ESAIM: M2AN 58 (2024) 2255–2286
https://doi.org/10.1051/m2an/2024038

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

CONTRACTION RATE ESTIMATES OF STOCHASTIC GRADIENT KINETIC LANGEVIN INTEGRATORS ⋆

Benedict Leimkuhler, Daniel Paulin and Peter A. Whalley**

Abstract. In previous work, we introduced a method for determining convergence rates for integration methods for the kinetic Langevin equation for M-∇Lipschitz m-log-concave densities [Leimkuhler *et al.*, *SIAM J. Numer. Anal.* **62** (2024) 1226–1258]. In this article, we exploit this method to treat several additional schemes including the method of Brunger, Brooks and Karplus (BBK) and stochastic position/velocity Verlet. We introduce a randomized midpoint scheme for kinetic Langevin dynamics, inspired by the recent scheme of Bou-Rabee and Marsden [arXiv:2211.11003, 2022]. We also extend our approach to stochastic gradient variants of these schemes under minimal extra assumptions. We provide convergence rates of $\mathcal{O}(m/M)$, with explicit stepsize restriction, which are of the same order as the stability thresholds for Gaussian targets and are valid for a large interval of the friction parameter. We compare the contraction rate estimates of many kinetic Langevin integrators from molecular dynamics and machine learning. Finally, we present numerical experiments for a Bayesian logistic regression example.

Mathematics Subject Classification. 65C05, 65C30, 65C40.

Received June 14, 2023. Accepted May 16, 2024.

1. Introduction

Efficient sampling of high dimensional probability distributions is required for applications such as Bayesian inference and molecular dynamics (see for example [5,34,41]). A popular approach is to employ a Markov chain constructed by discretizing a stochastic differential equation (SDE) and to approximate observable averages using the central limit theorem. Some common choices of SDEs include overdamped Langevin dynamics [3,55,56] and underdamped/kinetic Langevin dynamics (see below). Other popular methods for MCMC are based on combining Hamiltonian dynamics with Metropolis-Hastings accept/reject condition or combining simple dynamics with stochastic refreshments, which can be simulated exactly [4,10,51].

The focus of this article is on kinetic Langevin dynamics which is the stochastic differential equation system defined by

$$\begin{aligned} \mathrm{d}X_t &= V_t \,\mathrm{d}t, \\ \mathrm{d}V_t &= -\nabla U(X_t)\,\mathrm{d}t - \gamma V_t \,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}W_t, \end{aligned} \tag{1.1}$$

Keywords and phrases. Stochastic gradient, contractive numerical method, Wasserstein convergence, kinetic Langevin dynamics, underdamped Langevin dynamics, MCMC sampling, Brunger–Brooks–Karplus, stochastic Verlet, Bayesian logistic regression, MNIST classification.

⋆ *Dedicated to the memory of Assyr Abdulle.*

School of Mathematics, University of Edinburgh, Edinburgh EH9 2NX, Scotland.
**Corresponding author: p.a.whalley@sms.ed.ac.uk

where $X_t, V_t \in \mathbb{R}^d$, U is a "potential energy" function, $\gamma > 0$ is a friction parameter and W_t is the driving d-dimensional Brownian motion. It can be shown under mild assumptions on the potential U that the invariant measure π of this process has density proportional to $\exp\left(-U(X) - \frac{1}{2}\|V\|^2\right)$ [50]. Particle masses and a temperature parameter are typically included in the context of molecular dynamics, which we neglect here in order to simplify the presentation of results; if desired our analysis could easily be modified to include them. Overdamped Langevin dynamics is the high friction limit of kinetic Langevin dynamics following a time rescaling [50].

In this article, we focus our attention to proving convergence rates of numerical methods for kinetic Langevin sampling in Wasserstein distance (see [63]). We use coupling methods to establish these convergence rates (see [36]), more specifically synchronous coupling as in [19, 23, 35, 44, 47, 58]. Demonstrating that a certain coupling leads to contraction has become a widely used method for demonstrating convergence in terms of Wasserstein distance in both continuous-time scenarios and when discretizing Langevin dynamics and Hamiltonian Monte Carlo [7–9, 24, 30, 44, 53, 59]. The numerical methods we consider are the Euler–Maruyama discretisation (EM), the Brunger–Brooks–Karplus discretisation (BBK) [12], the stochastic position and velocity Verlet (SPV, SVV) [46], popular splitting methods including BAOAB and OBABO [13, 40], a randomized method based on the Hamiltonian integrator of [8] (rOABAO) and the stochastic Euler scheme (SES/EB) [16, 31, 62].

When considering MCMC methods the performance of a sampling scheme is often assessed by measuring the number of steps needed to achieve a certain level of accuracy in the Wasserstein distance metric. By combining the results of this paper with estimates of the stepsize-dependent bias of the numerical methods, it is possible to develop such non-asymptotic bounds in Wasserstein distance which can ultimately provide insight into the computational complexity, convergence rate, and accuracy of the sampling scheme. Bias estimates of some relevant numerical methods have been treated in [35, 44, 47, 58].

The aim of this article is to extend the results of [44] to obtain Wasserstein convergence estimates for a wide interval of the friction parameter whilst maintaining reasonable assumptions on the stepsize. We also propose here a new sampling scheme based on the randomized midpoint method of [6] for Hamiltonian Monte Carlo and we provide convergence results. Moreover, we discuss the use of stochastic gradients and how these proofs can be extended to that setting, which is particularly important in the context of machine learning. We demonstrate our results on an anisotropic Gaussian as well as a Bayesian logistic regression problem involving the MNIST dataset. Although we only treat convergence towards the invariant measure of the scheme in this article, we demonstrate the bias of the methods and discuss these results in combination with the convergence results achieved. We verify the convergence results for the anisotropic Gaussian example, by computing spectral gaps for the numerical methods.

It is also important to note that bias analysis of kinetic Langevin dynamics in the stochastic gradient setting has been considered in Proposition 4 of [60] using the techniques of [1, 66]. Assuming smooth test functions that are compactly supported, they achieve order one bias estimates in the stepsize for the stochastic gradient OABAO scheme. This aligns with what is observed in practice. It remains an open problem to achieve order one in stepsize (for both overdamped Langevin and kinetic Langevin dynamics) in Wasserstein distance, where bias estimates of order 1/2 have been shown in [17, 22, 35].

Using full gradients at each iteration can be computationally expensive in the case of large datasets. The predominant approach used in machine learning optimization is to rely on a stochastic approximation of the gradient (see *e.g.* [54] for one of the first applications of such approaches). In the context of sampling, there has been a great deal of interest to also use such ideas to improve the scalability of MCMC to large datasets, see *e.g.* [67], or the recent review paper [49]. Our contribution here is to generalize the contraction rates for all schemes in Table 1 to appropriate versions of these schemes using stochastic dynamics (see Tab. 3 for a summary of results). We allow for a flexible choice of unbiased gradient estimators (*i.e.* they do not necessarily have to be based on subsampling) and control errors *via* expected variability in the Jacobian of the stochastic gradient *versus* the Hessian of the true potential. It turns out that for all schemes, there is some reduction in convergence rate as the gradient noise increases (we observed this in our numerical experiments when using sub-sampling with very small batch sizes). Nevertheless, for a fixed level of gradient noise, the relative reduction in the contraction rate due to stochastic gradients becomes negligible as the stepsize decreases.

TABLE 1. The table provides our stepsize restrictions and optimal contraction rates of the discretized kinetic Langevin dynamics with stepsize h for an m-convex, M-∇ Lipschitz potential and previous results of [44] and other recent work for further integrators for comparison. We define $\eta = e^{-\gamma h}$.

Algorithm	Stepsize restriction	One-step contraction rate
EM	$\mathcal{O}(1/\gamma)$	$\mathcal{O}(mh/\gamma)$
BBK	$\mathcal{O}(1/\gamma)$	$\mathcal{O}(mh/\gamma)$
SPV	$\mathcal{O}(1/\gamma)$	$\mathcal{O}(mh/\gamma)$
SVV	$\mathcal{O}(1/\gamma)$	$\mathcal{O}(mh/\gamma)$
BAOAB	$\mathcal{O}((1-\eta)/\sqrt{M})$ [44]	$\mathcal{O}(mh^2/(1-\eta))$ [44]
OBABO	$\mathcal{O}((1-\eta)/\sqrt{M})$ [44]	$\mathcal{O}(mh^2/(1-\eta))$ [35,44]
rOABAO	$\mathcal{O}((1-\eta)/\sqrt{M})$	$\mathcal{O}(mh^2/(1-\eta))$
SES/EB	$\mathcal{O}(1/\gamma)$ [44,58]	$\mathcal{O}(mh/\gamma)$ [23,44,58]

2. Assumptions and definitions

2.1. Assumptions on the potential

We place assumptions on the target measure and the resulting potential U. These are strong but allow us to easily obtain quantitative convergence rates. We assume that the potential has a M-Lipschitz gradient and is m-strongly convex, which is equivalent to the following assumptions on the Hessian of U:

Assumption 2.1 (M-∇Lipschitz and m-convex)**.** *For all $x \in \mathbb{R}^d$ there exists m and M such that $0 < m < M < \infty$, and*

$$mI_d \prec \nabla^2 U(x) \prec MI_d.$$

These assumptions are widely used in the analysis of optimisation and sampling methods for gradient descent (see [11,21,25]). Analysis of sampling algorithms in the non-convex setting has also been studied (see [15,30,45]).

2.2. Modified Euclidean norms

We introduce a modified Euclidean norm as in [47] to establish convergence of the discretizations of kinetic Langevin dynamics. It is not possible to establish one-step contraction using the standard Euclidean norm; due to the fact that the generator of kinetic Langevin dynamics is hypoelliptic a specialized metric is needed. We introduce the modified Euclidean norm for $z = (x, v) \in \mathbb{R}^{2d}$

$$\|z\|_{a,b}^2 = \|x\|^2 + 2b\langle x, v\rangle + a\|v\|^2,$$

for $a, b > 0$ which is equivalent to the Euclidean norm when $b^2 < a/4$ with explicit constants given by

$$\frac{1}{2}\|z\|_{a,0}^2 \leq \|z\|_{a,b}^2 \leq \frac{3}{2}\|z\|_{a,0}^2.$$

2.3. Wasserstein distance

We introduce a notion of distance between probability measures to measure convergence. The metric we consider is the p-Wasserstein distance on $\mathcal{P}_p(\mathbb{R}^{2d})$ the space of probability measures with finite p-th moment, for $p \in [0, \infty)$. For probability measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^{2d})$ the p-Wasserstein distance with respect to the modified norm $\|\cdot\|_{a,b}$ (introduced in Sect. 2.2) is defined by

$$\mathcal{W}_{p,a,b}(\nu, \mu) = \left(\inf_{\xi \in \Gamma(\nu,\mu)} \int_{\mathbb{R}^{2d}} \|z_1 - z_2\|_{a,b}^p \, \mathrm{d}\xi(z_1, z_2)\right)^{1/p},$$

where $\Gamma(\mu,\nu)$ is the set of all couplings between μ and ν (the set of joint probability measures with marginals μ and ν).

If one is able to show contraction of any coupling of two paths of a numerical integrator for kinetic Langevin dynamics then one also obtains convergence in p-Wasserstein distance, as it is the infimum over all such couplings (see [18], Cor. 7 or [47], Cor. 20). The coupling technique we consider for quantitative contraction rates is synchronous coupling, where the two paths are generated using identical noise increments. This coupling is also used in [18,23,48].

