf{q})$

$$\boldsymbol{\Omega}^\top = \begin{bmatrix} \boldsymbol{I} - \eta C\mathbf{M}^{-1} - \frac{\eta^2}{2}\mathbf{M}^{-1}\Delta U(\mathbf{q}) & \eta\mathbf{M}^{-1} - \frac{\eta^3}{4}\mathbf{M}^{-2}\Delta U(\mathbf{q}) - \frac{\eta^2}{2}C\mathbf{M}^{-2} \\ -\eta\Delta U(\mathbf{q}) & \boldsymbol{I} - \frac{\eta^2}{2}\mathbf{M}^{-1}\Delta U(\mathbf{q}) \end{bmatrix} \tag{36}$$

Simple algebraic manipulations show that the transposed Jacobian can be rewritten as

$$\boldsymbol{\Omega}^\top = \begin{bmatrix} \boldsymbol{I} & \frac{\eta}{2}\mathbf{M}^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} - \eta C\mathbf{M}^{-1} & \mathbf{0} \\ -\eta\Delta U(\mathbf{q}) & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \frac{\eta}{2}\mathbf{M}^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \tag{37}$$

The determinant of the Jacobian is then easily calculated as the product of the three determinants

$$\det(\boldsymbol{\Omega}^\top) = \det \begin{bmatrix} \boldsymbol{I} & \frac{\eta}{2}\mathbf{M}^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \det \begin{bmatrix} \boldsymbol{I} - \eta C\mathbf{M}^{-1} & \mathbf{0} \\ -\eta\Delta U(\mathbf{q}) & \boldsymbol{I} \end{bmatrix} \det \begin{bmatrix} \boldsymbol{I} & \frac{\eta}{2}\mathbf{M}^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} = \det\left(\boldsymbol{I} - \eta C\mathbf{M}^{-1}\right), \tag{38}$$

and is independent on $\mathbf{z}_0$, satisfying the first condition. By taking the limit $C \to 0$, to prove that the integrator converge to a symplectic one is sufficient to prove that

$$\begin{bmatrix} \boldsymbol{I} & \frac{\eta}{2}\mathbf{M}^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \frac{\eta}{2}\mathbf{M}^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix}^\top = \begin{bmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}, \tag{39}$$

and that

$$\begin{bmatrix} \boldsymbol{I} & \mathbf{0} \\ -\eta\Delta U(\mathbf{q}) & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \mathbf{0} \\ -\eta\Delta U(\mathbf{q}) & \boldsymbol{I} \end{bmatrix}^\top = \begin{bmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}. \tag{40}$$

Simple calculations show that

$$\begin{bmatrix} \boldsymbol{I} & \frac{\eta}{2}\mathbf{M}^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \frac{\eta}{2}\mathbf{M}^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix}^\top = \begin{bmatrix} \frac{\eta}{2}\mathbf{M}^{-1} & -\boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \mathbf{0} \\ \frac{\eta}{2}\mathbf{M}^{-1} & \boldsymbol{I} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix},$$

and similarly we can prove eq. Equation (40), completing the proof that LEAPFROG is quasi-symplectic. This integrator has a theoretical order of convergence equal to one but its performance is on par with other integrators of higher order.

B.5.2. MT3

We then consider the quasi-symplectic scheme of third order that can be found in Section 2.3 of Milstein & Tretyakov (2003):

$$\begin{aligned} \boldsymbol{\theta}_1 &= \boldsymbol{\theta}_k + \frac{7}{24}\eta\mathbf{M}^{-1}\mathbf{r}_k \\ \mathbf{r}_1 &= \mathbf{r}_k + \frac{7}{24}\eta\left[-\nabla U(\boldsymbol{\theta}_1) - C\mathbf{M}^{-1}\mathbf{r}_1\right] \end{aligned} \tag{41}$$

$$\begin{aligned} \boldsymbol{\theta}_2 &= \boldsymbol{\theta}_k + \frac{25}{24}\eta\mathbf{M}^{-1}\mathbf{r}_k + \frac{\eta^2}{2}\mathbf{M}^{-1}\left[-\nabla U(\boldsymbol{\theta}_1) - C\mathbf{M}^{-1}\mathbf{r}_1\right] \\ \mathbf{r}_2 &= \mathbf{r}_k + \frac{2}{3}\eta\left[-\nabla U(\boldsymbol{\theta}_1) - C\mathbf{M}^{-1}\mathbf{r}_1\right] + \frac{3}{8}\eta\left[-\nabla U(\boldsymbol{\theta}_2) - C\mathbf{M}^{-1}\mathbf{r}_2\right] \end{aligned} \tag{42}$$

$$\begin{aligned} \boldsymbol{\theta}_3 &= \boldsymbol{\theta}_k + \eta\mathbf{M}^{-1}\mathbf{r}_k + \frac{17}{36}\eta^2\mathbf{M}^{-1}\left[-\nabla U(\boldsymbol{\theta}_1) - C\mathbf{M}^{-1}\mathbf{r}_1\right] + \frac{1}{36}\eta^2\mathbf{M}^{-1}\left[-\nabla U(\boldsymbol{\theta}_2) - C\mathbf{M}^{-1}\mathbf{r}_2\right] \\ \mathbf{r}_3 &= \mathbf{r}_k + \frac{2}{3}\eta\left[-\nabla U(\boldsymbol{\theta}_1) - C\mathbf{M}^{-1}\mathbf{r}_1\right] - \frac{2}{3}\eta\left[-\nabla U(\boldsymbol{\theta}_2) - C\mathbf{M}^{-1}\mathbf{r}_2\right] + \eta\left[-\nabla U(\boldsymbol{\theta}_3) - C\mathbf{M}^{-1}\mathbf{r}_3\right] \end{aligned} \tag{43}$$

$$\begin{aligned}\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_3 + \eta^{3/2}\mathbf{M}^{-1}\sqrt{2C}\left(\mathbf{w}_1^{(k)}/2 + \mathbf{w}_2^{(k)}\right) \\ - C\eta^{5/2}\mathbf{M}^{-2}\sqrt{2C}\,\mathbf{w}_1^{(k)}/6\end{aligned} \tag{44}$$

$$\begin{aligned}\mathbf{r}_{k+1} = \mathbf{r}_3 + \eta^{1/2}\sqrt{2C}\,\mathbf{w}_1^{(k)} - C\eta^{3/2}\mathbf{M}^{-1}\sqrt{2C}\,\left(\mathbf{w}_1^{(k)}/2 + \mathbf{w}_2^{(k)}\right) \\ + \eta^{5/2}\sum_{j=1}^{n}\left[\sum_{i=1}^{n}\left(\mathbf{M}^{-1}\sqrt{2C}\,\mathbf{e_j}\right)^i \frac{\partial}{\partial\theta^i}(-\nabla U(\boldsymbol{\theta}_3))\right]\mathbf{w}_{1,j}^{(k)}/6 \\ + \eta^{5/2}C^2\mathbf{M}^{-2}\sqrt{2C}\,\mathbf{w}_1^{(k)}/6\end{aligned} \tag{45}$$

B.5.3. LIE-TROTTER

Finally we consider an order-two Lie-Trotter splitting scheme (Abdulle et al., 2015)

$$\begin{cases}\boldsymbol{\theta}^* = \boldsymbol{\theta}_{i-1} + \frac{\eta}{2}\mathbf{M}^{-1}\mathbf{r}_{i-1} \\ \mathbf{r}^* = \mathbf{r}_{i-1} - \eta\nabla U(\boldsymbol{\theta}^*) \\ \boldsymbol{\theta}^i = \boldsymbol{\theta}^* + \frac{\eta}{2}\mathbf{M}^{-1}\mathbf{r}^* \\ \mathbf{r}_i = \exp(-C\mathbf{M}^{-1}\eta)\mathbf{r}^* + \sqrt{\mathbf{M}(1-\exp(-2C\mathbf{M}^{-1}\eta))}\mathbf{w}, \quad \mathbf{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})\end{cases} \tag{46}$$

We choose the particular case of a deterministic leapfrog for the Hamiltonian integration, but any other integrator ψ would have been valid. To derive the order of convergence, it is sufficient to use the result of Theorem B.5 as in Abdulle et al. (2014).

The scheme can be interpreted as the cascade of a purely deterministic symplectic integrator (the leapfrog step)

$$\begin{cases}\mathbf{r}_{i-1} \\ \boldsymbol{\theta}_{i-1}\end{cases} \to \begin{cases}\mathbf{r}^* = \mathbf{r}_{i-1} - \eta\nabla U(\boldsymbol{\theta}_{i-1} + \frac{\eta}{2}\mathbf{M}^{-1}\mathbf{r}_{i-1}) \\ \boldsymbol{\theta}^* = \boldsymbol{\theta}_{i-1} + \frac{\eta}{2}\mathbf{M}^{-1}\mathbf{r}_{i-1} + \frac{\eta}{2}\mathbf{M}^{-1}\left(\mathbf{r}_{i-1} - \eta\nabla U(\boldsymbol{\theta}_{i-1} + \frac{\eta}{2}\mathbf{M}^{-1}\mathbf{r}_{i-1})\right)\end{cases} \tag{47}$$

and the analytic step

$$\begin{cases}\mathbf{r}^* \\ \boldsymbol{\theta}^*\end{cases} \to \begin{cases}\mathbf{r}_i = \exp(-C\mathbf{M}^{-1}\eta)\mathbf{r}^* + \sqrt{\mathbf{M}(1-\exp(-2C\mathbf{M}^{-1}\eta))}\mathbf{w} \\ \boldsymbol{\theta}_i = \boldsymbol{\theta}^*\end{cases} \tag{48}$$

With calculations similar to the ones in Section B.5.1 it is easy to show that the purely deterministic leapfrog step is symplectic, i.e., its Jacobian $\boldsymbol{\Omega}_1$ satisfies $\boldsymbol{\Omega}_1^\top\mathbf{J}\boldsymbol{\Omega}_1 = \mathbf{J}$.

Concerning the analytical step instead, it is easy to show that the Jacobian of the transformation is equal to

$$\boldsymbol{\Omega}_2^\top = \begin{bmatrix}\exp(-C\mathbf{M}^{-1}\eta) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}\end{bmatrix}. \tag{49}$$

Consequently, the overall Jacobian of the two steps taken jointly is

$$\boldsymbol{\Omega}^\top = \boldsymbol{\Omega}_1^\top\boldsymbol{\Omega}_2^\top \tag{50}$$

and it is easily shown to be satisfying the two requirements for the quasi-symplecticity. The LIE-TROTTER scheme is then quasi-symplectic.

