Revisiting the Effects of Stochasticity for Hamiltonian Samplers




Giulio Franzese Dimitrios Milios Maurizio Filippone Pietro Michiardi

EURECOM, Sophia Antipolis

39th International Conference on Machine Learning, Baltimore 2022

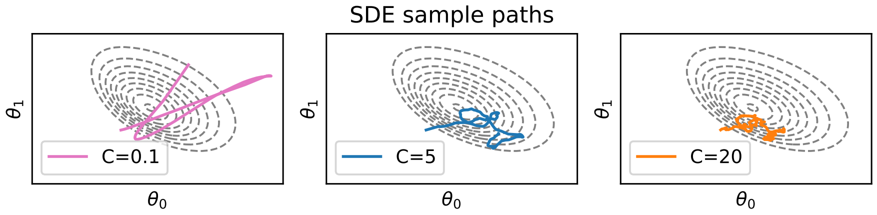
Stochastic Hamiltonian Monte Carlo

- Given **dataset** $\mathrm{D} = \{\mathbf{x}_i\}_{i=1}^{N}$, sample from $p(\boldsymbol{\theta}|\mathrm{D}) \propto \exp(-U(\boldsymbol{\theta}))$

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

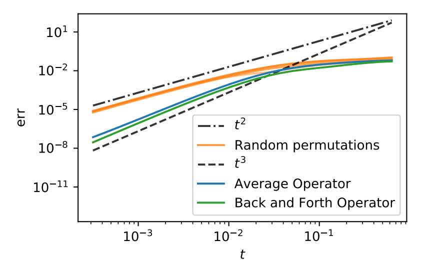
Hamiltonian *Stochastic Differential Equation* with friction term C:

$$d\boldsymbol{\theta}_t = \mathbf{r}_t dt, \quad d\mathbf{r}_t = -\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}_t)dt - C\mathbf{r}_t dt + \sqrt{2C}d\mathbf{w}_t$$



Mini-batches and Bottleneck

- Mini-batches $\mathrm{D}_1, \dots, \mathrm{D}_K$, decomposition $\mathcal{L} = \mathcal{L}_1 + \cdots + \mathcal{L}_K$



- Order-p operators:

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

- Permutations $\boldsymbol{\pi}^1, \dots \boldsymbol{\pi}^{K!}$,

Expected operator:

$$\mathcal{U} = \frac{1}{K!} \sum_{i=1}^{K!} \mathcal{U}_{\pi^i_K} \mathcal{U}_{\pi^i_{K-1}} \dots \mathcal{U}_{\pi^i_1} = \exp\left(\eta K \mathcal{L}\right) + \mathcal{O}(K\eta^3) + \mathcal{O}(K\eta^{p+1})$$

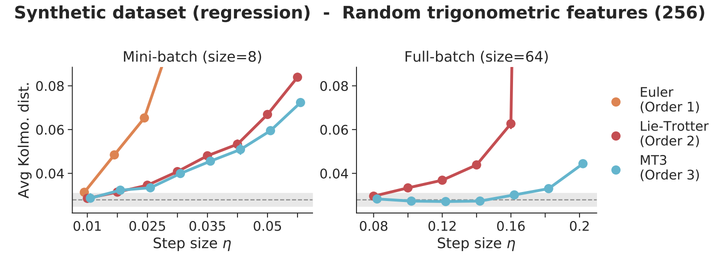
Theorem

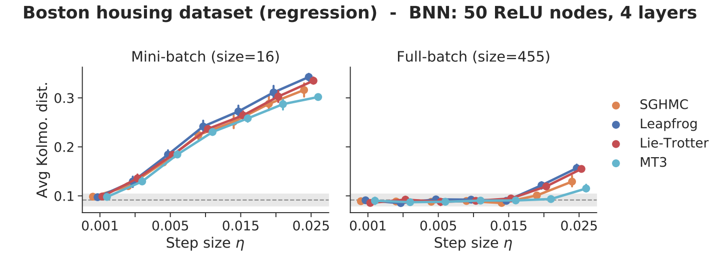
The ergodic error has expansion

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

- Integrators of order $p > 2$ are not beneficial, when mini-batches are used.

Mini-batches and Bottleneck – Empirical Results





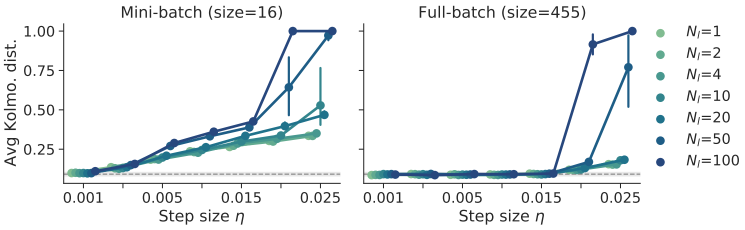
A Connection between HMC and the Lie-Trotter Integrator

HMC (partial momentum refreshment $\alpha > 0$)

$$\theta^1, r^* = \underbrace{\psi(\dots\psi(\psi(}_{N_l \text{ times}}\theta^0, r^0))), \qquad r^1 = \alpha r^* + \sqrt{1-\alpha^2}w$$

- For $\alpha = \exp^{-\eta N_l C}$ and $N_l = 1$: recover the Lie-Trotter SDE integrator
- Convergence rate: mini-batch bottleneck, rate **independent** on N_l.

Boston housing dataset (regression) - BNN: 50 ReLU nodes, 4 layers



Thank you!

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

ral/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

And for any questions, please feel free to contact any of the Authors.