Proposition 2.2 ([44], Prop. 2.3)**.** *Assume a numerical scheme for kinetic Langevin dynamics with a m-strongly convex M-∇Lipschitz potential U and transition kernel P_h. If any two synchronously coupled chains with initial conditions $(x_0,v_0)\in\mathbb{R}^{2d}$ and $(\tilde{x}_0,\tilde{v}_0)\in\mathbb{R}^{2d}$ of the numerical scheme have the contraction property*

$$\|(x_k-\tilde{x}_k,v_k-\tilde{v}_k)\|_{a,b}^2\leq C(1-c(h))^k\|(x_0-\tilde{x}_0,v_0-\tilde{v}_0)\|_{a,b}^2, \tag{2.1}$$

for a contraction rate $c(h)>0$, $C>0$, $\gamma^2\geq C_\gamma M$, $h\leq C_h(\gamma,\sqrt{M})$ and $a,b>0$ such that $b^2< a/4$, where $C_\gamma>0$ is a constant and $C_h(\gamma,\sqrt{M})>0$ is a constant depending on γ and $\sqrt{M}$ Then we have that for all $\gamma^2\geq C_\gamma M$, $h\leq C_h(\gamma,\sqrt{M})$, $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^{2d})$, and all $k\in\mathbb{N}$,

$$\mathcal{W}_{2,a,b}^2\big(\nu P_h^k,\mu P_h^k\big)\leq C(1-c(h))^k\mathcal{W}_{2,a,b}^2(\nu,\mu).$$

Further to this, P_h has a unique invariant measure which depends on the stepsize, π_h, where $\pi_h\in\mathcal{P}_2(\mathbb{R}^{2d})$.

Proof. The proof is given in Corollary 20 of [47], which relies on Corollary 5.22, Theorem 6.18 of [65]. □

3. Proof strategy

We use the proof strategy introduced in [44] to prove contraction of the numerical schemes considered in this article. We summarize the proof strategy and for further details we refer the reader to [44]. Our method relies on proving contraction of a "twisted Euclidean norm" or modified norm (as stated in Sect. 2.2) which is equivalent to the standard Euclidean norm with an explicit constant. The approach is to find constants $a,b>0$ such that $b^2<a/4$ and the contraction property

$$\|\tilde{z}_{k+1}-z_{k+1}\|_{a,b}^2<(1-c(h))\|\tilde{z}_k-z_k\|_{a,b}^2, \tag{3.1}$$

holds with explicit weak assumptions on the parameters γ and h. Now this is equivalent to showing that

$$\overline{z}_k^T\big((1-c(h))M-P^TMP\big)\overline{z}_k>0,\quad\text{where}\quad M=\begin{pmatrix}1&b\\b&a\end{pmatrix}, \tag{3.2}$$

$(\overline{z}_i)_{i\in\mathbb{N}}:=(\tilde{z}_i-z_i)_{i\in\mathbb{N}}$ and $\overline{z}_{k+1}=P\overline{z}_k$. P is determined by the scheme and implicitly depends on z_k and $\tilde{z}_k$ through a mean value theorem and the Hessian of the potential. If this relation holds for any choice of z_k and $\tilde{z}_k$ then it implies contraction. Therefore proving contraction with a rate $c(h)$ is equivalent to proving that the matrix $\mathcal{H}:=(1-c(h))M-P^TMP\succ 0$ is positive definite. We can use the symmetric structure of the matrix $\mathcal{H}$

$$\mathcal{H}=\begin{pmatrix}A&B\\B&C\end{pmatrix}, \tag{3.3}$$

to show that $\mathcal{H}$ is positive definite by applying the following Proposition 3.1.

Proposition 3.1 ([44])**.** *Let $\mathcal{H}$ be a symmetric matrix of the form (3.3), then $\mathcal{H}$ is positive definite if and only if $A\succ 0$ and $C-BA^{-1}B\succ 0$. Further if A, B and C commute then $\mathcal{H}$ is positive definite if and only if $A\succ 0$ and $AC-B^2\succ 0$.*

Proof. Proof given in [44]. □

4. Numerical integrators

We will consider several popular numerical integrators for kinetic Langevin dynamics, arising in the literatures of molecular dynamics and machine learning. The numerical methods are generally defined by $(x_k, v_k) \in \mathbb{R}^d \times \mathbb{R}^d$ for $k \in \mathbb{N}$ with initial conditions $(x_0, v_0) \in \mathbb{R}^d \times \mathbb{R}^d$ and given noise sequences.

Choices for the algorithm include:

- the Euler–Maruyama discretization (EM);
- splitting methods based on breaking the dynamics into parts which can be solved analytically (in the weak sense) [13, 40, 43];
- the stochastic Euler scheme (SES/EB) (see [16, 31, 62]), which is popular in the machine learning literature (see [19, 23, 58]) and is based on keeping the force constant and integrating exactly over the interval;
- the Brunger–Brooks–Karplus (BBK) scheme which uses a leapfrog-like approach to propagate position and velocity components, combined with implicit and explicit Euler steps in velocity [12, 32];
- the stochastic position and velocity Verlet schemes (SPV, SVV) based on integrating the force and the OU process together in a splitting scheme introduced in [46];
- a new randomized midpoint method based on a Hamiltonian integrator from [8].

We recommend [28, 32] for an introduction to many of these schemes. We next describe these algorithms by giving their respective update rules. We introduce the notation $\Psi_F(x, v)$ to represent the one-step phase-space map corresponding to the operator F with initial condition (x, v) for a step of size $h > 0$, where the operator could be for example the operators $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}, \mathcal{A}, \mathcal{V}$ introduced in the following sections.

4.1. Euler–Maruyama

The Euler–Maruyama discretization with initial conditions $(x_0, v_0) \in \mathbb{R}^{2d}$ and iterations $(x_k, v_k) \in \mathbb{R}^{2d}$ for $k \in \mathbb{N}$ are defined by the update rule

$$\begin{aligned} x_{k+1} &= x_k + hv_k, \\ v_{k+1} &= v_k - h\nabla U(x_k) - h\gamma v_k + \sqrt{2\gamma h}\xi_{k+1}, \end{aligned} \tag{4.1}$$

where $(\xi_k)_{k\in\mathbb{N}}$ are independent $\mathcal{N}(0, I_d)$ draws.

4.2. Splitting methods

Integrators studied in [13, 40] are defined by splitting the dynamics into parts given by $\mathcal{B}$ (integrating the velocity by the force), $\mathcal{A}$ (integrating the position by the velocity) and $\mathcal{O}$ (the solution in the weak sense to the OU process) with update rules given by

$$\begin{aligned} \mathcal{B} &: v \to v - h\nabla U(x), \\ \mathcal{A} &: x \to x + hv, \\ \mathcal{O} &: v \to \eta v + \sqrt{1-\eta^2}\xi, \end{aligned} \tag{4.2}$$

where

$$\eta := \exp(-\gamma h).$$

Then the schemes we will study in this framework are BAOAB and OBABO (the ordering given from left to right based on their application as operators defined in (4.2)). When there is a repeated letter in the ordering that means that each operator is taken with half a step, *i.e.* $h \to h/2$, $\eta \to \eta^{1/2}$. For example BAOAB performs two half steps of $\mathcal{B}$ and $\mathcal{A}$ and one full step of $\mathcal{O}$. For computational efficiency, in practice the same gradient evaluation for the last $\mathcal{B}$ step in a BAOAB iteration is used for the first B step in the next iteration, as the position is not updated.

4.3. Stochastic Euler scheme

The stochastic Euler scheme is based on keeping the force constant and integrating the dynamics exactly over the time interval. For initial conditions $(x_0, v_0) \in \mathbb{R}^{2d}$, the iterations $(x_k, v_k) \in \mathbb{R}^{2d}$ for $k \in \mathbb{N}$ are defined by the update rule

$$\begin{aligned} x_{k+1} &= x_k + \frac{1-\eta}{\gamma} v_k - \frac{\gamma h + \eta - 1}{\gamma^2} \nabla U(x_k) + \zeta_{k+1}, \\ v_{k+1} &= \eta v_k - \frac{1-\eta}{\gamma} \nabla U(x_k) + \omega_{k+1}, \end{aligned} \tag{4.3}$$

where $\eta := \exp(-\gamma h)$ and

$$\zeta_{k+1} = \sqrt{2\gamma} \int_0^h e^{-\gamma(h-s)}\, \mathrm{d}W_{h\gamma+s}, \qquad \omega_{k+1} = \sqrt{2\gamma} \int_0^h \frac{1 - e^{-\gamma(h-s)}}{\gamma}\, \mathrm{d}W_{h\gamma+s}.$$

$(\zeta_k, \omega_k)_{k\in\mathbb{N}}$ are i.i.d Gaussian random vectors with covariances matrix $\Sigma \otimes I_d$ with Σ given by

$$\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2 & \Sigma_3 \end{pmatrix}, \tag{4.4}$$

where

$$\begin{aligned} \Sigma_1 &= \frac{1}{\gamma}\left(2h - \frac{3 - 4\eta + \eta^2}{\gamma}\right), \\ \Sigma_2 &= \frac{1}{\gamma}(1-\eta)^2, \\ \Sigma_3 &= 1 - \eta^2, \end{aligned}$$

as defined in [28].

All of the schemes mentioned so far were studied in [44]. We now introduce some additional schemes, which include the BBK integrator [12] which is popular in molecular dynamics and the stochastic position and velocity Verlet methods [46].

4.4. BBK

For initial conditions $(x_0, v_0) \in \mathbb{R}^{2d}$, the iterations $(x_k, v_k) \in \mathbb{R}^{2d}$ for $k \in \mathbb{N}$ of the BBK method of [12] are defined by the update rule

$$\begin{aligned} x_{k+1} &= x_k + h\left(1 - \frac{\gamma h}{2}\right) v_k - \frac{h^2}{2} \nabla U(x_k) + \sqrt{2\gamma} \frac{h^{3/2}}{2} \xi_k, \\ v_{k+1} &= \frac{1 - \gamma h/2}{1 + \gamma h/2} v_k - \frac{h}{2(1+\gamma h/2)} (\nabla U(x_k) + \nabla U(x_{n+1})) + \frac{\sqrt{2\gamma h}}{2(1+\gamma h/2)} (\xi_{k+1} + \xi_k), \end{aligned}$$

which can be rewritten as [32]

$$\begin{aligned} &(B_1) \qquad v_{k+1/2} = v_k + \frac{h}{2}\left(-\nabla U(x_k) - \gamma v_k + \frac{\sqrt{2\gamma}}{\sqrt{h}} \xi_k\right), \\ &(A) \qquad x_{k+1} = x_k + h v_{k+1/2}, \\ &(B_2) \qquad v_{k+1} = v_{k+1/2} + \frac{h}{2}\left(-\nabla U(x_{k+1}) - \gamma v_{k+1} + \frac{\sqrt{2\gamma}}{\sqrt{h}} \xi_{k+1}\right). \end{aligned}$$

This can be viewed as an explicit Euler step followed by a position update followed by an implicit Euler step. We denote the explicit Euler step by B_1 and the implicit Euler step by B_2.

4.5. Stochastic position and velocity Verlet

The stochastic position and velocity Verlet schemes are defined through an alternative splitting of the dynamics based on keeping the B and O steps together in an exact integration. We define the operators involved in the update rule by

$$\mathcal{V}(h) : v \to \eta v - \frac{1-\eta}{\gamma}\nabla U(x) + \sqrt{1-\eta^2}\xi,$$
$$\mathcal{A}(h) : x \to x + hv,$$

where $\eta = e^{-\gamma h}$. Then the stochastic position Verlet is defined by $\mathcal{A}(h/2)\mathcal{V}(h)\mathcal{A}(h/2)$ and the stochastic velocity Verlet is defined by $\mathcal{V}(h/2)\mathcal{A}(h)\mathcal{V}(h/2)$.

4.6. Randomized midpoint method

Other algorithms for kinetic Langevin dynamics include the randomized midpoint methods considered in [61] and analyzed in [14], which have improved dimension dependence in non-asymptotic estimates. However, they involve multiple gradient evaluations at each step and cannot be analyzed in our framework; this problem has been discussed in [58]. For contractivity of algorithms involving several gradient evaluations we refer the reader to [57].

In the recent paper [8] the authors consider such a method for Hamiltonian Monte Carlo, whose discretization is closely related to the OBABO or OABAO discretization in the $\gamma \to \infty$ limit [35]. More precisely one could consider the following procedure.