C. Additional Derivations for Section 3

We present the Baker–Campbell–Hausdorff in Appendix C.1, useful for technical derivations in the rest of the section. Appendix C.2 contains results on the average of back and forth of multiple operators (Theorem 3.1). In Appendix C.3 we study the effect of chaining multiple (different) numerical integrators. Finally, Appendix C.4 combines the elements of the previous Sections to provide the proof for Theorem 3.2. Appendix C.5 contains the Proof of Theorem 3.3.

C.1. Baker–Campbell–Hausdorff formula

We here present the Baker–Campbell–Hausdorff formula, and in particular the series expansion due to Dynkin (1947), that will prove useful in the technical derivations. Given two differential operators $\mathcal{A}, \mathcal{B}$, we have that

$$\exp(\mathcal{A})\exp(\mathcal{B}) = \exp(\mathcal{Z}) \tag{51}$$

where $\mathcal{Z} = \mathcal{A} + \mathcal{B} + \frac{1}{2}[\mathcal{A},\mathcal{B}] + \frac{1}{12}\left([\mathcal{A},[\mathcal{A},\mathcal{B}]] + [\mathcal{B},[\mathcal{B},\mathcal{A}]]\right) - \frac{1}{24}[\mathcal{B},[\mathcal{A},[\mathcal{A},\mathcal{B}]]]$
$- \frac{1}{720}\left([\mathcal{B},[\mathcal{B},[\mathcal{B},[\mathcal{B},\mathcal{A}]]]] + [\mathcal{A},[\mathcal{A},[\mathcal{A},[\mathcal{A},\mathcal{B}]]]]\right) + \frac{1}{360}\left([\mathcal{A},[\mathcal{B},[\mathcal{B},[\mathcal{B},\mathcal{A}]]]] + [\mathcal{B},[\mathcal{A},[\mathcal{A},[\mathcal{A},\mathcal{B}]]]]\right)$
$+ \frac{1}{120}\left([\mathcal{B},[\mathcal{A},[\mathcal{B},[\mathcal{A},\mathcal{B}]]]] + [\mathcal{A},[\mathcal{B},[\mathcal{A},[\mathcal{B},\mathcal{A}]]]]\right) + \cdots$

C.2. Average of back and forth operators — Proof of Theorem 3.1

Theorem 3.1 provides a technical support for the proofs of the order of convergence of chains of operators.

Proof. The proof of Theorem 3.1 is as follows:

$$\exp\left(\eta K \mathcal{L}_i\right) = \mathcal{I} + \eta K \mathcal{L}_i + \frac{\eta^2 K^2 \mathcal{L}_i^2}{2} + \mathcal{O}(\eta^3) \tag{52}$$

Consequently:

$$\prod_{i=1}^{K}\left(\mathcal{I} + \eta K \mathcal{L}_i + \frac{\eta^2 K^2 \mathcal{L}_i^2}{2} + \mathcal{O}(\eta^3)\right) = \mathcal{I} + \eta\sum_{i=1}^{K} K\mathcal{L}_i + \frac{\eta^2 K^2}{2}\sum_{i=1}^{K}\mathcal{L}_i^2 + \eta^2 K^2 \sum_{\substack{i_1=1\ldots K-1,\\ i_2=i_1+1\ldots K}} \mathcal{L}_{i_1}\mathcal{L}_{i_2} + \mathcal{O}((K-1)\eta^3) \tag{53}$$

and

$$\prod_{i=K}^{1}\left(\mathcal{I} + \eta K \mathcal{L}_i + \frac{\eta^2 K^2 \mathcal{L}_i^2}{2} + \mathcal{O}(\eta^3)\right) = \mathcal{I} + \eta\sum_{i=1}^{K} K\mathcal{L}_i + \frac{\eta^2 K^2}{2}\sum_{i=1}^{K}\mathcal{L}_i^2 + \eta^2 K^2 \sum_{\substack{i_1=2\ldots K,\\ i_2=1\ldots i_1-1}} \mathcal{L}_{i_1}\mathcal{L}_{i_2} + \mathcal{O}((K-1)\eta^3) \tag{54}$$

The fundamental equality of interest is:

$$\begin{aligned}
\mathcal{L}^2 &= \left(\sum_{i_1=1}^{K}\mathcal{L}_{i_1}\right)\left(\sum_{i_2=1}^{K}\mathcal{L}_{i_2}\right) = \left(\sum_{\substack{i_1=1\ldots K,\\ i_2=1\ldots K}}\mathcal{L}_{i_1}\mathcal{L}_{i_2}\right) = \left(\sum_{\substack{i_1=1\ldots K,\\ i_2=i_1\ldots K}}\mathcal{L}_{i_1}\mathcal{L}_{i_2} + \sum_{\substack{i_1=1\ldots K,\\ i_2=1\ldots i_1-1}}\mathcal{L}_{i_1}\mathcal{L}_{i_2}\right) \\
&= \left(\sum_{\substack{i_1=1\ldots K,\\ i_2=i_1}}\mathcal{L}_{i_1}\mathcal{L}_{i_2} + \sum_{\substack{i_1=1\ldots K,\\ i_2=i_1+1\ldots K}}\mathcal{L}_{i_1}\mathcal{L}_{i_2} + \sum_{\substack{i_1=1\ldots K,\\ i_2=1\ldots i_1-1}}\mathcal{L}_{i_1}\mathcal{L}_{i_2}\right) \\
&= \left(\sum_{\substack{i_1=1\ldots K,\\ i_2=i_1}}\mathcal{L}_{i_1}\mathcal{L}_{i_2} + \sum_{\substack{i_1=1\ldots K,\\ i_2=i_1+1\ldots K}}\mathcal{L}_{i_1}\mathcal{L}_{i_2} + \sum_{\substack{i_1=2\ldots K,\\ i_2=1\ldots i_1-1}}\mathcal{L}_{i_1}\mathcal{L}_{i_2}\right)
\end{aligned}$$

By means of simple algebric considerations we can thus obtain the desired result, as

$$\begin{aligned}
\mathcal{P} &= \frac{\prod_{i=K}^{1}\exp\left(\eta K\mathcal{L}_i\right) + \prod_{i=1}^{K}\exp\left(\eta K\mathcal{L}_i\right)}{2} \\
&= \frac{\prod_{i=1}^{K}\left(\mathcal{I} + \eta K\mathcal{L}_i + \frac{\eta^2K^2\mathcal{L}_i^2}{2} + \mathcal{O}(\eta^3)\right) + \prod_{i=K}^{1}\left(\mathcal{I} + \eta K\mathcal{L}_i + \frac{\eta^2K^2\mathcal{L}_i^2}{2} + \mathcal{O}(\eta^3)\right)}{2} \\
&= \mathcal{I} + \eta\sum_{i=1}^{K}K\mathcal{L}_i + \frac{\eta^2K^2}{2}\left(\left(\sum_{i_1=1}^{K}\mathcal{L}_{i_1}\right)\left(\sum_{i_2=1}^{K}\mathcal{L}_{i_2}\right)\right) + \mathcal{O}((K-1)\eta^3) = \exp(\eta K\mathcal{L}) + \mathcal{O}(K\eta^3)
\end{aligned}$$

□

Also, we stress that this bound is sharp, i.e. the back and forth combination of operators **is not** $\mathcal{O}(\eta^4)$. In fact, this is a known result (Castella et al., 2009). We fully elaborate considering the case $K = 2$. Given differential operator $\mathcal{C} = \mathcal{A} + \mathcal{B}$, we have:

$$\exp(\eta\mathcal{A})\exp(\eta\mathcal{B}) = \mathcal{I} + \eta(\mathcal{A}+\mathcal{B}) + \frac{\eta^2}{2}(\mathcal{A}^2 + 2\mathcal{A}\mathcal{B} + \mathcal{B}^2) + \frac{\eta^3}{6}\underbrace{\left(\mathcal{A}^3 + 2\mathcal{A}^2\mathcal{B} + \mathcal{A}\mathcal{B}^2 + \mathcal{B}\mathcal{A}^2 + 2\mathcal{A}\mathcal{B}^2 + \mathcal{B}^3\right)}_{q_0} + \mathcal{O}(\eta^4), \tag{55}$$

and similarly:

$$\exp(\eta\mathcal{B})\exp(\eta\mathcal{A}) = \mathcal{I} + \eta(\mathcal{B}+\mathcal{A}) + \frac{\eta^2}{2}(\mathcal{B}^2 + 2\mathcal{B}\mathcal{A} + \mathcal{A}^2) + \frac{\eta^3}{6}\underbrace{\left(\mathcal{B}^3 + 2\mathcal{B}^2\mathcal{A} + \mathcal{B}\mathcal{A}^2 + \mathcal{A}\mathcal{B}^2 + 2\mathcal{B}\mathcal{A}^2 + \mathcal{A}^3\right)}_{q_1} + \mathcal{O}(\eta^4). \tag{56}$$

While the term $\mathcal{C}^2 = \mathcal{A}^2 + \mathcal{A}\mathcal{B} + \mathcal{B}\mathcal{A} + \mathcal{B}^2$ can be obtained by averaging back and forth, the term $\mathcal{C}^3 = \mathcal{A}^3 + \mathcal{A}^2\mathcal{B} + \mathcal{A}\mathcal{B}\mathcal{A} + \mathcal{A}\mathcal{B}^2 + \mathcal{B}\mathcal{A}^2 + \mathcal{B}\mathcal{A}\mathcal{B} + \mathcal{B}^2\mathcal{A} + \mathcal{B}^3$ is not. To see this intuitively, it is sufficient to notice that in the average $\frac{q_0+q_1}{2}$ terms $\mathcal{A}\mathcal{B}\mathcal{A}$, $\mathcal{B}\mathcal{A}\mathcal{B}$ are missing.

To increase the order of the considered splitting it is necessary to chain multiple stages with *negative* step coefficients. As acknowledged in (Castella et al., 2009), this poses a serious limitation whenever semi-groups are considered, as in our case. To solve the matter we should consider *imaginary* step-sizes, that we leave as an open possibility for future works.

C.3. Randomized sequence of numerical integration steps

The first aspect we consider is the net effect of chaining multiple (different) numerical integrators on the transformation of ϕ.

Theorem C.1. *Consider a set of numerical integrators $\{\psi_i\}_{i=1}^K$ with corresponding functionals $\mathcal{U}_i$. If we build the stochastic variable $\mathbf{z}_{fin}$ starting from $\mathbf{z}_0$ through the following chain*

$$\mathbf{z}_{fin} = \psi_K(\psi_{K-1}(\dots\psi_1(\mathbf{z}_0))), \tag{57}$$

then the generic function ϕ is transformed as follows

$$\mathbb{E}\left[\phi(\mathbf{z}_{fin})\right] = \mathcal{U}_1\mathcal{U}_2\dots\mathcal{U}_K\phi(\mathbf{z}_0) \tag{58}$$

Notice the backward order of the differential operators and the difference with the ordering assumed in Chen et al. (2015).