Fix a stepsize h then sample $u \sim [0, h]$ and compute

$$\mathcal{A} : x \to x + uv,$$
$$\mathcal{B} : v \to v - h\nabla U(x),$$
$$\mathcal{A} : x \to x + (h-u)v,$$

which is the following update

$$x_{k+1} = x_k + hv_k - h(h-u)\nabla U(x_k + uv_k),$$
$$v_{k+1} = v_k - h\nabla U(x_k + uv_k),$$

then only considering the randomness in the gradient evaluation we arrive at the Verlet scheme considered in [8]. We define $r\mathcal{ABA}$ to be the update

$$x_{k+1} = x_k + hv_k - \frac{h^2}{2}\nabla U(x_k + uv_k),$$
$$v_{k+1} = v_k - h\nabla U(x_k + uv_k),$$

where $u \sim \mathcal{U}(0, h)$ as introduced in [8]. The key difference being that the gradient is evaluated at a random midpoint in the interval of numerical integration. We define the kinetic Langevin dynamics integrator rOABAO to be $\mathcal{O}(r\mathcal{ABA})\mathcal{O}$.

We remark that we can achieve contraction rates by coupling two trajectories which have common noise (in Brownian increment and randomized midpoint) $(\zeta_k, u_k)_{k\in\mathbb{N}}$ with the previously introduced methods. The convergence rates will be established in Section 6.

5. Convergence rates

Theorem 5.1. *For the numerical schemes for kinetic Langevin dynamics given in Table 2 with an m-strongly convex, M-∇Lipschitz potential U we consider any sequence of synchronously coupled random variables with initial conditions $(x_0, v_0) \in \mathbb{R}^{2d}$ and $(\tilde{x}_0, \tilde{v}_0) \in \mathbb{R}^{2d}$.*

TABLE 2. Constants for contraction of each scheme. Implicit refers to the implicit assumption on γ through the stepsize restriction h_0, the value of γ^2 must be greater than a certain constant multiple of M. For example for $h_0 = (1-\eta)/\alpha\sqrt{M}$ is satisfied when $\gamma \geq 2\alpha\sqrt{M}$ and $h < 1/(2\gamma)$ and for $h \geq 1/(2\gamma)$ we have $h_0 \geq 1/(6\alpha\sqrt{M})$.

Algorithm	h_0	γ_0	b	$c(h)$	C	s
EM	$1/2\gamma$	$2\sqrt{M}$	$1/\gamma$	$mh/2\gamma$	1	k
BBK	$1/4\gamma$	$\sqrt{12M}$	$h/2 + 1/\gamma$	$mh/4\gamma$	7	$k-1$
SPV	$1/2\gamma$	$\sqrt{11M}$	$h/(1-\eta)$	$mh/4\gamma$	7	$k-1$
SVV	$1/2\gamma$	$\sqrt{11M}$	$h/(1-\eta)$	$mh/4\gamma$	7	$k-1$
BAOAB	$(1-\eta)/2\sqrt{M}$	implicit	$h/(1-\eta)$	$h^2m/4(1-\eta)$	7	$k-1$
OBABO	$(1-\eta)/4\sqrt{M}$	implicit	$h/(1-\eta)$	$h^2m/4(1-\eta)$	7	$k-1$
rOABAO	$(1-\eta)/2\sqrt{M}$	implicit	$h/(1-\eta)$	$h^2m/4(1-\eta)$	7	$k-1$
SES/EB	$1/2\gamma$	$5\sqrt{M}$	$1/\gamma$	$mh/4\gamma$	1	k

Under stepsize restrictions $h < h_0$ and $\gamma \geq \gamma_0$ for constants given in Table 2 we have the contraction

$$\|(x_k - \tilde{x}_k, v_k - \tilde{v}_k)\|_{a,b} \leq C(1 - c(h))^{s/2}\|(x_0 - \tilde{x}_0, v_0 - \tilde{v}_0)\|_{a,b},$$

with norm given with constants $a = 1/M$ and b, where b, the contraction rate $c(h)$, the preconstant C and the number of steps s are given in Table 2 and are specific to each scheme.

The proofs for EM, BAOAB, OBABO, SES are given in [44]; for all other schemes the proofs are given here.

Referring to Table 2 we have that the convergence rate $c(h)$ is proportional to m/γ for small h, which is shown to match the convergence rate of the continuous dynamics for large γ (see for example [23]). We have that the convergence rate is hm/γ for all the schemes apart from BAOAB, OBABO and rOABAO, which have convergence rates which are faster than the continuous dynamics for large values of γ and h (as originally noted in [44]). This is due to the fact that the $\mathcal{O}$ step is integrated exactly separately and one can take the high friction limit. Since the $\mathcal{O}$ step also leaves the measure invariant the bias in these types of schemes comes from the discretization error of the Hamiltonian integrator and hence retains high order asymptotic bias [43]. However, these splitting schemes are only strong order 1 and this can be seen particularly for large values of friction when the convergence rates are higher than for the continuous dynamics. (It fails to approximate the continuous dynamics, but it is accurate in the sampling context [44]).

6. PROOF OF THEOREM 5.1

In the following contraction rate proofs we follow the structure of [44] for the additional schemes that we are analysing.

Proof for rOABAO. Using the fact that $(r\mathcal{OABAO})^n = \mathcal{O}(r\mathcal{ABAO})^{n-1}r\mathcal{ABAO}$, we can instead proof contraction of $r\mathcal{ABAO}$. This is done to simplify the problem to one gradient evaluation per step. Denoting two synchronous realisations (synchronous in the sense of $(u_k)_{k\in\mathbb{N}}$ and $(\xi_k)_{k\in\mathbb{N}}$) of rABAO as (x_j, v_j) and $(\tilde{x}_j, \tilde{v}_j)$ for $j \in \mathbb{N}$. Further we denote $\overline{x} := (\tilde{x}_j - x_j)$ and $\overline{v} = (\tilde{v}_j - v_j)$. We define $\overline{z}_j = (\overline{x}_j, \overline{v}_j)$ for $j = k, k+1$ for $k \in \mathbb{N}$ by the update rule

$$\overline{x}_{k+1} = \overline{x}_k + h\overline{v}_k - \frac{1}{2}h^2Q(\overline{x}_k + u_k\overline{v}_k), \quad \overline{v}_{k+1} = \eta\overline{v}_k - h\eta Q(\overline{x}_k + u_k\overline{v}_k).$$

Q, by mean value theorem, is defined by $Q = \int_{t=0}^{1} \nabla^2 U(\tilde{x}_k + u_k\tilde{v}_k + t(x_k - \tilde{x}_k + u_k(v_k - \tilde{v}_k)))\,\mathrm{d}t$, then $\nabla U(\tilde{x}_k + u_k\tilde{v}_k) - \nabla U(x_k + u_k v_k) = Q(\overline{x} + u_k\overline{v})$. We use the notation of (3.3), where

$$\begin{aligned}
A &= -c(h)I_d + Q\big(2b\eta h + h^2\big) + Q^2\bigg(-a\eta^2h^2 - b\eta h^3 - \frac{1}{4}h^4\bigg),\\
B &= (b(1-\eta) - h - bc(h))I_d + Q\bigg(a\eta^2 h + \frac{3}{2}b\eta h^2 + b\eta h u + \frac{1}{2}h^2 u + \frac{1}{2}h^3\bigg)\\
&\quad + Q^2\bigg(-a\eta^2h^2u - b\eta h^3 u - \frac{1}{4}h^4 u\bigg),\\
C &= \big(a\big(1-\eta^2\big) - ac(h) - 2b\eta h - h^2\big)I_d + Q\big(2a\eta^2 hu + 3b\eta h^2 u + h^3 u\big)\\
&\quad + Q^2\bigg(-a\eta^2h^2u^2 - b\eta h^3u^2 - \frac{1}{4}h^4u^2\bigg),
\end{aligned}$$

for this scheme and $\eta = \exp\{-\gamma h\}$, further for the ease of notation we have used $u := u_k$. We have chosen b, such that B simplifies to

$$B = -bc(h)I_d + Q\bigg(a\eta^2 h + \frac{3}{2}b\eta h^2 + b\eta h u + \frac{1}{2}h^2 u + \frac{1}{2}h^3\bigg) + Q^2\bigg(-a\eta^2h^2u - b\eta h^3 u - \frac{1}{4}h^4 u\bigg).$$

It is sufficient to prove that $A \succ 0$ and that $C - BA^{-1}B \succ 0$, noting that A, B and C all commute as they are all polynomials in Q. It is therefore sufficient to show that $A \succ 0$ and $AC - B^2 \succ 0$ and A is positive definite by the proof of Theorem 6.1 from [44]. We use $P_{AC-B^2}(\lambda)$ to denote the eigenvalues of $AC - B^2$, where λ are eigenvalues of Q, where we know $\lambda \in [m, M]$ due to the assumptions on U. $AC - B^2 \succ 0$ is shown as follows

$$\begin{aligned}
\frac{P_{AC-B^2}(\lambda)}{h\lambda} &= \frac{\big(\eta^2-1\big)c(h)}{hM\lambda} + h\bigg(\frac{(1+\eta)^2}{M} - \frac{\eta^2\lambda}{M^2} + \frac{c(h)^2}{h^2M\lambda} - \frac{2c(h)\eta^2 u}{hM}\bigg)\\
&\quad + h^2\bigg(\frac{c(h)}{h}\bigg(-\frac{1+2\eta}{M} + \frac{1}{\lambda} + \frac{2\eta}{(1-\eta)\lambda} + \frac{\eta^2\lambda}{M^2}\bigg) + u\bigg(\frac{\eta^2\lambda}{M} + \frac{2\eta^3\lambda}{(1-\eta)M} + \frac{c(h)\eta^2\lambda u}{hM}\bigg)\bigg)\\
&\quad + h^3\bigg(-\frac{\eta^2}{(1-\eta)^2} - \frac{3\eta}{(1-\eta)^2} - \frac{1}{1-\eta} - \frac{c(h)^2}{h^2(1-\eta)^2\lambda} - \frac{\lambda(1-\eta^2)}{4M} - \frac{\eta\lambda}{(1-\eta)M}\bigg)\\
&\quad + h^3u\bigg(\frac{c(h)}{h}\bigg(-1 + \frac{2\eta}{(1-\eta)^2} + \frac{1}{1-\eta} - \frac{3\eta}{1-\eta} - \frac{2\eta^2\lambda}{(1-\eta)M}\bigg) + u\bigg(-\frac{\lambda}{4} - \frac{\eta^2\lambda}{(1-\eta)^2} - \frac{\eta\lambda}{1-\eta}\bigg)\bigg)\\
&\quad + h^4\bigg(\frac{c(h)}{h}\bigg(\frac{3\eta}{(1-\eta)^2} + \frac{1}{1-\eta} + \frac{\lambda}{4M} + \frac{\eta\lambda}{(1-\eta)M}\bigg)\bigg)\\
&\quad + h^4\bigg(u\bigg(\frac{3\eta^2\lambda}{(1-\eta)^2} + \frac{2\eta\lambda}{1-\eta} + \frac{\lambda}{2(1-\eta)} + \frac{c(h)\lambda u}{4h} + \frac{c(h)\eta\lambda u}{h(1-\eta)}\bigg)\bigg)\\
&\quad + h^5\bigg(-\frac{\eta^2\lambda}{4(1-\eta)^2} - \frac{2c(h)\eta\lambda u}{h(1-\eta)^2} - \frac{c(h)\lambda u}{2h(1-\eta)}\bigg)\\
&\geq -\frac{(1+\eta)h}{4M} + h\bigg(\frac{1+3\eta/2}{M}\bigg) + h^5\bigg(-\frac{\eta^2\lambda}{4(1-\eta)^2}\bigg)\\
&\quad + h^3\bigg(\bigg(-\frac{3\eta+5/4}{(1-\eta)^2} - \frac{1-\eta^2}{4}\bigg) + u\frac{c(h)}{h}\bigg(-\frac{2\eta^2}{1-\eta}\bigg) + u^2M\bigg(-\frac{1}{4} - \frac{1}{(1-\eta)^2}\bigg)\bigg)\\
&\geq \frac{3h}{4M} + \frac{5\eta h}{4M} + \bigg(-\frac{3\eta h}{4M} - \frac{5h}{16M} - \frac{h}{16M} - \frac{h}{32M} - \frac{6h}{64M}\bigg) > 0,
\end{aligned}$$

where we have imposed the restriction $h < \frac{1-\eta}{\sqrt{4M}}$ and used $u < h$. Therefore contraction of rABAO holds and all computations can be checked using symbolic computing. We now bound the remaining terms to achieve a contraction result for rOABAO. We bound $\mathcal{O}$ operator on $\|\cdot\|_{a,b}$ by

$$\|\Phi_{\mathrm{O}}(\tilde{x},\tilde{v}) - \Phi_{\mathrm{O}}(x,v)\|_{a,b}^2 \leq 3\|(\overline{x},\overline{v})\|_{a,b}^2,$$

where we have used the norm equivalence in Section 2.2. We bound

$$\begin{aligned}\|\Phi_{\mathrm{rABAO}}(\tilde{x},\tilde{v}) - \Phi_{\mathrm{rABAO}}(x,v)\|_{a,b}^2 &= \left\|\left(\overline{x} + h\overline{v} - \frac{1}{2}h^2 Q(\overline{x}+u\overline{v}), \eta^{1/2}\overline{v} - h\eta^{1/2}Q(\overline{x}+u\overline{v})\right)\right\|_{a,b}^2,\\ &\leq 3\left(\|\overline{x}+h\overline{v}\|^2 + \frac{1}{4}h^4M^2\|\overline{x}+u\overline{v}\|^2 + a\eta\left(\|\overline{v}\|^2 + h^2M^2\|\overline{x}+u\overline{v}\|^2\right)\right)\\ &\leq 14\|(\overline{x},\overline{v})\|_{a,b}^2,\end{aligned}$$

and we can combine these estimates to achieve the contraction result of rOABAO. □

Proof for BBK. Using the relation

$$(\Phi_{\mathrm{BBK}})^n = \Phi_{\mathrm{B}_2} \circ \Phi_{\mathrm{A}} \circ (\Phi_{\mathrm{B}_1} \circ \Phi_{\mathrm{B}_2} \circ \Phi_{\mathrm{A}})^{n-1} \circ \Phi_{\mathrm{B}_1},$$

where

$$\begin{aligned}\Phi_{\mathrm{A}}(x,v) &= (x + hv, v),\\ \Phi_{\mathrm{B}_1}(x,v) &= \left(x, v + \frac{h}{2}\left(-\nabla U(x) - \gamma v + \frac{\sqrt{2\gamma}}{\sqrt{h}}\xi_1\right)\right),\\ \Phi_{\mathrm{B}_2}(x,v) &= \left(x, \left(1+\frac{h}{2}\gamma\right)^{-1}\left(v - \frac{h}{2}\nabla U(x) + \frac{\sqrt{2\gamma}}{\sqrt{h}}\xi_2\right)\right),\end{aligned}$$

and $\xi_1, \xi_2 \sim \mathcal{N}(0_d, I_d)$. We can instead prove contraction of $\Phi_{\mathrm{B}_1} \circ \Phi_{\mathrm{B}_2} \circ \Phi_{\mathrm{A}}$, by doing this we only have to deal with a single evaluation of the Hessian at each step. We will denote two synchronous realisations of $\Phi_{\mathrm{B}_1} \circ \Phi_{\mathrm{B}_2} \circ \Phi_{\mathrm{A}}$ as (x_j, v_j) and $(\tilde{x}_j, \tilde{v}_j)$ for $j \in \mathbb{N}$. Further we denote $\overline{x} := (\tilde{x}_j - x_j)$, $\overline{v} = (\tilde{v}_j - v_j)$ and $\overline{z} = (\overline{x}, \overline{v})$. We define $\overline{z}_j = (\overline{x}_j, \overline{v}_j)$ for $j = k, k+1$ by the update rule

$$\overline{x}_{k+1} = \overline{x}_k + h\overline{v}_k, \quad \overline{v}_{k+1} = \frac{1 - \frac{h}{2}\gamma}{1 + \frac{h}{2}\gamma}\left(\overline{v}_k - \frac{h}{2}Q(\overline{x}_k + h\overline{v}_k)\right) - \frac{h}{2}Q(\overline{x}_k + h\overline{v}_k),$$

where Q is defined by mean value theorem, $Q = \int_{t=0}^1 \nabla^2 U(\tilde{x}_k + h\tilde{v}_k + t(x_k - \tilde{x}_k + h(v_k - \tilde{v}_k)))\,\mathrm{d}t$, then $\nabla U(\tilde{x}_k + h\tilde{v}_k) - \nabla U(x_k + hv_k) = Q(\overline{x} + h\overline{v})$. We use the notation of (3.3), where

$$\begin{aligned}A &= -c(h)I_d + \frac{2bhQ}{\frac{1}{2}\gamma h + 1} - \frac{ah^2Q^2}{\left(\frac{1}{2}\gamma h + 1\right)^2},\\ B &= \left(b\frac{\gamma h}{\frac{1}{2}\gamma h + 1} - h - bc(h)\right)I_d + Q\left(\frac{ah\left(1 - \frac{1}{2}\gamma h\right)}{\left(\frac{1}{2}\gamma h + 1\right)^2} + \frac{2bh^2}{\frac{1}{2}\gamma h + 1}\right) - \frac{ah^3Q^2}{\left(\frac{1}{2}\gamma h + 1\right)^2},\\ C &= \left(a(1 - c(h)) - h^2 - \frac{a\left(1 - \frac{1}{2}\gamma h\right)^2}{\left(\frac{1}{2}\gamma h + 1\right)^2} - \frac{2bh\left(1 - \frac{1}{2}\gamma h\right)}{\frac{1}{2}\gamma h + 1}\right)I_d\end{aligned}$$

$$+ Q\left(\frac{2ah^2\left(1-\frac{1}{2}\gamma h\right)}{\left(\frac{1}{2}\gamma h+1\right)^2}+\frac{2bh^3}{\frac{1}{2}\gamma h+1}\right)-\frac{ah^4Q^2}{\left(\frac{1}{2}\gamma h+1\right)^2}.$$

This motivates the choice of $b=(\frac{\gamma h}{2}+1)/\gamma$. We have chosen b such that B simplifies to

$$B=-bc(h)I_d+Q\left(\frac{ah\left(1-\frac{1}{2}\gamma h\right)}{(\frac{1}{2}\gamma h+1)^2}+\frac{2bh^2}{\frac{1}{2}\gamma h+1}\right)-\frac{ah^3Q^2}{(\frac{1}{2}\gamma h+1)^2}.$$

It is sufficient to prove that $A\succ 0$ and that $C-BA^{-1}B\succ 0$, noting that A, B and C commute as they are all polynomials in Q. It is therefore sufficient to prove that $A\succ 0$ and $AC-B^2\succ 0$. First considering A with our choice of $c(h)$, a and b, we have that

$$P_A(\lambda)=-c(h)+\frac{2h\lambda}{\gamma}-\frac{h^2\lambda^2}{M\left(\frac{1}{2}\gamma h+1\right)^2}\geq h\lambda\left(\frac{7}{4\gamma}-h\right)>0,$$

where $h<\frac{7}{4\gamma}$. What remains is to show that $AC-B^2\succ 0$. This discretization is more complicated that the previous discretisations, we find that expanding P_{AC-B^2} in terms of a is convenient to show positive definiteness (as in [44]). By using for example symbolic computing one can check that $P_{AC-B^2}(\lambda)=(c_0+c_1a+c_2a^2)/h\lambda$, where

$$\begin{aligned}
(c_1+c_2a)/h\lambda=&\frac{2}{\gamma}-\frac{2}{\gamma\left(\frac{1}{2}\gamma h+1\right)^2}+\frac{h}{\left(\frac{1}{2}\gamma h+1\right)^2}+\frac{h\left(1-\frac{1}{2}\gamma h\right)}{\left(\frac{1}{2}\gamma h+1\right)^2}-\frac{2c(h)}{\gamma}\\
&+\frac{2bc(h)\left(1-\frac{1}{2}\gamma h\right)}{\left(\frac{1}{2}\gamma h+1\right)^2}-\frac{2hc(h)\left(1-\frac{1}{2}\gamma h\right)}{\left(\frac{1}{2}\gamma h+1\right)^2}+\frac{2h^2\lambda}{\gamma\left(\frac{1}{2}\gamma h+1\right)^2}+\frac{c(h)}{h\lambda\left(\frac{1}{2}\gamma h+1\right)^2}-\frac{c(h)}{h\lambda}\\
&-\frac{\gamma c(h)}{\lambda\left(\frac{1}{2}\gamma h+1\right)^2}-\frac{2bh^2c(h)\lambda}{\left(\frac{1}{2}\gamma h+1\right)^2}+\frac{\frac{1}{4}\gamma^2hc(h)}{\lambda\left(\frac{1}{2}\gamma h+1\right)^2}+\frac{h^3c(h)\lambda}{\left(\frac{1}{2}\gamma h+1\right)^2}+\frac{c(h)^2}{h\lambda}\\
&+a\left(\frac{hc(h)\lambda}{\left(\frac{1}{2}\gamma h+1\right)^2}-\frac{h\lambda}{\left(\frac{1}{2}\gamma h+1\right)^2}\right),\\
\geq&\frac{7h}{8\left(\frac{1}{2}\gamma h+1\right)^2}-\frac{2c(h)}{\gamma}-\frac{\gamma c(h)}{\lambda\left(\frac{1}{2}\gamma h+1\right)^2}-\frac{2bh^2c(h)\lambda}{\left(\frac{1}{2}\gamma h+1\right)^2}\\
\geq&\frac{64}{81}h\left(\frac{7}{8}-\frac{1}{4}-\frac{h^2m\lambda\left(\frac{\gamma h}{2}+1\right)}{2\gamma^2}\right)-\frac{mh}{2\gamma^2}\geq\frac{64}{81}h\left(\frac{5}{8}-\frac{1}{64}\right)-\frac{h}{8}\geq\frac{7h}{20},
\end{aligned}$$

where we have relied on the fact that $\gamma^2\geq 4M\geq 4m$, and, for $h<\frac{1}{4\gamma}$, that $1-\frac{1}{2}\gamma h>\frac{7}{8}$ and $1/(\frac{1}{2}\gamma h+1)>\frac{8}{9}$.

Now considering the remaining terms we have that

$$c_0=-\frac{b^2c(h)^2}{h\lambda}+\frac{4bhc(h)}{\gamma}-\frac{4h}{\gamma^2}-\frac{2h^2c(h)}{\gamma}+\frac{2c(h)}{\gamma\lambda}\geq-\frac{4h}{\gamma^2},$$

where we can combine this with the previous estimate to obtain

$$\frac{P_{AC-B^2}(\lambda)}{h\lambda}\geq h\left(\frac{7}{20M}-\frac{4}{\gamma^2}\right),$$

which holds for $\gamma\geq\sqrt{12M}$. Hence $AC-B^2\succ 0$ and our contraction results hold.