Proof. We give the proof for $K = 2$ since the general result can be obtained by an extension of this case. In this case $\mathbf{z}_{fin} = \psi_2(\psi_1(\mathbf{z}_0))$. Define $\mathbf{z}_1 = \psi_1(\mathbf{z}_0)$. Since

$$\mathbb{E}\left[\phi(\mathbf{z}_{fin})\right] = \int \phi(\mathbf{z}_{fin})p(\mathbf{z}_{fin}|\mathbf{z}_0)d\mathbf{z}_{fin} = \int \phi(\mathbf{z}_{fin})p(\mathbf{z}_{fin}|\mathbf{z}_1)p(\mathbf{z}_1|\mathbf{z}_0)d\mathbf{z}_1 d\mathbf{z}_{fin} =$$
$$\int\left(\int \phi(\mathbf{z}_{fin})p(\mathbf{z}_{fin}|\mathbf{z}_1)d\mathbf{z}_{fin}\right)p(\mathbf{z}_1|\mathbf{z}_0)d\mathbf{z}_1 = \int\left(\mathcal{U}_2\phi(\mathbf{z}_1)\right)p(\mathbf{z}_1|\mathbf{z}_0)d\mathbf{z}_1 = \mathcal{U}_1\mathcal{U}_2\phi(\mathbf{z}_0)$$

the result is proven. □

Next, we consider the case in which we consider different chains of numerical integrators and apply one of them according to some probability distribution.

Theorem C.2. *Consider multiple sets of numerical integrators $\{\psi_{i,j}\}_{i=1}^K$ with corresponding functionals $\mathcal{U}_{i,j}$, with $j = 1\dots S$. Choose an index $j^* \in [1,\dots S]$ according to some probability distribution $p(j^* = j) = w_j$. Build the stochastic variable $\mathbf{z}_{fin}$ starting from $\mathbf{z}_0$ through the chain corresponding to the sampled index*

$$\mathbf{z}_{fin} = \psi_{K,j^*}(\psi_{K-1,j^*}(\dots\psi_{1,j^*}(\mathbf{z}_0))), \tag{59}$$

then the generic function ϕ is transformed as follows

$$\mathbb{E}\left[\phi(\mathbf{z}_{fin})\right] = \sum_{j=1}^{S} w_j \mathcal{U}_{1,j}\mathcal{U}_{2,j}\dots\mathcal{U}_{K,j}\phi(\mathbf{z}_0). \tag{60}$$

The result is derived considering the previous theorem and the law of total expectation. Indeed

$$\mathbb{E}\left[\phi(\mathbf{z}_{fin})\right] = \mathbb{E}_{j^*}\left[\mathbb{E}\left[\phi(\psi_{K,j^*}(\psi_{K-1,j^*}(\dots\psi_{1,j^*}(\mathbf{z}_0))))\right]\right] =$$

$$\mathbb{E}_{j^*}\left[\mathcal{U}_{1,j^*}\mathcal{U}_{2,j^*}\dots\mathcal{U}_{K,j^*}\phi(\mathbf{z}_0)\right] = \sum_{j=1}^{S} w_j \mathcal{U}_{1,j}\mathcal{U}_{2,j}\dots\mathcal{U}_{K,j}\phi(\mathbf{z}_0).$$

C.4. Proof of Theorem 3.2

As stated in Theorem 3.2, we consider numerical integrators ψ_i with order p. Consequently, we have representation $\mathcal{U}_i = \exp(\eta K\mathcal{L}_i) + \mathcal{O}(\eta^{p+1})$ of the operators $\mathcal{U}_i$ corresponding to the integrators ψ_i. The proof of Theorem 3.2 is obtained considering Theorem 3.1 and Theorem C.2.

Consider $K!$ sets of numerical integrators as follows:

$$\{\psi_{\pi_j^1}\}_{j=1}^{K}, \quad \{\psi_{\pi_j^2}\}_{j=1}^{K}, \dots \{\psi_{\pi_j^{K!}}\}_{j=1}^{K}, \tag{61}$$

where $\boldsymbol{\pi}^1, \dots \boldsymbol{\pi}^{K!}$ are all the possible permutations of the K operators, sampled uniformly with probability $\frac{1}{K!}$. If, in accordance with the hypotheses of Theorem 3.2, the initial condition has stochastic evolution through a chain of randomly permuted integrators, then the propagation operator is:

$$\mathcal{U} = \sum_{j=1}^{K!} \frac{1}{K!} \mathcal{U}_{\pi_K^j}\mathcal{U}_{\pi_{K-1}^j}\dots\mathcal{U}_{\pi_1^j}. \tag{62}$$

The proof is a direct consequence of Theorem C.2.

It is possible to rewrite the set of all possible permutations as the union of two sets of equal cardinality, such that each element of one of the two set has a (unique) element of the other set that is its lexicographic reverse.

$$\{\psi_{\pi_j^1}\}_{j=1}^{K}, \quad \{\psi_{\pi_j^2}\}_{j=1}^{K}, \dots \{\psi_{\pi_j^{K!/2}}\}_{j=1}^{K}$$
$$\{\psi_{\pi_j^1}\}_{j=K}^{1}, \quad \{\psi_{\pi_j^2}\}_{j=K}^{1}, \dots \{\psi_{\pi_j^{K!/2}}\}_{j=K}^{1}.$$

We rewrite the sum then as:

$$\mathcal{U} = \sum_{j=1}^{K!/2} \frac{2}{K!} \left(\frac{1}{2}\mathcal{U}_{\pi_K^j}\mathcal{U}_{\pi_{K-1}^j}\dots\mathcal{U}_{\pi_1^j} + \frac{1}{2}\mathcal{U}_{\pi_1^j}\mathcal{U}_{\pi_2^j}\dots\mathcal{U}_{\pi_K^j}\right).$$

Then, by Theorem 3.1, since:

$$\frac{1}{2}\mathcal{U}_{\pi_K^j}\mathcal{U}_{\pi_{K-1}^j}\dots\mathcal{U}_{\pi_1^j} + \frac{1}{2}\mathcal{U}_{\pi_1^j}\mathcal{U}_{\pi_2^j}\dots\mathcal{U}_{\pi_K^j} = \exp\left(\eta K\mathcal{L}\right) + \mathcal{O}(K\eta^3) + \mathcal{O}(K\eta^{p+1}) \tag{63}$$

we have

$$\mathcal{U} = \exp\left(\eta K\mathcal{L}\right) + \mathcal{O}(K\eta^3) + \mathcal{O}(K\eta^{p+1}). \tag{64}$$

Higher than order-3? It is natural to question whether the randomized ordering bound can be improved to some order > 3, by carefully collecting together different terms. Unfortunately, the answer is negative. It is indeed simple to see that whenever $K = 2$, even by considering a perfect numerical integrator ($p = \infty$) the rate of convergence is order-3. This claim is proven by considering Theorem 3.1 with $K = 2$, and noticing that in this particular case average of back and forth and randomized ordering coincides. Then, for generic K, it is not possible to achieve any higher order.

The intuition of the result lies in the missing of terms $\mathcal{L}_i\mathcal{L}_j\mathcal{L}_i$ in the products. Interestingly, and mimicking the results discussed at the end ofAppendix C.2, a possible explored solution to increase the order of convergence is to consider negative evolution operators (Zhang, 2012). However, this is not compatible with our settings unless we are willing to consider complex valued numerical schemes.

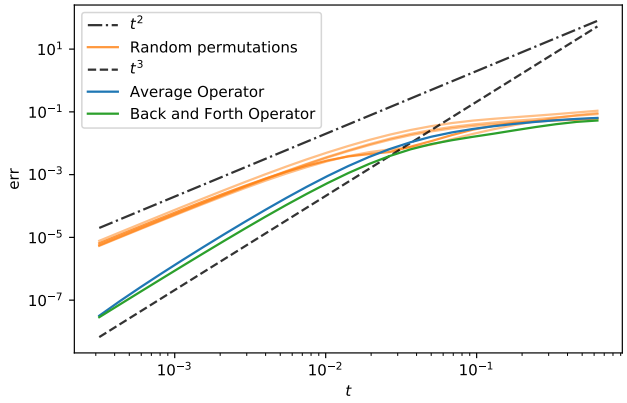

Figure 6. Local error as function of time t for operator splitting in a continuous-time Markov chain.

EXAMPLE: RANDOMIZED OPERATOR SPLITTING FOR A CONTINUOUS-TIME MARKOV CHAIN

We shall demonstrate Theorem 3.2 on a discrete linear operator featuring an analytic solution, so that we can appreciate the effect of operator splitting in isolation from other sources of error. More specifically, we consider a continuous-time Markov chain with infinitesimal generator matrix $\mathbf{Q}$, so that the transition probability matrix at time t will be:

$$\mathbf{P}_t = \exp(t\mathbf{Q}). \tag{65}$$

This matrix exponential will allows us to calculate the state space probabilities after time t exactly; starting from an initial distribution vector $\boldsymbol{p}_0$ we have:

$$\boldsymbol{p}_t^\top = \boldsymbol{p}_0^\top \mathbf{P}_t. \tag{66}$$

In this setting, there is not any time-discretization or Monte Carlo error. Of course, this practice is only applicable for small-size toy models; in our example, we consider a Markov chain with 8 states, or equivalently $\mathbf{Q} \in \mathbb{R}^{8\times 8}$.

In the experiment of Figure 6, we measure the mean absolute error of the state probabilities for different splitting strategies. We consider a randomly generated $\mathbf{Q}$ and a random initial vector $\boldsymbol{p}_0$, which both remain fixed for the rest of the experiment. Given a decomposition of $\mathbf{Q}$ into six matrices and an arbitrary permutation π, the state probabilities at time t are approximated as follows:

$$\tilde{\boldsymbol{p}}_t^\top = \boldsymbol{p}_0^\top \prod_{i=1}^{6} \exp\left(t\mathbf{Q}_{\pi_i}\right). \tag{67}$$

The yellow curves of Figure 6 show the error as a function of t for 5 random permutations; we see that as the step size decreases, the curves tend to have the same slope as a quadratic curve (shown as a dashed-dotted black line for reference). The blue curve corresponds to the average operator of 5000 random permutations; this has convergence of order $\mathcal{O}(t^3)$, as predicted by Theorem 3.2. In fact, the slope is remarkably similar to the back-and-forth scheme (green curve) whose local error is also known to be order-three (Castella et al., 2009).