Now we need to compute the prefactors. First we consider

$$\begin{aligned}\|\Phi_{\mathrm{B}_2}\circ\Phi_{\mathrm{A}}(x,v)-\Phi_{\mathrm{B}_2}\circ\Phi_{\mathrm{A}}(\tilde{x},\tilde{v})\|_{a,b}^2 &\le \left\|\left(\overline{x}+h\overline{v},\left(1+\frac{h}{2}\gamma\right)^{-1}\left(\overline{v}-\frac{h}{2}Q(\overline{x}+h\overline{v})\right)\right)\right\|_{a,b}^2\\ &\le 3\left(\|\overline{x}\|^2+h^2\|\overline{v}\|^2+a\left(1+\frac{h}{2}\gamma\right)^{-2}\left(\|\overline{v}\|^2+\frac{h^2}{2}M^2\left(\|\overline{x}\|^2+h^2\|\overline{v}\|^2\right)\right)\right)\\ &\le 7\|(\overline{x},\overline{v})\|_{a,b}^2,\end{aligned}$$

then observe

$$\begin{aligned}\|\Phi_{\mathrm{B}_1}(x,v)-\Phi_{\mathrm{B}_1}(\tilde{x},\tilde{v})\|_{a,b}^2 \le \left\|\left(\overline{x},\overline{v}-\frac{h}{2}Q\overline{x}-\gamma\frac{h}{2}\overline{v}\right)\right\|_{a,b}^2 &\le \frac{3}{2}\left(\|\overline{x}\|^2+2a\left(1-\frac{\gamma h}{2}\right)^2\|\overline{v}\|^2+a\frac{h^2}{2}M^2\|\overline{x}\|^2\right)\\ &\le 6\|(\overline{x},\overline{v})\|_{a,b}^2,\end{aligned}$$

and combine these estimates to achieve the contraction result for BBK. □

Proofs for SPV and SVV. We now turn our attention to the stochastic position Verlet (SPV) method, where we use the fact that

$$(\Phi_{\mathrm{SPV}})^n=\Phi_{\mathrm{A}}\circ\Phi_{\mathrm{V}}\circ(\Phi_{\mathrm{A}}\circ\Phi_{\mathrm{V}})^{n-1}\circ\Phi_{\mathrm{A}},$$

where

$$\Phi_{\mathrm{V}}(x,v)=\left(x,\eta v-\frac{1-\eta}{\gamma}\nabla U(x)+\sqrt{1-\eta^2}\xi\right),$$

and $\xi\sim\mathcal{N}(0_d,I_d)$. We can instead prove contraction of $\Phi_{\mathrm{A}}\circ\Phi_{\mathrm{V}}$. This simplifies the estimation task but will introduce a prefactor to take into account the estimation of the head and tail operators $\mathcal{AV}$ and $\mathcal{A}$. We denote two synchronous realizations of $\mathcal{AV}$ as (x_j,v_j) and $(\tilde{x}_j,\tilde{v}_j)$ for $j\in\mathbb{N}$. Next, write $\overline{x}:=(\tilde{x}_j-x_j)$, $\overline{v}=(\tilde{v}_j-v_j)$ and $\overline{z}=(\overline{x},\overline{v})$, where $\overline{z}_j=(\overline{x}_j,\overline{v}_j)$ for $j=k,k+1$ for $k\in\mathbb{N}$, and $\overline{z}_k$ are defined by the update rule

$$\overline{x}_{k+1}=\overline{x}_k+h\overline{v}_k,\quad \overline{v}_{k+1}=\eta\overline{v}_k-\frac{1-\eta}{\gamma}Q(\overline{x}_k+h\overline{v}_k),$$

where $\eta=\exp(-\gamma h)$ and we define Q from the mean value theorem as $Q=\int_{t=0}^1\nabla^2U(\tilde{x}_k+h\tilde{v}_k+t(x_k-\tilde{x}_k+h(v_k-\tilde{v}_k)))\,\mathrm{d}t$, then $\nabla U(\tilde{x}_k+h\tilde{v}_k)-\nabla U(x_k+hv_k)=Q(\overline{x}+h\overline{v})$. We use the notation of (3.3), where

$$\begin{aligned}A&=-c(h)I_d+\frac{2b(1-\eta)}{\gamma}Q-\frac{a(1-\eta)^2}{\gamma^2}Q^2,\\ B&=(b(1-\eta)-h-bc(h))I_d+\left(\frac{a\eta(1-\eta)}{\gamma}+\frac{2bh(1-\eta)}{\gamma}\right)Q-\frac{ah(1-\eta)^2}{\gamma^2}Q^2,\\ C&=\left(a(1-\eta^2)-ac(h)-2b\eta h-h^2\right)I_d+\frac{2h(1-\eta)}{\gamma}(a\eta+bh)Q-\frac{a(1-\eta)^2h^2Q^2}{\gamma^2},\end{aligned}$$

which motivates the choice of $b=\frac{h}{1-\eta}$. We have chosen b, such that B simplifies to

$$B=-bc(h)I_d+\left(\frac{a\eta(1-\eta)}{\gamma}+\frac{2bh(1-\eta)}{\gamma}\right)Q-\frac{ah(1-\eta)^2}{\gamma^2}Q^2.$$

It is sufficient to prove that $A\succ 0$ and that $C-BA^{-1}B\succ 0$, noting that A, B and C commute as they are all polynomials in Q; it is sufficient to prove that $A\succ 0$ and $AC-B^2\succ 0$. We will use P_A and P_{AC-B^2} to denote

the eigenvalues of the respective matrices, which will be polynomials in terms of the eigenvalues of Q, which we know belong in $[m, M]$. First considering A with our choice of $c(h)$, a and b we have that

$$P_A(\lambda) = -\frac{mh}{4\gamma} + \frac{2h}{\gamma}\lambda - \frac{(1-\eta)^2}{M\gamma^2}\lambda^2 \geq \frac{h\lambda}{\gamma}\left(\frac{7}{4} - \frac{(1-\eta)^2}{h\gamma}\right) > 0,$$

as $\frac{h\gamma}{2} \leq 1-\eta \leq h\gamma$ for $h < \frac{1}{2\gamma}$.

Next, $AC - B^2 \succ 0$ is shown as follows

$$\begin{aligned}
\frac{P_{AC-B^2}(\lambda)}{\lambda} &= -\frac{(1-\eta)^2\lambda}{\gamma^2M^2} + h\left(\frac{2\big(1-\eta^2\big)}{\gamma M} + \frac{c(h)}{h}\left(\frac{(1-\eta)^2\lambda}{\gamma^2M^2} - \frac{1-\eta^2}{\lambda M}\right)\right) \\
&\quad + h^2\left(-\frac{2c(h)}{h\gamma M}\big(1-\eta^2\big) + \frac{c(h)^2}{h^2M\lambda} + \frac{\lambda\big(1-\eta^2\big)}{\gamma^2M}\right) \\
&\quad + h^3\left(-\frac{2(1+\eta)}{(1-\eta)\gamma} + \frac{c(h)}{h}\left(\frac{(1+\eta)}{(1-\eta)\lambda} - \frac{\big(1-\eta^2\big)\lambda}{\gamma^2M}\right)\right) + h^4\left(\frac{2c(h)(1+\eta)}{\gamma h(1-\eta)} - \frac{c(h)^2}{h^2(1-\eta)^2\lambda}\right) \\
&\geq -\frac{(1-\eta)h}{\gamma M} + h\left(\frac{2(1-\eta)(1+\eta)}{\gamma M} - \frac{1-\eta^2}{4\gamma M}\right) + h^3\left(-\frac{2(1+\eta)}{(1-\eta)\gamma}\right) \\
&\geq \frac{3\big(1-\eta^2\big)h}{4\gamma M} - \frac{8\big(1-\eta^2\big)h}{\gamma^3} > 0,
\end{aligned}$$

where we have imposed the restriction $\gamma \geq \sqrt{11M}$. Hence $AC - B^2 \succ 0$ and our contraction results hold. All computations can be checked using symbolic computing. Now we wish to explicitly compute estimates of the prefactors and the remaining terms, first, we consider

$$\begin{aligned}
\|\Phi_{\mathrm{V}} \circ \Phi_{\mathrm{A}}(x,v) - \Phi_{\mathrm{V}} \circ \Phi_{\mathrm{A}}(\tilde{x},\tilde{v})\|_{a,b}^2 &\leq \left\|\left(\overline{x} + \frac{h}{2}\overline{v}, \eta\overline{v} - \frac{1-\eta}{\gamma}Q\left(\overline{x} + \frac{h}{2}\overline{v}\right)\right)\right\|_{a,b}^2 \\
&\leq 3\left(\left(1 + 2\left(\frac{1-\eta}{\gamma}\right)^2 M\right)\|\overline{x}\|^2 + \left(a\eta^2 + \frac{h^2}{4}\left(1 + 2\left(\frac{1-\eta}{\gamma}\right)^2 M\right)\right)\|\overline{v}\|^2\right) \\
&\leq 7\|(\overline{x},\overline{v})\|_{a,b}^2.
\end{aligned}$$

and

$$\|\Phi_{\mathrm{A}}(x,v) - \Phi_{\mathrm{A}}(\tilde{x},\tilde{v})\|_{a,b}^2 \leq \left\|\left(\overline{x} + \frac{h}{2}\overline{v}, \overline{v}\right)\right\|_{a,b}^2 \leq 7\|(\overline{x},\overline{v})\|_{a,b}^2.$$

We can combine these estimates to achieve the contraction result for SPV.

Now for the stochastic velocity Verlet we have that $(\mathcal{V}\mathcal{A}\mathcal{V})^n = \mathcal{V}(\mathcal{A}\mathcal{V}\mathcal{V})^{n-1}\mathcal{A}\mathcal{V}$. We will now focus our attention on proving contraction of $\mathcal{A}\mathcal{V}\mathcal{V}$. From the fact that $\mathcal{V}(h/2)\mathcal{V}(h/2) = \mathcal{V}(h)$, contraction of $\mathcal{A}\mathcal{V}\mathcal{V}$ is shown in our argument for the stochastic position Verlet.

Now we wish to explicitly compute estimates of the prefactors and the remaining terms, first, we consider

$$\|\Phi_{\mathrm{V}}(x,v) - \Phi_{\mathrm{V}}(\tilde{x},\tilde{v})\|_{a,b}^2 \leq \left\|\left(\overline{x}, \eta^{1/2}\overline{v} - \frac{1-\eta^{1/2}}{\gamma}Q\overline{x}\right)\right\|_{a,b}^2 \leq 6\|(\overline{x},\overline{v})\|_{a,b}^2.$$

and

$$\|\Phi_{\mathrm{V}} \circ \Phi_{\mathrm{A}}(x,v) - \Phi_{\mathrm{V}} \circ \Phi_{\mathrm{A}}(\tilde{x},\tilde{v})\|_{a,b}^2 \leq \left\|\left(\overline{x} + h\overline{v}, \eta^{1/2}\overline{v} - \frac{1-\eta^{1/2}}{\gamma}Q(\overline{x} + h\overline{v})\right)\right\|_{a,b}^2$$

$$\leq 3\left(\left(1+2\left(\frac{1-\eta^{1/2}}{\gamma}\right)^2 M\right)\|\overline{x}\|^2 + \left(a\eta + h^2\left(1+2\left(\frac{1-\eta^{1/2}}{\gamma}\right)^2 M\right)\right)\|\overline{v}\|^2\right)$$
$$\leq 7\|(\overline{x},\overline{v})\|_{a,b}^2.$$

If we combine these estimates we have the required result. □

7. Stochastic gradients

In many machine learning and statistics applications, the cost of a gradient evaluation is high as it requires an evaluation of the entire data set. Instead, stochastic gradients are used, where one takes a random sub-sample of the data-set to approximate the gradient with an unbiased estimate. An analysis of convergence rates of the discretizations with stochastic gradients is performed in [35].

Definition 7.1. A *stochastic gradient approximation* of a potential U is defined by a function $\mathcal{G} : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$ and a probability distribution ρ on a Polish space Ω, satisfying that $\mathcal{G}$ is measurable on $(\Omega, \mathcal{F})$, and that for every $x \in \mathbb{R}^d$, for $W \sim \rho$,

$$\mathbb{E}(\mathcal{G}(x, W)) = \nabla U(x).$$

The function $\mathcal{G}$ and the distribution ρ together define the stochastic gradient, which we denote as $(\mathcal{G}, \rho)$.

In many applications, this can dramatically reduce the computational cost as the approximation will usually come at a fraction of the workload. The numerical schemes considered in [44] and in this article are roughly one gradient evaluation per sample, roughly meaning when negating the extra gradient evaluations at the head and tail of the simulation of the algorithm (for the first and last sample). This is done by using the same gradient evaluation in consecutive velocity updates when the position has not been updated, for an increase in computational efficiency. We treat this case when it also comes to stochastic gradients to improve computational efficiency, for example in the BAOAB scheme the last B and first B of each iteration will share an estimate of the force (using the same stochastic gradient evaluation). For clarity, a stochastic gradient version of each algorithm is provided in Appendix A.