C.5. Proof of Theorem 3.3

We are considering numerical integrators whose corresponding operators are

$$\mathcal{U}_i = \exp(\eta K \mathcal{L}_i) + \mathcal{O}(\eta^{p+1}) \tag{68}$$

Applying the randomized ordering scheme to the considered setting induces then an operator (Theorem 3.2)

$$\mathcal{U} = \exp(\eta K \mathcal{L}) + \mathcal{O}(K\eta^{\min(p+1,3)}). \tag{69}$$

To apply Theorem B.5 we first make a change of variable $h = K\eta$. Then we expand the operator as

$$\mathcal{U} = \mathcal{I} + h\mathcal{L} + \frac{h^2}{2}\mathcal{L}^2 + O(K^{-\min(p,2)} h^{\min(p+1,3)}). \tag{70}$$

Then, the application of Theorem B.5 provides the ergodic order of convergence is desired result, as the ergodic error is

$$O(K^{-\min(p,2)}h^{\min(p,2)}) = O(\eta^{\min(p,2)}).$$

D. Additional Derivations for Section 4

We here provide the proofs to the claims stated in the main paper concerning the convergence rates of HMC schemes with and without mini-batches. We prove in Appendix D.1 that the convergence rate of HMC is $\mathcal{O}(\eta^p)$. In Appendix D.2 we prove that the convergence rate of HMC with mini-batches $\mathcal{O}(\eta^{\min(p,2)})$.

D.1. Proof of HMC convergence rate

We shall rewrite for simplicity the HMC scheme

$$\begin{aligned}\mathbf{z}^* &= \left[\mathbf{r}^{*\top}, \boldsymbol{\theta}^{*\top}\right]^\top = \underbrace{\psi(\dots\psi(\psi(}_{N_l \text{times}}\mathbf{z}_0))),\\ \left[\mathbf{r}_1^\top, \boldsymbol{\theta}_1^\top\right] &= \left[\exp(-CN_l\eta)\mathbf{r}^{*\top} + \sqrt{1-\exp(-2CN_l\eta)}\mathbf{w}^\top, \boldsymbol{\theta}^{*\top}\right].\end{aligned} \tag{71}$$

where we selected C such that $\exp(-CN_l\eta) = \alpha$. By doing so we can formally levarage the results concerning SDE integration schemes. The pure Hamiltonian steps transform the functions ϕ with operator

$$\mathcal{P} = \exp(\eta N_l \mathcal{H}) + \mathcal{O}(N_l\eta^{p+1}). \tag{72}$$

The dynamics of an SDE with generator term $\mathcal{D}$ are simulated exactly, and consequently the overall operator that transforms the ϕ functions can then be expressed as

$$\begin{aligned}&\mathcal{U} = \mathcal{P}\exp(\eta N_l\mathcal{D}) = (\exp(\eta N_l\mathcal{H}) + \mathcal{O}(N_l\eta^{p+1}))\exp(\eta N_l\mathcal{D}) =\\ &\exp(\eta N_l\mathcal{H})\exp(\eta N_l\mathcal{D}) + \mathcal{O}(N_l\eta^{p+1}).\end{aligned} \tag{73}$$

To apply directly Theorem B.5, see also Abdulle et al. (2014; 2015), it is convenient to rewrite (73) as a unique step with step-size $h = N_l\eta$, i.e.

$$\mathcal{U} = \exp(h\mathcal{H})\exp(h\mathcal{D}) + \mathcal{O}(N_l^{-p}h^{p+1}) \tag{74}$$

We expand in Taylor series (look at Abdulle et al. (2015) for a similar exposition) the terms of the product

$$\begin{aligned}&\exp(h\mathcal{H})\exp(h\mathcal{D}) = \sum_{n=0}^{\infty} h^n\frac{(\mathcal{H})^n}{n!}\sum_{m=0}^{\infty}h^m\frac{(\mathcal{D})^m}{m!} =\\ &\sum_{k=0}^{\infty}h^k\left(\sum_{n+m=k}\frac{(\mathcal{H})^n}{n!}\frac{(\mathcal{D})^m}{m!}\right).\end{aligned}$$

Simple operator algebra shows that

$$(\exp(h\mathcal{H})\exp(h\mathcal{D}))^\dagger = \sum_{k=0}^{\infty}h^k\left(\sum_{n+m=k}\frac{(\mathcal{D}^\dagger)^m}{m!}\frac{(\mathcal{H}^\dagger)^n}{n!}\right). \tag{75}$$

Then

$$\mathcal{U} = \sum_{k=0}^{\infty}h^k\left(\sum_{n+m=k}\frac{(\mathcal{D}^\dagger)^m}{m!}\frac{(\mathcal{H}^\dagger)^n}{n!}\right) + \mathcal{O}(N_l^{-p}h^{p+1}). \tag{76}$$

We use the fact that $\mathcal{H}^\dagger\rho_{ss} = \mathcal{D}^\dagger\rho_{ss} = 0$ and with the help of Theorem B.5 prove that the ergodic error is $\mathcal{O}(N_l^{-p}h^p)$. By simple substitution this is the desired $\mathcal{O}(\eta^p)$.

D.2. Proof of HMC convergence rate with mini-batches

We then proceed to prove the generic ergodic error when considering HMC with mini-batches. As stated in the main text, we assume that $N_l = TK$ with T a positive integer. We define the HMC scheme with mini-batches as follows.

Proposition D.1. *Consider a class of deterministic numerical integrators with order p. Split the datasets into K batches and consider ψ_i that solves $\mathcal{H}_i = -K\left(\nabla_{\boldsymbol{\theta}}^\top U_i(\boldsymbol{\theta})\right)\nabla_{\mathbf{r}} + \left(\left(\mathbf{M}^{-1}\mathbf{r}\right)^\top\right)\nabla_{\boldsymbol{\theta}}$. Sample independently $T = \frac{N_l}{K}$ random permutations $\{\boldsymbol{\pi}^j\}_{j=1}^{N_l}, \boldsymbol{\pi}^j \in \mathbb{P}$ and apply the scheme*

$$\mathbf{z}^* = \psi_{\pi_K^T} \circ \dots \psi_{\pi_1^T} \dots \psi_{\pi_K^2} \circ \dots \psi_{\pi_1^2} \circ \psi_{\pi_K^1} \circ \dots \psi_{\pi_1^1} \mathbf{z}_0, \quad \left[\boldsymbol{\theta}_1^\top, \mathbf{r}_1^\top\right] = \left[\boldsymbol{\theta}^{*\top}, \mathbf{w}^\top,\right] \tag{77}$$

We proceed then to prove the desired result. We have

$$\mathbb{E}(\phi(\mathbf{z}^*)) = \prod_{i=\pi_K^1,\dots,\pi_1^1} \left(\exp(\eta K \mathcal{H}_i) + \mathcal{O}(\eta^{p+1})\right) \cdots \prod_{i=\pi_K^T,\dots,\pi_1^T} \left(\exp(\eta K \mathcal{H}_i) + \mathcal{O}(\eta^{p+1})\right) \phi(\mathbf{z}_0) \tag{78}$$

where the expectation are not yet taken w.r.t the random permutations. As the random permutations are independent, taking the expected value w.r.t them,

$$\mathbb{E}[\phi(\mathbf{z}^*)] = \mathbb{E}_{\boldsymbol{\pi}^1}[\prod_{i=\pi_K^1,\dots,\pi_1^1} \left(\exp(\eta K \mathcal{H}_i) + \mathcal{O}(\eta^{p+1})\right)] \dots \mathbb{E}_{\boldsymbol{\pi}^T}[\prod_{i=\pi_K^T,\dots,\pi_1^T} \left(\exp(\eta K \mathcal{H}_i) + \mathcal{O}(\eta^{p+1})\right)]\phi(\mathbf{z}_0) = \tag{79}$$

$$\left(\exp(\eta K \mathcal{H}) + \mathcal{O}(K\eta^{\min(p+1,3)})\right) \dots \left(\exp(\eta K \mathcal{H}) + \mathcal{O}(K\eta^{\min(p+1,3)})\right) \phi(\mathbf{z}_0) = \tag{80}$$

$$\exp(\eta T K \mathcal{H}) + \mathcal{O}(TK\eta^{\min(p+1,3)})\phi(\mathbf{z}_0). \tag{81}$$

The $\mathcal{D}$ step can be solved analytically, then the overall operator that transform functions is

$$\mathcal{S} = \exp(\eta T K \mathcal{D}) \left(\exp(\eta T K \mathcal{H}) + \mathcal{O}(TK\eta^{\min(p+1,3)})\right) = \exp(\eta T K \mathcal{D}) \exp(\eta T K \mathcal{H}) + \mathcal{O}(TK\eta^{\min(p+1,3)})$$

Similarly to the full batch case, we define $h = TK\eta$, then the operator is rewritten as

$$\mathcal{S} = \exp(h\mathcal{D}) \exp(h\mathcal{H}) + \mathcal{O}((TK)^{-\min(p,2)} h^{\min(p+1,3)})$$

The ergodic error is then of $\mathcal{O}(\eta^{\min(p,2)})$.

E. Toy Example

We remark that Theorem 4 in Chen et al. (2015) is not aligned with the results we present in this work. Indeed, the work in Chen et al. (2015) suggests that the only effects on the asymptotic invariant measure of an SG-MCMC algorithm are due to the order of the numerical integrators, independently of mini-batching. Instead, our results in Theorem 3.3, show that an explicit bottleneck of order two is present due to mini-batches, i.e. due to the stochastic nature of gradient.

In Appendix E.1, we present a simple toy example with known analytic solution, which contradicts the claims of Chen et al. (2015), while our geometrical interpretation is compatible nevertheless. We consider an analytical integrator that corresponds to an arbitrarily high order numerical integration. We show that when considering mini-batches the stationary distribution is not the desired one, hinting at the presence of a bottleneck in accordance with our theory.

E.1. Model details

In this section we introduce a toy example – whose dynamics has closed form solutions – to understand the role of mini-batching. For simplicity we consider a one dimensional case.

The prior distribution for the parameter θ is

$$p(\theta) = \mathcal{N}(\theta; 0, \sigma_\theta^2), \tag{82}$$

while the likelihood for any given observation is

$$p(x|\theta) = \mathcal{N}(x; \theta, \sigma_x^2), \tag{83}$$

where σ_θ, σ_x are arbitrary positive constants.

We consider the simplest case in which the dataset is composed by two datapoints, i.e. $N = 2$, $\mathrm{D} = \{x_1, x_2\}$. Simple calculations show that the posterior distribution is of the form

$$p(\theta|\mathrm{D}) = \mathcal{N}\left(\theta; \frac{x_1 + x_2}{v}, \sigma_l^2\right), \tag{84}$$

where $v = \frac{\sigma_x^2}{\sigma_\theta^2} + 2$ and $\sigma_l^2 = \left(\frac{1}{\sigma_\theta^2} + \frac{2}{\sigma_x^2}\right)^{-1}$.

The potential $U(\theta)$ corresponding to the distribution Equation (84) has expression

$$U(\theta) = \frac{1}{2\sigma_l^2}\left(\theta - \left(\frac{x_1 + x_2}{v}\right)\right)^2 + \text{const.} \tag{85}$$

Notice that, up to constants of θ, the potential can be rewritten as

$$U(\theta) = \frac{1}{4\sigma_l^2}\left(\theta - \frac{2x_1}{v}\right)^2 + \frac{1}{4\sigma_l^2}\left(\theta - \frac{2x_2}{v}\right)^2. \tag{86}$$

Consequently the gradient has form

$$\nabla U(\theta) = \frac{1}{\sigma_l^2}\left(\theta - \frac{x_1 + x_2}{v}\right) = \frac{1}{2\sigma_l^2}\left(\theta - \frac{2x_1}{v}\right) + \frac{1}{2\sigma_l^2}\left(\theta - \frac{2x_2}{v}\right). \tag{87}$$

The potential and gradient with a sampled mini-batch corresponding to Equation (86) and Equation (87) are respectively

$$U_i(\theta) = \frac{1}{2\sigma_l^2}\left(\theta - \frac{2x_i}{v}\right)^2, \tag{88}$$

and

$$\nabla U_i(\theta) = \frac{1}{\sigma_l^2}\left(\theta - \frac{2x_i}{v}\right). \tag{89}$$

As shown in Appendix E.2, it is possible to build analytical integrators, that provide perfect simulation of paths. These integrators, that correspond to arbitrarily high order numerical integrators, are used for a numerical simulation with $\eta = 0.4, C = 2, \sigma_x^2 = 2, \sigma_\theta^2 = 0.5$ and two sampled datapoints $x_1 = 4, x_2 = -3.2$. The resulting posterior distributions are summarized in the histograms (100K samples) of Figure 7. Even considering a perfect integrator, when considering mini-batches, the stationary distribution is not the desired one. Importantly, this is a different result from the one suggested in Chen et al. (2015) where it is claimed that asymptotically in time only the order of the numerical integrator determines the convergence rate.

We further support our numerical evidence with a through theoretical exploration in Appendix E.2.

E.2. Derivation of the analytical integrators

We consider the dynamics of Equation (3), where for simplicity $\mathbf{M} = 1$. We then have, remembering $\mathbf{z}(t) = \begin{bmatrix} r(t) \\ \theta(t) \end{bmatrix}$, the following SDE

$$d\mathbf{z}(t) = \begin{bmatrix} -C & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} r(t) \\ \sigma_l^{-2}(\theta(t) - \frac{x_1+x_2}{v}) \end{bmatrix} dt + \begin{bmatrix} \sqrt{2C}dw(t) \\ 0 \end{bmatrix}, \tag{90}$$

that we rewrite in the more convenient form

$$d\mathbf{z}(t) = \begin{bmatrix} -C & -\sigma_l^{-2} \\ 1 & 0 \end{bmatrix} \left(\mathbf{z}(t) - \begin{bmatrix} 0 \\ \bar{x} \end{bmatrix}\right) dt + \begin{bmatrix} \sqrt{2C}dw(t) \\ 0 \end{bmatrix}, \tag{91}$$

where $\bar{x} = \frac{x_1+x_2}{v}$.

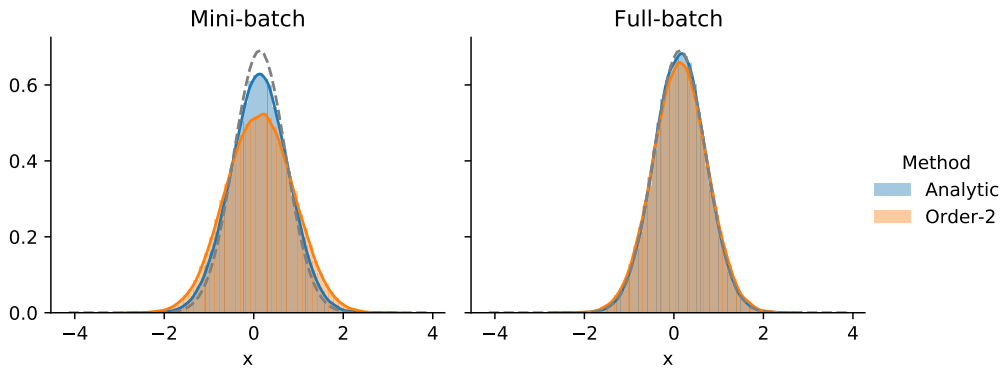

Figure 7. Histograms of stationary distributions (100K samples). The grey dotted line denotes the true posterior density. Irrespectively of order, mini-batching prevents from convergence to true posterior.

Starting from initial conditions $\mathbf{z}(0) = \mathbf{z}_0$, we can shown that the system has analytic solution

$$\mathbf{z}(t) = \exp(t\mathbf{A})\left(\mathbf{z}_0 - \begin{bmatrix}0\\ \bar{x}\end{bmatrix}\right) + \begin{bmatrix}0\\ \bar{x}\end{bmatrix} + \int_{s=0}^{t} \exp((t-s)\mathbf{A}) \begin{bmatrix}\sqrt{2C}dw(s)\\ 0\end{bmatrix} \tag{92}$$

where for simplicity we have defined $\mathbf{A} = \begin{bmatrix}-C & -\sigma_l^{-2}\\ 1 & 0\end{bmatrix}$. Considering that $\sigma_l^{-2}, C > 0$ simple calculations show that the eigenvalues of $\mathbf{A}$ have negative real part, ensuring that as $t \to \infty$ we have $\exp(t\mathbf{A}) \to \mathbf{0}$.

A probabilistically equivalent representation of eq. Equation (92) is the following

$$\mathbf{z}(t) = \exp(t\mathbf{A})\left(\mathbf{z}_0 - \begin{bmatrix}0\\ \bar{x}\end{bmatrix}\right) + \begin{bmatrix}0\\ \bar{x}\end{bmatrix} + \mathbf{n}, \tag{93}$$

where

$$\mathbf{n} \sim \mathcal{N}\left(\mathbf{0}, \int_{s=0}^{t} \exp((t-s)\mathbf{A}) \begin{bmatrix}2C & 0\\ 0 & 0\end{bmatrix} \exp\left((t-s)\mathbf{A}^\top\right) ds\right). \tag{94}$$

Equation Equation (93) guarantees that it is possible to build a numerical integrator for any chosen step size η that simulates exactly the dynamics, corresponding to an arbitrarily high order of weak integration.

Convergence to the posterior

We then explore the transition probability starting from a given initial state $\mathbf{z}_0$ induced by such a numerical integrator with step size η

$$p(\mathbf{z}(\eta)|\mathbf{z}(0) = \mathbf{z}_0) =$$
$$\mathcal{N}\left(\mathbf{z}(\eta); \exp(\eta\mathbf{A})\left(\mathbf{z}_0 - \begin{bmatrix}0\\ \bar{x}\end{bmatrix}\right) + \begin{bmatrix}0\\ \bar{x}\end{bmatrix}, \int_{s=0}^{\eta} \exp((\eta-s)\mathbf{A}) \begin{bmatrix}2C & 0\\ 0 & 0\end{bmatrix} \exp\left((\eta-s)\mathbf{A}^\top\right) ds\right). \tag{95}$$

As shown in Appendix E.2.1, the covariance matrix of the distribution Equation (95) can be expressed as

$$\int_{s=0}^{\eta} \exp((\eta-s)\mathbf{A}) \begin{bmatrix}2C & 0\\ 0 & 0\end{bmatrix} \exp\left((\eta-s)\mathbf{A}^\top\right) ds = \begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix} - \exp(\eta\mathbf{A}) \begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix} \exp\left(\eta\mathbf{A}^\top\right), \tag{96}$$

and consequently, we can rewrite Equation (95) as

$$p(\mathbf{z}(\eta)|\mathbf{z}(0)=\mathbf{z}_0) = \\ \mathcal{N}\left(\mathbf{z}(\eta); \exp\left(\eta\mathbf{A}\right)\left(\mathbf{z}_0 - \begin{bmatrix}0\\ \bar{x}\end{bmatrix}\right) + \begin{bmatrix}0\\ \bar{x}\end{bmatrix}, \mathbf{\Sigma}(\eta)\right), \tag{97}$$

where we introduced the $\mathbf{\Sigma}(\eta) = \begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix} - \exp\left(\eta\mathbf{A}\right)\begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix}\exp\left(\eta\mathbf{A}^\top\right)$.

Importantly, we have that $\rho_{ss}(\mathbf{z})$ is the stationary distribution of the stochastic process, if and only if the following holds:

$$\rho_{ss}(\mathbf{z}^*) = \int p(\mathbf{z}(\eta)=\mathbf{z}^*|\mathbf{z}(0)=\mathbf{z})\rho_{ss}(\mathbf{z})d\mathbf{z} \tag{98}$$

where equality must hold for all η. To check that eq. Equation (98) holds with:

$$\rho_{ss}(\mathbf{z}) = \mathcal{N}(r;0,1)\,\mathcal{N}\left(\theta; \frac{x_1+x_2}{v}, \sigma_l^2\right) = \mathcal{N}\left(\mathbf{z}; \begin{bmatrix}0\\ \bar{x}\end{bmatrix}, \begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix}\right) \tag{99}$$

we substitute directly:

$$\begin{aligned}
&\int p(\mathbf{z}(\eta)=\mathbf{z}^*|\mathbf{z}(0)=\mathbf{z})\rho_{ss}(\mathbf{z})d\mathbf{z} = \\
&\int \mathcal{N}\left(\mathbf{z}^*; \exp\left(\eta\mathbf{A}\right)\left(\mathbf{z} - \begin{bmatrix}0\\ \bar{x}\end{bmatrix}\right) + \begin{bmatrix}0\\ \bar{x}\end{bmatrix}, \mathbf{\Sigma}(\eta)\right)\mathcal{N}\left(\mathbf{z}; \begin{bmatrix}0\\ \bar{x}\end{bmatrix}, \begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix}\right)d\mathbf{z} \\
&= \mathcal{N}\left(\mathbf{z}^*; \exp\left(\eta\mathbf{A}\right)\left(\begin{bmatrix}0\\ \bar{x}\end{bmatrix} - \begin{bmatrix}0\\ \bar{x}\end{bmatrix}\right) + \begin{bmatrix}0\\ \bar{x}\end{bmatrix}, \mathbf{\Sigma}(\eta) + \exp\left(\eta\mathbf{A}\right)\begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix}\exp\left(\eta\mathbf{A}^\top\right)\right) \\
&= \mathcal{N}\left(\mathbf{z}^*; \begin{bmatrix}0\\ \bar{x}\end{bmatrix}, \begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix}\right)
\end{aligned}$$

proving that the stationary distribution is indeed the desired one.