In our convergence rate estimates, we impose the assumption that the variance of the Jacobian of the stochastic gradient is bounded.

Assumption 7.2. *We assume that the Jacobian of the stochastic gradient $\mathcal{G}$, $D_x\mathcal{G}(x, W)$ exists and it is measurable on $(\Omega, \mathcal{F})$. We also assume there exists $C_G > 0$ such that for $W \sim \rho$,*

$$\sup_{x\in\mathbb{R}^d} \mathbb{E}\|D_x\mathcal{G}(x, W) - \nabla^2 U(x)\|^2 \leq C_G.$$

Our results extend to the stochastic gradient setting by including a coupling in the mini-batches or the stochastic gradients in the same way as OABAO in [35]. Further the results for the other schemes in this article and [44] are generalized to the case of stochastic gradients when the same stochastic gradient is chosen as in Appendix A for each algorithm. In this way, there is still one gradient evaluation per step.

We remark that these assumptions hold when $\mathcal{G}$ is of the form $\mathcal{G}(x, W) = \sum_{i\in W} \nabla U_i(x)$, where $W \in \Omega \subset [N]^B$, B is the batch size, and $(U_i)_{i\in[N]}$ are strongly convex and gradient-Lipschitz, this is the setting of minibatching in many Bayesian learning problems. The contraction results of Theorem 5.1 are extended to the stochastic gradient setting in Theorem 7.3. We remark that our assumptions are more flexible than the assumptions imposed in [35], where they assume that the stochastic gradient is universally gradient Lipschitz and strongly convex over the entire state space Ω.

Theorem 7.3. *Consider the numerical schemes for stochastic gradient kinetic Langevin dynamics given in Appendix A and Table 3, where the potential U is m-strongly convex and M-∇Lipschitz. Assume a stochastic gradient approximation defined by $(\mathcal{G}, \rho)$ (see Def. 7.1) satisfying Assumption 7.2 with constant C_G. We consider*

TABLE 3. Contraction rates $c(h)$ and preconstants $C(h)$ in (7.1).

Algorithm	$c(h)$	$C(h)$
EM	$mh/2\gamma - 2h^2C_G/M$	1
BBK	$mh/4\gamma - 4h^2C_G/M$	$7 + 3h^2C_G/M$
SPV	$mh/4\gamma - 4h^2C_G/M$	$7 + 12h^2C_G/M$
SVV	$mh/4\gamma - 4h^2C_G/M$	$7 + 6h^2C_G/M$
BAOAB	$h^2m/4(1-\eta) - 5h^2C_G(\eta/M + \frac{1}{4}h^2)$	$7 + 3h^2C_G/M$
OBABO	$h^2m/4(1-\eta) - 4h^2C_G/M$	$8 + 3h^2C_G/M$
rOABAO	$h^2m/4(1-\eta) - 5h^2C_G(\eta/M + \frac{1}{4}h^2)$	$8 + 8h^2C_G/M$
SES/EB	$mh/4\gamma - 4h^2C_G/M$	1

any sequence of synchronously coupled random variables (in Brownian increment and stochastic gradient) with initial conditions $(x_0, v_0) \in \mathbb{R}^{2d}$ *and* $(\tilde{x}_0, \tilde{v}_0) \in \mathbb{R}^{2d}$.

Under stepsize restrictions $h < h_0$ *and* $\gamma \geq \gamma_0$, *where* h_0 *and* γ_0 *are given in Table 2, and given initial conditions* $(x_0, v_0) \in \mathbb{R}^{2d}$ *and* $(\tilde{x}_0, \tilde{v}_0) \in \mathbb{R}^{2d}$ *we have the expected contraction*

$$\left(\mathbb{E}\|(x_k - \tilde{x}_k, v_k - \tilde{v}_k)\|_{a,b}^2\right)^{1/2} \leq C(h)(1 - c(h))^{s/2}\|(x_0 - \tilde{x}_0, v_0 - \tilde{v}_0)\|_{a,b}, \tag{7.1}$$

with norm given with constants $a = 1/M$ *and* b, *where the contraction rate* $c(h)$ *and the preconstant* $C(h)$ *are given in Table 3 and* b *and the number of steps* s *are given in Table 2 with all parameters specific to each scheme.*

Remark 7.4. Compared to Theorem 5.1 with deterministic gradients, Theorem 7.3 demonstrates expected contraction, because the randomness from the stochastic gradients can be integrated out. This allows us to make Assumption 7.2 less restrictive than it would need to be otherwise. Rather than deterministic contraction we have contraction in expectation.

Remark 7.5. Our analysis suggests a reduction in the convergence rate for large gradient noises, which we have observed in numerical experiments when using sub-sampling and very small batches. For large gradient noise C_G and stepsize h it is possible that these bounds become vacuous and the loss of convergence was also confirmed in our experiments.

Remark 7.6. The implementation of the BAOAB algorithm and other algorithms considered in Section A is non-Markovian, because the last B step of each iteration and the first B step of the next iteration share the same stochastic gradient sample. This is not an issue in our convergence rate framework as we consider convergence of a different operator, which is Markovian, for example $\mathcal{ABAO}$ for BAOAB, which does not share stochastic gradients with consecutive iterations. We simplify the problem into proving convergence of an operator which only has a single gradient evaluation and hence is Markovian in the stochastic gradient setting.

Proof of Theorem 7.3. For stochastic gradients, we synchronously couple Brownian increments as well as the stochastic gradients. We wish to instead consider expected contraction of the update rule we used to prove contraction in the full gradient setting, *i.e.* for synchronously coupled (in stochastic gradient and Brownian increment) iterates $(x_l, v_l), (\tilde{x}_l, \tilde{v}_l) \in \mathbb{R}^{2d}$ for $l \in \mathbb{N}$ and $(\overline{x}_l, \overline{v}_l) = (\tilde{x}_l, \tilde{v}_l) - (x_l, v_l)$ and for $k \in \mathbb{N}$

$$\mathbb{E}\|(\overline{x}_{k+1}, \overline{v}_{k+1})\|_{a,b}^2 \leq (1 - c(h))\|(\overline{x}_k, \overline{v}_k)\|_{a,b}^2$$

then we have

$$\mathbb{E}\left(\overline{z}_k^T P^T M P \overline{z}_k\right) \leq (1 - c(h))\overline{z}_k^T M \overline{z}_k.$$

Now if $\tilde{Q}$ is defined through the mean value theorem of $D_x\mathcal{G}$ (the Jacobian of $\mathcal{G}$) and is a random variable in W, such that $\mathbb{E}(\tilde{Q}) = Q$. Then P^TMP is of the form

$$\mathcal{P}\Big(\tilde{Q}\Big) = \begin{pmatrix} P_1\Big(\tilde{Q}\Big) & P_2\Big(\tilde{Q}\Big) \\ P_2\Big(\tilde{Q}\Big) & P_3\Big(\tilde{Q}\Big) \end{pmatrix},$$

where P_1, P_2 and P_3 are quadratics in $\tilde{Q}$ of the form

$$\begin{aligned} P_1\Big(\tilde{Q}\Big) &= a_0 + a_1\tilde{Q} + a_2\tilde{Q}^2, \\ P_2\Big(\tilde{Q}\Big) &= b_0 + b_1\tilde{Q} + b_2\tilde{Q}^2, \\ P_3\Big(\tilde{Q}\Big) &= c_0 + c_1\tilde{Q} + c_2\tilde{Q}^2. \end{aligned}$$

Then we have

$$\mathbb{E}\big(P^TMP\big) = \mathcal{P}(Q) + \begin{pmatrix} a_2\mathbb{E}\Big(\tilde{Q}-Q\Big)^2 & b_2\mathbb{E}\Big(\tilde{Q}-Q\Big)^2 \\ b_2\mathbb{E}\Big(\tilde{Q}-Q\Big)^2 & c_2\mathbb{E}\Big(\tilde{Q}-Q\Big)^2 \end{pmatrix},$$

in combination with the Theorem 5.1 result we have that

$$\begin{aligned} \mathbb{E}\|(x_{k+1}, v_{k+1})\|_{a,b}^2 &\leq (1-c(h))\|(x_k, v_k)\|_{a,b}^2 + z_k^T \begin{pmatrix} a_2\mathbb{E}\Big(\tilde{Q}-Q\Big)^2 & b_2\mathbb{E}\Big(\tilde{Q}-Q\Big)^2 \\ b_2\mathbb{E}\Big(\tilde{Q}-Q\Big)^2 & c_2\mathbb{E}(\tilde{Q}-Q)^2 \end{pmatrix} z_k \\ &= (1-c(h))\|(x_k, v_k)\|_{a,b}^2 + z_k^T\mathcal{R}\Big(\tilde{Q}\Big)z_k, \end{aligned}$$

where we use the notation

$$\mathcal{R}\Big(\tilde{Q}\Big) := \begin{pmatrix} a_2\mathbb{E}\Big(\tilde{Q}-Q\Big)^2 & b_2\mathbb{E}\Big(\tilde{Q}-Q\Big)^2 \\ b_2\mathbb{E}\Big(\tilde{Q}-Q\Big)^2 & c_2\mathbb{E}\Big(\tilde{Q}-Q\Big)^2 \end{pmatrix}.$$

Then we will bound the remainder term $z^T\mathcal{R}(\tilde{Q})z$ separately for each scheme and we refer the reader to the contraction estimate proofs in [44] for the coefficients a_2, b_2 and c_2 for Euler–Maruyama, BAOAB, OBABO and SES and to Section 6 for the schemes analyzed in this article. We remark that we analyze the update rules for which we proved contraction for all the schemes, which aren't necessarily the same as the scheme for example we analyze $\mathcal{ABAO}$ for BAOAB. Throughout these estimates we use the equivalence of norms in Section 2.2 and the stepsize and parameter restrictions imposed in the contraction estimates of the respective schemes. We define $\mathrm{Var}(\tilde{Q}) := \mathbb{E}(\tilde{Q}-Q)^2$, to be the variance of $\tilde{Q}$.