FAILURE OF CONVERGENCE IN THE CASE OF MINI-BATCHES

In the following, we show that when considering mini-batches instead, we fail to converge to the true posterior. Importantly this is true even with the analytical solution for the steps of the integrator, stressing that numerical integration and mini-batching have two independent effects. With mini-batches, at every step of the integration a x_1, x_2 are sampled with probability $\frac{1}{2}$.

Instead of simulating eq. Equation (91) at every step of the numerical integration, the following stochastic process is then considered:

$$d\mathbf{z}(t) = \begin{bmatrix}-C & -\sigma_l^{-2}\\ 1 & 0\end{bmatrix}\left(\mathbf{z}(t) - \begin{bmatrix}0\\ \bar{x}_i\end{bmatrix}\right)dt + \begin{bmatrix}\sqrt{2C}dw(t)\\ 0\end{bmatrix} \tag{100}$$

where, $\bar{x}_i = \frac{2x_i}{v}$, and $i=1$ or $i=2$ is sampled randomly.

Similarly to the previous case it is possible to construct a numerical integrator that solves exactly the dynamics, this time with the potential computed using randomly only one of the two datapoints. The transition probability induced by such an integrator will then be equal to:

$$\begin{aligned}
p(\mathbf{z}(\eta)=\mathbf{z}^*|\mathbf{z}(0)=\mathbf{z}_0) &= \frac{1}{2}\mathcal{N}\left(\mathbf{z}^*; \exp\left(\eta\mathbf{A}\right)\left(\mathbf{z}_0 - \begin{bmatrix}0\\ \bar{x}_1\end{bmatrix}\right) + \begin{bmatrix}0\\ \bar{x}_1\end{bmatrix}, \mathbf{\Sigma}(\eta)\right) \\
&\quad + \frac{1}{2}\mathcal{N}\left(\mathbf{z}^*; \exp\left(\eta\mathbf{A}\right)\left(\mathbf{z}_0 - \begin{bmatrix}0\\ \bar{x}_2\end{bmatrix}\right) + \begin{bmatrix}0\\ \bar{x}_2\end{bmatrix}, \mathbf{\Sigma}(\eta)\right) \qquad (101) \\
&= \frac{1}{2}f_1(\mathbf{z}^*, \eta, \mathbf{z}_0) + \frac{1}{2}f_2(\mathbf{z}^*, \eta, \mathbf{z}_0), \qquad (102)
\end{aligned}$$

where f_1, f_2 are the two Gaussian transition probabilities. Importantly, this allows to state the following central statement: the distribution $p(\theta|\mathcal{D})$ is not the stationary distribution of the process. This claim can be proven by contradiction. Suppose

that the stationary distribution is the one of interest, $\mathcal{N}\left(\mathbf{z}; \begin{bmatrix}0\\ \bar{x}\end{bmatrix}, \begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix}\right)$. Performing one step of integration the new distribution will be of the form:

$$\rho_{new}(\mathbf{z}^*) = \int \left(\frac{1}{2}f_1(\mathbf{z}^*, \eta, \mathbf{z}) + \frac{1}{2}f_2(\mathbf{z}^*, \eta, \mathbf{z})\right)\rho_{ss}(\mathbf{z})d\mathbf{z} \tag{103}$$

$$= \frac{1}{2}\mathcal{N}\left(\mathbf{z}^*; \exp(\eta\mathbf{A})\left(\begin{bmatrix}0\\ \bar{x}\end{bmatrix} - \begin{bmatrix}0\\ \bar{x}_1\end{bmatrix}\right) + \begin{bmatrix}0\\ \bar{x}\end{bmatrix}, \begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix}\right)$$

$$+ \frac{1}{2}\mathcal{N}\left(\mathbf{z}^*; \exp(\eta\mathbf{A})\left(\begin{bmatrix}0\\ \bar{x}\end{bmatrix} - \begin{bmatrix}0\\ \bar{x}_2\end{bmatrix}\right) + \begin{bmatrix}0\\ \bar{x}\end{bmatrix}, \begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix}\right) \tag{104}$$

$$\neq \rho_{ss}(\mathbf{z}^*). \tag{105}$$

By definition, we should have the functional equivalence between ρ_{new} and ρ_{ss}. Equation Equation (103) implies however that after the mixture transition probability the new density ρ_{new} will be a mixture of Gaussian distributions, and consequently $\rho_{new} \neq \rho_{ss}$. This concludes the demonstration that the stationary distribution is not the one of interest.

E.2.1. Derivation of eq. Equation (96)

Define:

$$\mathbf{K}(t) = \int_{s=0}^{t} \exp\left((t-s)\mathbf{A}\right)\begin{bmatrix}2C & 0\\ 0 & 0\end{bmatrix}\exp\left((t-s)\mathbf{A}^\top\right)ds, \tag{106}$$

and

$$\mathbf{S}(t) = \begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix} - \exp(t\mathbf{A})\begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix}\exp\left(t\mathbf{A}^\top\right). \tag{107}$$

Immediately we see that $\mathbf{K}(0) = \mathbf{S}(0) = \mathbf{0}$.

The derivative of $\mathbf{K}(t)$ has expression:

$$\frac{d\mathbf{K}(t)}{dt} = \exp\left((t-t)\mathbf{A}\right)\begin{bmatrix}2C & 0\\ 0 & 0\end{bmatrix}\exp\left((t-t)\mathbf{A}^\top\right)$$

$$+ \mathbf{A}\int_{s=0}^{t} \exp\left((t-s)\mathbf{A}\right)\begin{bmatrix}2C & 0\\ 0 & 0\end{bmatrix}\exp\left((t-s)\mathbf{A}^\top\right)ds$$

$$+ \int_{s=0}^{t} \exp\left((t-s)\mathbf{A}\right)\begin{bmatrix}2C & 0\\ 0 & 0\end{bmatrix}\exp\left((t-s)\mathbf{A}^\top\right)ds\mathbf{A}^\top$$

$$= \begin{bmatrix}2C & 0\\ 0 & 0\end{bmatrix} + \mathbf{A}\mathbf{K}(t) + \mathbf{K}(t)\mathbf{A}^\top.$$

Similarly we have:

$$\frac{d\mathbf{S}(t)}{dt} = -\mathbf{A}\exp(t\mathbf{A})\begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix}\exp\left(t\mathbf{A}^\top\right) - \exp(t\mathbf{A})\begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix}\exp\left(t\mathbf{A}^\top\right)\mathbf{A}^\top$$

$$= -\mathbf{A}\left(\begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix} - \mathbf{S}(t)\right) - \left(\begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix} - \mathbf{S}(t)\right)\mathbf{A}^\top.$$

Since $\mathbf{A}\begin{bmatrix}1 & 0\\ 0 & \sigma_l^2\end{bmatrix} = \begin{bmatrix}-C & -1\\ 1 & 0\end{bmatrix}$, we can prove

$$\frac{d\mathbf{S}(t)}{dt} = \begin{bmatrix}2C & 0\\ 0 & 0\end{bmatrix} + \mathbf{A}\mathbf{S}(t) + \mathbf{S}(t)\mathbf{A}^\top.$$

Definining the difference matrix $\mathbf{H}(t) = \mathbf{K}(t) - \mathbf{S}(t)$ we recognize that it satisfies the differential equation

$$\frac{d\mathbf{H}(t)}{dt} = \mathbf{A}\mathbf{H}(t) + \mathbf{H}(t)\mathbf{A}^\top$$

that has explicit solution

$$\mathbf{H}(t) = \exp(t\mathbf{A})\,\mathbf{H}(0)\exp\left(t\mathbf{A}^\top\right) \tag{108}$$

Since $\mathbf{H}(0) = \mathbf{0}$ we conclude $\mathbf{H}(t) = \mathbf{0}$ proving the desired equality.

F. Experiments: Additional Details

F.1. Experimental setup

Throughout our experimental campaign, we work on a number of datasets from the UCI repository[3]. For each dataset, we have considered 5 random splits into training and test sets; for the regression datasets, we have adopted the splits in Mukhoti et al. (2018) which are available online[4] under the Creative Commons licence. The primary point of interest is in this work is the predictive distribution over the test set and how this deviates from the true posterior. Table 1 and Table 2 summarize the details of the regression and classification datasets in this work. In the same tables we report some error metrics; the latter are not a central part of the evaluation of this paper, but they are simply indicative of the performance of SHMC. Our implementation is loosely based on the PYSGMCMC framework[5].

Network architecture. Throughout this experimental campaign, we consider BNNs featuring L layers, where each layer is defined as follows:

$$f_l(\mathbf{x}) = \frac{1}{\sqrt{D_{l-1}}}\mathbf{W}_l\,\varphi(f_{l-1}(\mathbf{x})) + \mathbf{b}_l, \quad l \in \{1, ..., L+1\}, \tag{109}$$

where φ is a non-linear activation function. The model parameters are summarized as $\boldsymbol{\theta} \equiv \{\mathbf{W}_l \in \mathbb{R}^{D_l \times D_{l-1}}, \mathbf{W}_l \in \mathbb{R}^{D_l \times D_{l-1}}\}$, and they denote the matrix of weights and the vector of biases for layer l. Note that we divide by the square root of the input dimension D_{l-1}; this scheme is known as the *NTK parameterization* (Jacot et al., 2018; Lee et al., 2020), and it ensures that the asymptotic variance neither explodes nor vanishes. In this context, we place a standard Gaussian prior for both weights and biases, i.e. $p(\boldsymbol{\theta}) = \mathcal{N}(0, I)$.

Regression and classification tasks. For regression tasks we consider BNNs with 4 layers and 50 nodes per layer with RELU activation. Table 1 outlines the predictive performance of SHMC for four regression datasets (BOSTON, CONCRETE, ENERGY and YACHT), as well as the noise variance σ^2 used in each case. We have drawn 200 samples from the posterior and the regression performance is evaluated in terms of *root mean squared error* (RMSE) and *mean negative log-likelihood* MNLL. In order to put these values into perspective, we also show the results for the adaptive SGHMC scheme from Springenberg et al. (2016) featuring a network of identical structure as in our setup. In both cases, the full-batch gradient was used.

Table 1. Regression results for BNNs with 4 layers and 50 nodes per layer with RELU activation. For SGHMC we follow Springenberg et al. (2016).

DATASET	Training Size	Test Size	σ^2	Test RMSE(↓) SHMC	Test RMSE(↓) SGHMC	Test MNLL(↓) SHMC	Test MNLL(↓) SGHMC
Boston	455	51	0.2	2.821±0.61	2.825±0.63	2.600±0.09	2.600±0.09
Concrete	927	103	0.05	4.907±0.39	4.833±0.46	2.971±0.07	2.948±0.09
Energy	691	77	0.01	0.501±0.07	0.489±0.07	1.191±0.07	1.120±0.03
Yacht	277	31	0.005	0.420±0.12	0.436±0.13	1.180±0.02	1.176±0.02

For classification we consider IONOSPHERE and VEHICLE from UCI, for which we sample from RELU BNNs with 2 layers and 50 nodes per layer. We have drawn 200 samples from the posterior, and the predictive accuracy and test MNLL of SHMC can be seen in Table 2, where we also show the performance of the adaptive version of SGHMC (Springenberg et al., 2016).

[3] https://archive.ics.uci.edu/ml/index.php
[4] https://github.com/yaringal/DropoutUncertaintyExps
[5] https://github.com/MFreidank/pysgmcmc

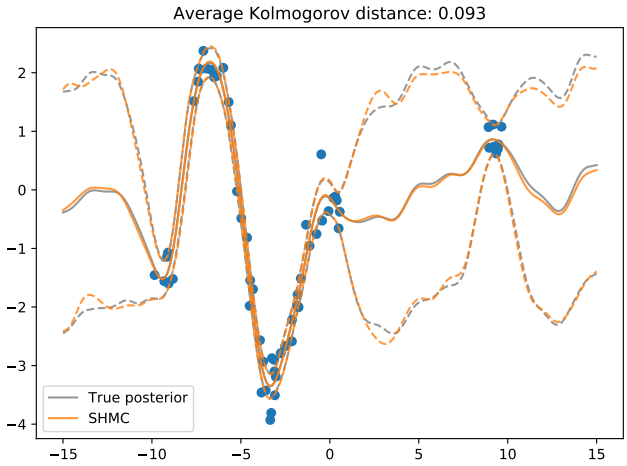

Figure 8. True and SHMC predictive posterior on a synthetic dataset.

The results have been very similar for these two cases; this has been expected, as both methods are supposed to converge to the same posterior.

Table 2. Classification results for BNNs with 2 layers and 50 nodes per layer with RELU activation. For SGHMC we follow Springenberg et al. (2016).

DATASET	Classes	Training Size	Test Size	Test accuracy(↑) SHMC	Test accuracy(↑) SGHMC	Test MNLL(↓) SHMC	Test MNLL(↓) SGHMC
Ionosphere	2	301	34	0.920±0.03	0.916±0.02	0.294±0.04	0.295±0.04
Vehicle	4	646	200	0.777±0.02	0.783±0.02	0.588±0.03	0.588±0.03

Synthetic regression example. We also consider as simple regression model applied on the synthetic one-dimensional dataset of Figure 8. In this case we choose a linear model with fixed basis functions, in order to demonstrate SHMC convergence in a fully controlled environment where the true posterior can be calculated analytically. We consider $D = 256$ trigonometric basis functions: $f(x) = \sqrt{2/D}\,\mathbf{w}^\top \cos(\omega x - \pi/4)$, where $\mathbf{w} \in \mathbb{R}^{D\times 1}$ contains the weights of D features and $\omega \in \mathbb{R}^{D\times 1}$ is a vector of fixed frequencies. In the experiments that follow, we have a Gaussian likelihood with variance 0.1 and prior $p(\mathbf{w}) = \mathcal{N}(0, I_D)$. The predictive posterior can be seen in Figure 8, where the test set consists of 200 uniformly distributed points.

F.2. Comparison framework and convergence to posterior

Comparing predictive distributions. In this work, we explore the behavior of a number of methods by comparing the predictive distribution given by a particular setting to the distribution of an oracle. The comparison is performed in terms of one-dimensional predictive marginals using 200 samples: for each test point we evaluate the Kolmogorov distance between the predictive distribution and the oracle. We then report the average distance over the test set. The Kolmogorov distance takes values between 0 and 1. In Figure 8 we show an one-dimensional regression example where the average distance values over the test set is smaller than 0.1.

Self-distance. When we compare the distance between empirical distribution, this will never be exactly zero. In order to determine whether a given sample is a sufficiently good approximation for a target distribution, we have to compare the corresponding distance value to the *self-distance*. The latter can be evaluated by resampling from the oracle posterior distribution: for each oracle we consider 5 independent runs producing 200 samples each, which are used to estimate the Kolmogorov self-distance. In the results of Figure 9, Figure 10 and Figure 11, the shaded area denotes the 0.05 and 0.95-quantiles of the self-distance distribution.

Methods summary. The main focus of this experimental campaign is to explore how SHMC is affected by changes of step size and batch size, and to observe whether the introduction of mini-batches induces a convergence bottleneck in practice or not. For SHMC, we consider 4 integrators: LEAPFROG which is provably of order 1, SYMMETRIC (Chen et al., 2015) and LIE-TROTTER (Section 4) which are of order 2, and MT3 (Milstein & Tretyakov, 2003) which is a third order integrator. The SYMMETRIC scheme refers to the symmetric splitting integrator proposed by Chen et al. (2015), which is compatible with our SHMC framework. We also compare against SGHMC (Chen et al., 2014).

Computational summary. At this point, we note that each block in Figure 9, Figure 10 and Figure 11 corresponds to 5 random splits or seeds. For most datasets, drawing 200 samples for the smallest step size (i.e. 0.001) required 15–20 minutes. Such a small step size has been necessary in order to demonstrate convergence for small batch sizes. A full exploration of the methods and integrators for a single split (or seed) of a single dataset required approximately one day of computation on a computer cluster featuring Intel® Xeon® CPU @2.00GHz. It has not been possible to do this kind of exploration in a meaningful way for larger datasets, yet we believe that the existing results are sufficient to demonstrate our theoretical claims.

Convergence. In order to reason about convergence in the experiments that follow, we have taken great care to approximate the true posterior with SGHMC, which we treat as oracle. For the oracles we consider full-batch and $\eta = 0.005$ in all cases, except for the linear model where the true posterior has been analytically tractable. The simulation time has been determined by adjusting the thinning parameter so that lag-1 autocorrelation (ACF(1)) was considered as acceptable. In practice we keep one sample every 500 steps (i.e. thinning), and we discard the first 2000 steps. Then for each oracle, the simulation time was doubled one more time, but that resulted in no significant difference in the predictive distribution. The summary of ACF(1) values that correspond to the oracles used in this work can be found in Table 3. As a final remark, we note that for each combination of dataset and model, all random walks cover the same simulation time. The CPU time is thus determined by the step size η: a larger value for η implies that a smaller number of steps is required to simulate a certain system, as the thinning parameter is adjusted accordingly.

Table 3. Lag 1 autocorrelation for the SGHMC algorithms used as oracles.

DATASET	ACF(1)
Boston	0.09
Concrete	0.23
Energy	0.05
Yacht	0.05
Ionosphere	0.17
Vehicle	0.18

F.3. Extended regression and classification results

In this section we present the results of step size and batch size exploration for the entirety of datasets considered. In all cases, we set $C = 5$; we find this to be a reasonable choice, as we see in the exploration of Appendix F.4.

Figure 9 focuses on the comparison between SHMC with different integrators (LEAPFROG, SYMMETRIC, LIE-TROTTER and MT3) and SGHMC. This is an extended version of Figure 4 of the main paper; we include all the datasets considered, as well as a more fine grained exploration of the batch size. We note that the SYMMETRIC and LIE-TROTTER schemes, which are both of second order, respond similarly to the changes of step size and batch size.

Figure 10 is an extended version of Figure 5 of the main paper. It contains a complete account of the exploration of the integration length N_l for the (generalized) LIE-TROTTER integrator.

F.4. Exploration of the friction constant C

The user-specified constant C appears twice in Equation (3): in the friction term and in the stochastic diffusion term of the SDE. Although Theorem 2.1 implies that any choice for $C > 0$ will maintain the desired stationary distribution (i.e. $\rho_{ss}(\boldsymbol{\theta}) \propto \exp(-U(\boldsymbol{\theta}))$), the transient dynamics of the SDE do change as we have seen in the sample paths of Figure 2. It is generally expected that as C approaches 0, the sample paths approach the deterministic behavior of an ODE. On the other

hand, if C is too large, the stochastic process degenerates to the standard Brownian motion. It is expected that either of these extremes would hurt the usability of SDE simulation as a sampling scheme, but the exact effect of C is not apparent.

In this section, we experimentally examine how a certain choice for C affects practical convergence to the desired posterior in conjunction with the step size η as well as the mini-batch size. Figure 11 summarizes an extensive exploration of C for all the models and datasets considered in this work using the LEAPFROG integrator, with C varying from 0.5 to 100. Again, we measure the average Kolmogorov distance from the true predictive posterior for the test points, and we want to see whether the curve for varying η and batch size approaches the band of self-distance.

As a general remark, the desired convergence properties are not too sensitive to the constant C. Although, the optimal value for C does depend on the dynamics induced by a particular dataset and model, it appears that there is a wide range of values that can be considered as acceptable. For most of the cases considered, a value between 5 and 10 seems to produce reasonable results. Nevertheless, we think that some manual exploration would still be required for a new dataset.