(1) For the Euler–Maruyama $a_2 > 0$ and $b_2 = c_2 = 0$, therefore we have for $z = (x, v) \in \mathbb{R}^{2d}$

$$z^T\mathcal{R}\Big(\tilde{Q}\Big)z \leq h^2aC_G\|x\|^2 \leq 2h^2aC_G\|z\|_{a,b}^2.$$

(2) For BAOAB we have for $\mathcal{ABAO}$

$$\begin{aligned} z^T\mathcal{R}\Big(\tilde{Q}\Big)z &= ah^2\left(\eta^2 + b\eta hM + \frac{h^2}{4}M\right)\left(x + \frac{h}{2}v\right)^T \mathrm{Var}\Big(\tilde{Q}\Big)\left(x + \frac{h}{2}v\right) \\ &\leq 4ah^2C_G\left(\eta^2 + b\eta hM + \frac{h^2}{4}M\right)\|(x, v)\|_{a,b}^2 \leq 5ah^2C_G\left(\eta + \frac{h^2}{4}M\right)\|(x, v)\|_{a,b}^2. \end{aligned}$$

(3) For OBABO we have for $\mathcal{ABOB}$

$$z^T\mathcal{R}\Big(\tilde{Q}\Big)z = a(\eta+1)^2\frac{h^2}{4}\Big((x+hv)^T\mathrm{Var}\Big(\tilde{Q}\Big)(x+hv)\Big) \leq a(\eta+1)^2h^2C_G\|(x,v)\|_{a,b}^2 \leq 4ah^2C_G\|(x,v)\|_{a,b}^2.$$

(4) For SES we have

$$z^T\mathcal{R}\Big(\tilde{Q}\Big)z = \left(\frac{a(1-\eta)^2}{\gamma^2} + \frac{2b(1-\eta)(-1+\eta+\gamma h)}{\gamma^3} + \frac{(-1+\eta+\gamma h)^2}{\gamma^4}\right)x^T\mathrm{Var}\Big(\tilde{Q}\Big)x \leq 4ah^2C_G\|(x,v)\|_{a,b}^2.$$

(5) For rOABAO we have for $r\mathcal{ABAO}$

$$\begin{aligned} z^T\mathcal{R}\Big(\tilde{Q}\Big)z &= \left(a\eta^2h^2 + b\eta h^3 + \frac{1}{4}h^4\right)\Big((x+uv)^T\mathrm{Var}\Big(\tilde{Q}\Big)(x+uv)\Big) \\ &\leq 4\left(a\eta^2h^2 + b\eta h^3 + \frac{1}{4}h^4\right)C_G\|(x,v)\|_{a,b}^2 \leq 5h^2\left(a\eta + \frac{1}{4}h^2\right)C_G\|(x,v)\|_{a,b}^2. \end{aligned}$$

(6) For BBK we have for $\mathcal{AB}_2\mathcal{B}_1$

$$z^T\mathcal{R}\Big(\tilde{Q}\Big)z = \frac{ah^2}{\left(\frac{1}{2}\gamma h+1\right)^2}(x+hv)^T\mathrm{Var}\Big(\tilde{Q}\Big)(x+hv) \leq 4ah^2C_G\|(x,v)\|_{a,b}^2.$$

(7) For SPV and SVV we have for $\mathcal{AB}_s$

$$z^T\mathcal{R}\Big(\tilde{Q}\Big)z = \frac{a(1-\eta)^2}{\gamma^2}(x+hv)^T\mathrm{Var}\Big(\tilde{Q}\Big)(x+hv) \leq 4ah^2C_G\|(x,v)\|_{a,b}^2.$$

Due to the stochastic gradient, we have to recompute the preconstants for BAOAB, OBABO, rOABAO, BBK, SPV and SVV. We will use the fact that

$$\mathbb{E}\left\|\tilde{Q}x\right\|^2 = \mathbb{E}\left\|\Big(\tilde{Q}-Q\Big)x\right\|^2 + \|Qx\|^2 \leq \big(C_G+M^2\big)\|x\|^2$$

to compute the preconstants, we will now perform the calculation for all necessary schemes, we refer you to [44] and Section 6 for the operators at the head and tail of the contraction estimates that we will need to bound.

(1) For BAOAB in [44], contraction of $\mathcal{ABAO}$ was proven and hence we need to bound $\mathcal{BAO}$ and $\mathcal{AB}$. For $k\in\mathbb{N}$ and iteration $(x_k,v_k),(\tilde{x}_k,\tilde{v}_k)\in\mathbb{R}^{2d}$ we first estimate

$$\begin{aligned} \mathbb{E}\|\Phi_{\mathrm{AB}}(\tilde{x}_k,\tilde{v}_k) - \Phi_{\mathrm{AB}}(x_k,v_k)\|_{a,b}^2 &\leq 3\left(\left(1+\frac{ah^2}{2}\big(C_G+M^2\big)\right)\|\overline{x}_k\|^2 + \left(a+\frac{h^2}{4}+\frac{ah^4}{8}\big(M^2+C_G\big)\right)\|\overline{v}_k\|^2\right) \\ &\leq 7\|(\overline{x}_k,\overline{v}_k)\|_{a,b}^2 + 3\left(\frac{1}{2}ah^2C_G\|\overline{x}_k\|^2 + a\frac{h^4}{8}C_G\|\overline{v}_k\|^2\right) \\ &\leq \big(7+3ah^2C_G\big)\|(\overline{x}_k,\overline{v}_k)\|_{a,b}^2, \end{aligned}$$

and where we have defined $\overline{x}_k = \tilde{x}_k - x_k$ and $\overline{v}_k = \tilde{v}_k - v_k$. We then estimate

$$\begin{aligned} \mathbb{E}\|\Phi_{\mathrm{BAO}}(\tilde{x}_k,\tilde{v}_k) - \Phi_{\mathrm{BAO}}(x_k,v_k)\|_{a,b}^2 &\leq 3\left(\left(1+\frac{h^2}{4}\left(a\eta^2+\frac{h^2}{2}\right)\big(C_G+M^2\big)\right)\|\overline{x}_k\|^2 + \left(a\eta^2+\frac{h^2}{2}\right)\|\overline{v}_k\|^2\right) \\ &\leq \left(7+\frac{h^2}{2}C_G\left(a\eta^2+\frac{h^2}{2}\right)\right)\|(\overline{x}_k,\overline{v}_k)\|_{a,b}^2. \end{aligned}$$

Now combining these constants we have the relevant preconstant for the constraction estimate for BAOAB.

(2) Similarly we have the following bounds to compute preconstants for OBABO:

$$\mathbb{E}\|\Phi_{\mathrm{ABO}}(\tilde{x}_k,\tilde{v}_k)-\Phi_{\mathrm{ABO}}(x_k,v_k)\|_{a,b}^2 \le 3\left(\left(1+\frac{ah^2}{2}\big(C_G+M^2\big)\right)\|\overline{x}_k\|^2+\left(a+h^2+\frac{ah^4}{2}\big(C_G+M^2\big)\right)\|\overline{v}_k\|^2\right)$$
$$\le \big(8+3ah^2C_G\big)\|(\overline{x}_k,\overline{v}_k)\|_{a,b}^2,$$

and then we estimate

$$\mathbb{E}\|\Phi_{\mathrm{OB}}(\tilde{x}_k,\tilde{v}_k)-\Phi_{\mathrm{OB}}(x_k,v_k)\|_{a,b}^2 \le 3\left(\left(\frac{1}{2}+\frac{ah^2}{4}\big(C_G+M^2\big)\right)\|\overline{x}_k\|^2+a\|\overline{v}_k\|^2\right)$$
$$\le \big(6+2ah^2C_G\big)\|(\overline{x}_k,\overline{v}_k)\|_{a,b}^2.$$

Now combining these constants we have the relevant preconstant for the contraction estimate for OBABO.

(3) Now considering rOABAO the preconstant bound for $\mathcal{O}$ is unaffected so we just need to estimate the operator $r\mathcal{ABA}\mathcal{O}$. We have

$$\mathbb{E}\|\Phi_{\mathrm{rABAO}}(\tilde{x}_k,\tilde{v}_k)-\Phi_{\mathrm{rABAO}}(x_k,v_k)\|_{a,b}^2 \le \big(14+14ah^2C_G\big)\|(\overline{x}_k,\overline{v}_k)\|_{a,b}^2,$$

and we can combine this wiht the bound for $\mathcal{O}$ given in Section 6.

(4) For BBK we have the estimates

$$\mathbb{E}\|\Phi_{\mathrm{B}_2}\circ\Phi_{\mathrm{A}}(\tilde{x}_k,\tilde{v}_k)-\Phi_{\mathrm{B}_2}\circ\Phi_{\mathrm{A}}(x_k,v_k)\|_{a,b}^2 \le \big(7+3ah^2C_G\big)\|(\overline{x}_k,\overline{v}_k)\|_{a,b}^2,$$

and

$$\mathbb{E}\|\Phi_{\mathrm{B}_1}(\tilde{x}_k,\tilde{v}_k)-\Phi_{\mathrm{B}_1}(x_k,v_k)\|_{a,b}^2 \le \big(6+2ah^2C_G\big)\|(\overline{x}_k,\overline{v}_k)\|_{a,b}^2,$$

which we can combine to get the desired preconstant.

(5) For SPV we estimate

$$\mathbb{E}\|\Phi_{\mathrm{V}}\circ\Phi_{\mathrm{A}}(\tilde{x}_k,\tilde{v}_k)-\Phi_{\mathrm{V}}\circ\Phi_{\mathrm{A}}(x_k,v_k)\|_{a,b}^2 \le \big(7+12ah^2C_G\big)\|(\overline{x}_k,\overline{v}_k)\|_{a,b}^2,$$

and we can use the previous estimate of $\mathcal{A}$ to estimate the preconstant.

(6) Finally for SVV we estimate

$$\mathbb{E}\|\Phi_{\mathrm{V}}(\tilde{x}_k,\tilde{v}_k)-\Phi_{\mathrm{V}}(x_k,v_k)\|_{a,b}^2 \le \big(6+3ah^2C_G\big)\|(\overline{x}_k,\overline{v}_k)\|_{a,b}^2,$$

and

$$\mathbb{E}\|\Phi_{\mathrm{V}}\circ\Phi_{\mathrm{A}}(\tilde{x}_k,\tilde{v}_k)-\Phi_{\mathrm{V}}\circ\Phi_{\mathrm{A}}(x_k,v_k)\|_{a,b}^2 \le \big(7+6ah^2C_G\big)\|(\overline{x}_k,\overline{v}_k)\|_{a,b}^2.$$

We have all desired preconstants and penalty terms for the contraction rate when the gradient is a stochastic estimate. □

Proposition 7.7. *Consider the numerical schemes for stochastic gradient kinetic Langevin dynamics given in Appendix A and Table 3, where the potential U is m-strongly convex and M-∇Lipschitz. Assume a stochastic gradient approximation defined by $(\mathcal{G},\rho)$ (see Def. 7.1) satisfying Assumption 7.2 with constant C_G. We use P_h to denote the marginal transition kernel of the numerical schemes. For the constants given in Tables 2 and 3 we have for any two synchronously coupled chains, (x_k,v_k) and $(\tilde{x}_k,\tilde{v}_k)$ under the assumptions specific to the schemes imposed in Theorem 7.3 we have for all $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^{2d})$, and all $k\in\mathbb{N}$,*

$$\mathcal{W}_2^2\big(\nu P_h^k,\mu P_h^k\big) \le 3C(h)\max\left\{M,\frac{1}{M}\right\}(1-c(h))^k\mathcal{W}_2^2(\nu,\mu).$$

Proof. We remark that the stochastic gradients are independent from position and hence can be marginalized out in the following estimates over the extended state space. We first denote $\hat{P}_h$ to be the transition kernel for which contraction is proved in Theorem 5.1 or [44]. For example stochastic gradient $\mathcal{ABAO}$ for BAOAB. From Theorem 7.3 and following Corollary 20 of [47] we know that for $z_k = (x_k, v_k)$, $\tilde{z}_k = (\tilde{x}_k, \tilde{v}_k)$ such that $z_0 = (x_0, v_0) \sim \mu$ and $\tilde{z}_0 = (\tilde{x}_0, \tilde{v}_0) \sim \nu$ and $(z_0, \tilde{z}_0)$ is a $\mathcal{W}_2$ optimal coupling of μ and ν then under $\hat{P}_h$

$$\mathcal{W}_{2,a,b}^2\left(\mu\hat{P}_h^k, \nu\hat{P}_h^k\right) \leq \mathbb{E}\|z_k - \tilde{z}_k\|_{a,b}^2 \leq (1 - c(h))^k \mathcal{W}_{2,a,b}^2(\mu, \nu),$$

then we can use the equivalence of norms in Section 2.2 and the preconstant estimates of Table 3 to achieve the desired result for P_h. □

Remark 7.8. We remark that the contraction rate of BAOAB and rOABAO can be upper bounded by a simpler form, for example, $\mathcal{O}(mh^2/(1-\eta) - h^2 C_G/M)$, but we have included the more detailed estimate because it has the property that as you take the friction parameter $\gamma \to \infty$ then the contraction rate is of the same order as the overdamped Langevin dynamics scheme as discussed in Section 7.1.

7.1. Overdamped Langevin dynamics

If we take the limit of the friction parameter ($\gamma \to \infty$) in (1.1), with a time-rescaling ($t' = \gamma t$) we have the overdamped Langevin dynamics given by (see [50], Sect. 6.5)

$$\mathrm{d}X_t = -\nabla U(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t. \tag{7.2}$$

This equation has invariant measure with density proportional to $\exp(-U(x))$, like the marginal of the underdamped counterpart.