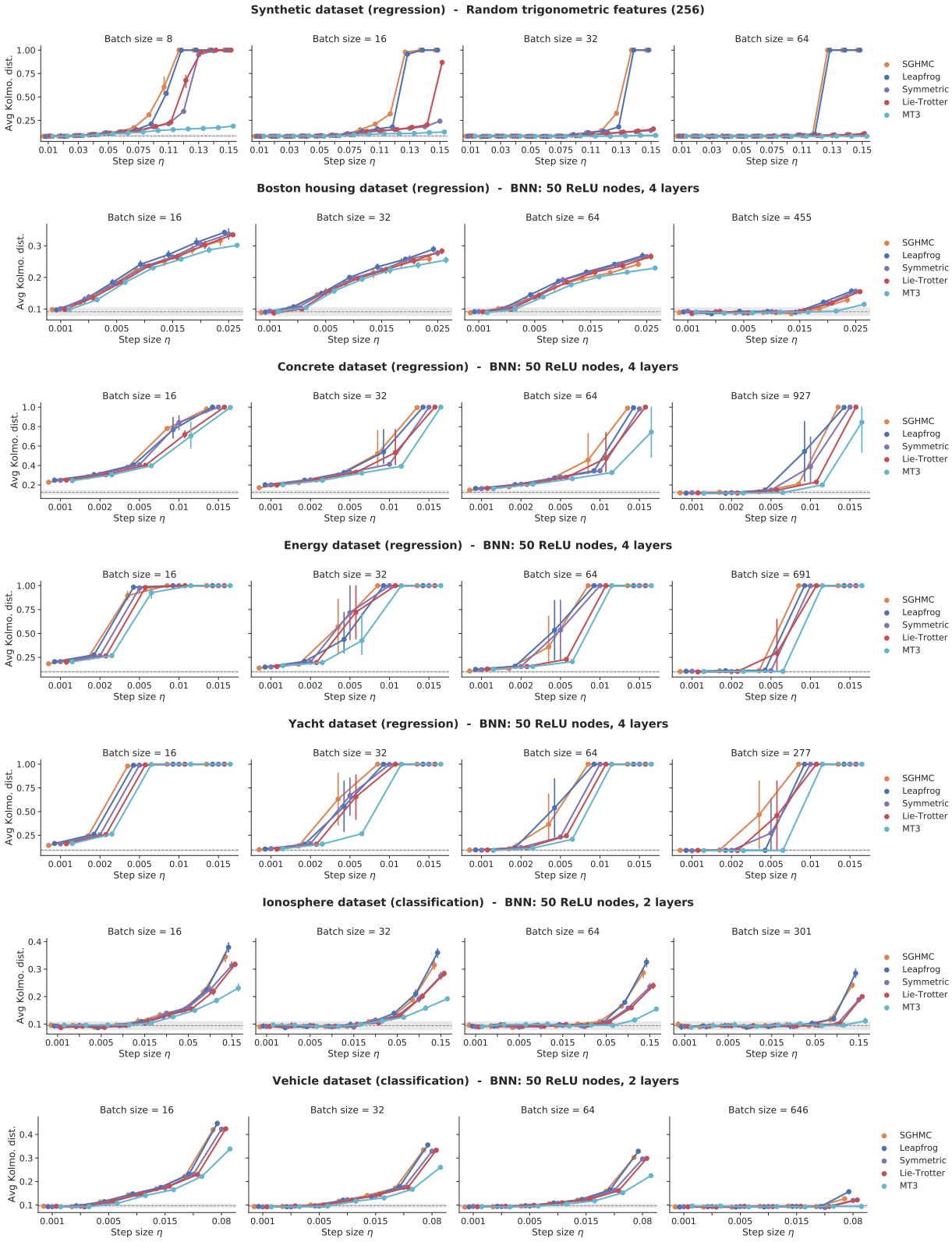

Figure 9. Exploration of step size and batch size for different Hamiltonian-based methods; the grey dotted line denotes the self-distance for the distribution of the oracle.

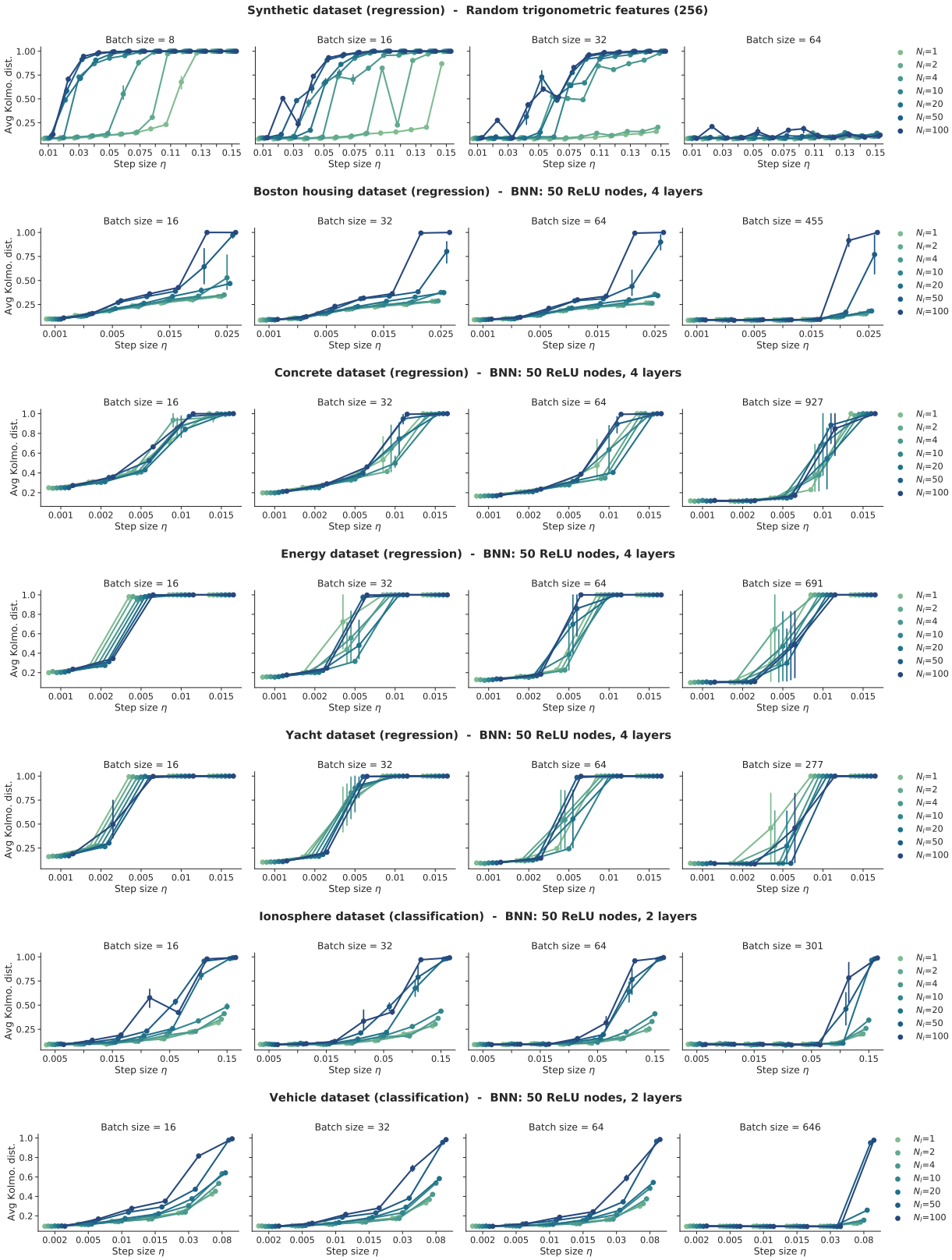

Figure 10. Generalized LIE-TROTTER: Exploration of step size and batch size for different values of the (deterministic) integration length N_l. The grey dotted line denotes the self-distance for the distribution of the oracle.

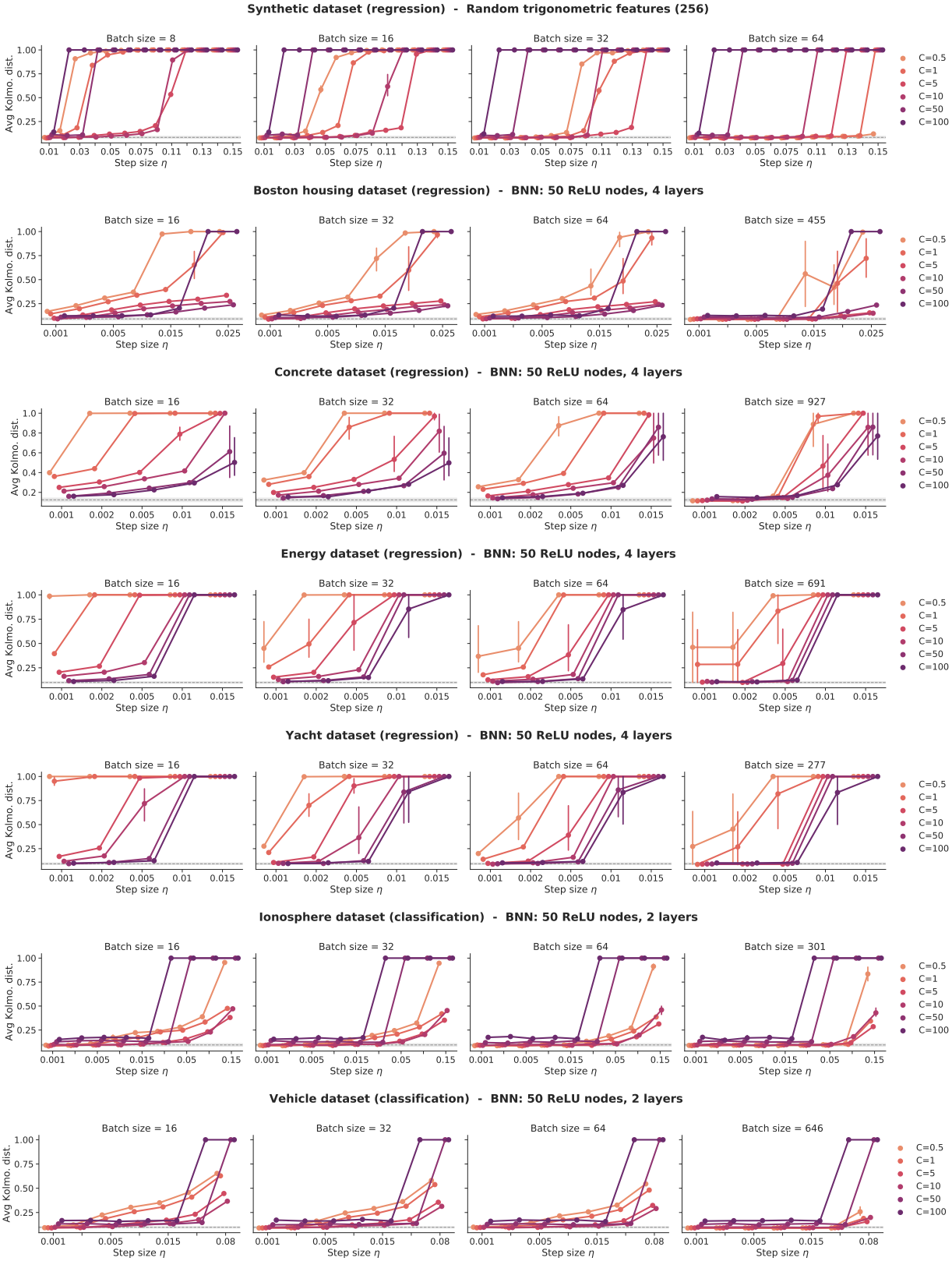

Figure 11. Exploration of step size and batch size for different values of a scalar friction coefficient C. The grey dotted line denotes the self-distance for the distribution of the oracle. LEAPFROG is used in all cases.