Two discretizations of the SDE (7.2) considered in [44] and linked to the BAOAB and OBABO schemes through high friction limits are the Euler–Maruyama (EM) method which is defined by the update rule

$$x_{k+1} = x_k - h\nabla U(x_k) + \sqrt{2h}\xi_{k+1}, \tag{7.3}$$

and the BAOAB limit method of Leimkuhler and Matthews (LM) [40, 42] which is defined by the update rule

$$x_{k+1} = x_k - h\nabla U(x_k) + \sqrt{2h}\frac{\xi_{k+1} + \xi_k}{2}.$$

Now we can apply coupling arguments to the overdamped dynamics in a simpler way (using the standard Euclidean distance), but we will also consider the case of stochastic gradients. Coupling arguments in the overdamped setting have been extensively studied with and without stochastic gradients (see [18, 20, 21, 25–27, 29]).

Proposition 7.9. *Consider the Euler–Maruyama and Leimkuhler–Matthews schemes for stochastic gradient Langevin dynamics, where the potential U is m-strongly convex and M-∇Lipschitz. Assume a stochastic gradient approximation defined by $(\mathcal{G}, \rho)$ (see Def. 7.1) satisfying Assumption 7.2 with constant C_G. We consider any sequence of synchronously coupled random variables (in Brownian increment and stochastic gradient) with initial conditions $x_0 \in \mathbb{R}^d$ and $y_0 \in \mathbb{R}^d$. We have the contraction property*

$$\mathbb{E}\|x_k - y_k\|^2 \leq \left(1 - \left(hm(2 - hM) - h^2 C_G\right)\right)^k \|x_0 - y_0\|^2.$$

Proof. If we first consider two chains x_k and y_k with shared noise such that

$$x_{k+1} = x_k - h\mathcal{G}(x_k, W_k) + \sqrt{2h}\xi_k, \quad y_{k+1} = y_k - h\mathcal{G}(y_k, W_k) + \sqrt{2h}\xi_k,$$

where $W_k \sim \rho$ and $\xi_k \sim \mathcal{N}(0_d, I_d)$ for all $k \in \mathbb{N}$ and this can be either the Euler–Maruyama or Leimkuhler–Matthews method, we have chosen Euler–Maruyama. Then we have that

$$\begin{aligned}\mathbb{E}\|x_{k+1} - y_{k+1}\|^2 &= \mathbb{E}\|x_k - y_k + h(-\mathcal{G}(x_k, W_k)) - (-\mathcal{G}(y_k, W_k))\|^2 \\ &= \|x_k - y_k\|^2 - 2h\langle x_k - y_k, Q(x_k - y_k)\rangle \\ &\quad + h^2\big\langle x_k - y_k, Q^2(x_k - y_k)\big\rangle + h^2\Big\langle x_k - y_k, \mathrm{Var}\Big(\tilde{Q}\Big)(x_k - y_k)\Big\rangle,\end{aligned}$$

where $\tilde{Q} = \int_{t=0}^{1} D_x\mathcal{G}(x_k + t(y_k - x_k), W_k)\,\mathrm{d}t$ and $Q = \mathbb{E}(\tilde{Q})$. Q has eigenvalues which are bounded between m and M, so $Q^2 \preceq MQ$, and hence

$$h^2\big\langle x_k - y_k, Q^2(x_k - y_k)\big\rangle \leq h^2 M\langle x_k - y_k, Q(x_k - y_k)\rangle.$$

If we impose Assumption 7.2 on the Jacobian of the stochastic gradient then we have that

$$h^2\Big\langle x_k - y_k, \mathrm{Var}\Big(\tilde{Q}\Big)(x_k - y_k)\Big\rangle \leq h^2 C_G\|x_k - y_k\|^2,$$

and therefore

$$\begin{aligned}\|x_{k+1} - y_{k+1}\|^2 &\leq \|x_k - y_k\|^2 - h(2 - hM)\langle x_k - y_k, Q(x_k - y_k)\rangle + h^2\Big\langle x_k - y_k, \mathrm{Var}\Big(\tilde{Q}\Big)(x_k - y_k)\Big\rangle \\ &\leq \|x_k - y_k\|^2\big(1 - hm(2 - hM) + h^2 C_G\big).\end{aligned}$$

□

In the same way as Proposition 7.7 the contraction property of Proposition 7.9 implies convergence in Wasserstein distance in the stochastic gradient setting. Now if we take the limit as $\gamma \to \infty$ for the BAOAB and rOABAO scheme we get a contribution from the stochastic gradient of $\mathcal{O}(h^4 C_G)$ in the convergence rate estimate, and for the overdamped analysis we have a contribution of $\mathcal{O}(h^4 C_G)$ in the high friction limit of BAOAB and rOABAO. However for OBABO we get a contribution of $\mathcal{O}\big(h^2 C_G/M\big)$, which agrees with the overdamped Langevin analysis for the largest choice of stepsize.

8. Overdamped limit

Now reflecting on the contraction rates achieved in [44] we also consider the γ-limit convergent (GLC) property, *i.e.* the convergence of the integrator obtained in the $\gamma \to \infty$ limit.

8.1. BAOAB and OBABO

As originally noted in [40] the high friction limit of the BAOAB method is

$$x_{k+1} = x_k - \frac{h^2}{2}\nabla U(x_k) + \frac{h}{2}(\xi_k + \xi_{k+1}),$$

which is simply the Leimkuhler–Matthews scheme of [40] with stepsize $h^2/2$. As studied in [44] this imposes stepsize restrictions $h^2 \leq 2/M$ due to the analysis of the overdamped counterpart. The limiting contraction rate is given by

$$\lim_{\gamma\to\infty} c(h) = \frac{h^2 m}{4},$$

which agrees with contraction rate estimates for the overdamped dynamics.

Similarly for the OBABO scheme the limiting method

$$x_{k+1} = x_k - \frac{h^2}{2}\nabla U(x_k) + h\xi_{k+1},$$

is the Euler–Maruyama scheme for overdamped Langevin with stepsize $h^2/2$, and the limit contraction rate is consistent with the contraction rate for the underdamped dynamics as noted in [44].

8.2. BBK integrator

Taking the limit as $\gamma \to \infty$ and by considering two consecutive iterations ($v_{k+1} = -v_k$ in this limit) one arrives at the following update rule

$$x_{k+2} = x_k - \frac{h^2}{2}(\nabla U(x_{k+1}) + \nabla U(x_k)) + \frac{\sqrt{2\gamma}h^{3/2}}{2}(\xi_k + \xi_{k+1}),$$

and hence the method is not GLC as this does not converge to overdamped dynamics as the stepsize is taken to zero.

8.3. Stochastic position Verlet and stochastic velocity Verlet

If one takes the limit as $\gamma \to \infty$ for the stochastic position and velocity Verlet then we get the operators

$$\mathcal{V}(h) : v \to \xi,$$
$$\mathcal{A}(h) : x \to x + hv,$$

hence these schemes do not converge to the overdamped dynamics as one takes the stepsize to zero.

8.4. rOABAO

The rOABAO scheme is GLC and, interestingly, by taking the high friction limit one arrives at the scheme

$$x_{k+1} = x_k - \frac{h^2}{2}\nabla U(x_k + u\xi_k) + h\xi_k,$$

where $u \sim \mathcal{U}(0, h)$, which has the correct invariant measure and is a randomized midpoint version of the Euler–Maruyama scheme for overdamped Langevin dynamics and the one-step HMC scheme of [8].

9. Numerical experiments

To quantify and validate our convergence results and contraction rates we approximate the spectral gap of the numerical scheme $c(h)$ for an Anisotropic Gaussian example. We then compare this to the continuous dynamics *via*

$$\frac{1 - c(h)}{h},$$

which converges to the spectral gap of the continuous dynamics as $h \to 0$, and is normalized by stepsize. We also compare the bias of the numerical integrators in a Bayesian classification application.

9.1. Anisotropic Gaussian

We first consider a simple low-dimensional example to compare the convergence rates, the anisotropic Gaussian distribution on $\mathbb{R}^2$ with potential $U : \mathbb{R}^2 \mapsto \mathbb{R}$ given by $U(x, y) = \frac{1}{2}mx^2 + \frac{1}{2}My^2$. This potential satisfies Assumption 2.1 with constants M and m respectively. For this example we can analytically solve for the contraction rates, which coincide with the convergence rates of $\mathbb{E}(X_n)$. We can do this by computing the spectral gap of the transition matrix P, by $1 - |\lambda_{\max}|$, where $\lambda_{\max}$ is the largest eigenvalue of the matrix P due to Gelfand's formula. This converges to the spectral gap of the continuous dynamics as $h \to 0$.

The dependence of the convergence rate on the friction parameter γ is given in Figure 1. We will study how this changes for the discretisations with contour plots of stepsize *versus* contraction rate for all the numerical methods we consider. If we take a slice of our contour plots for small stepsizes then this will coincide with Figure 1. This is given in Figure 2.

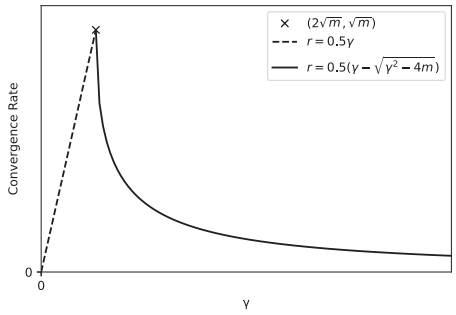

FIGURE 1. Contraction rate of continuous kinetic Langevin dynamics for an anisotropic Gaussian with parameters m and M.

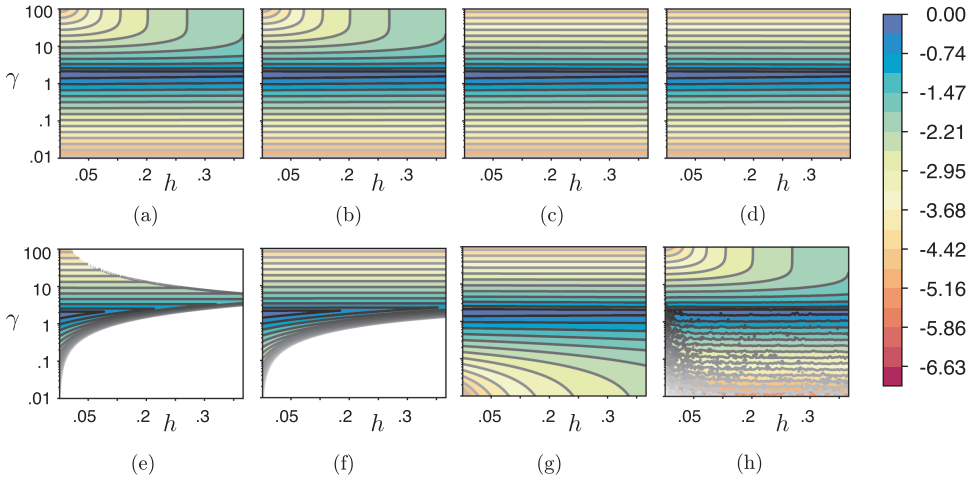

FIGURE 2. Contour plots of $\ln\left(\frac{1-c(h)}{h}\right)$ for various schemes in the case of an anisotropic Gaussian with parameters $m = 1$ and $M = 10$. Regions of white indicate instability. The rOABAO contour plot is approximate and all other plots are exact (analytic). (a) BAOAB. (b) OBABO. (c) SPV. (d) SVV. (e) EM. (f) SES/EB. (g) BBK. (h) rOABAO.

Due to the fact that each update matrix P for the anisotropic Gaussian using the rOABAO scheme is in fact a random matrix, we estimate the contraction rate using [33, 38], where

$$\lim_{N\to\infty} \log \|P_1 P_2 \dots P_N\|/N \to \log(1 - c(h)),$$

where P_i for $i \in \mathbb{N}$ is the transition matrix of the ith iteration. We approximate this limit by Monte Carlo simulations with a random $u \sim [0, h]$ from the randomized midpoint at each stage to approximate the spectral radius.

Figure 2 illustrates the exact synchronously coupled contraction rates for all the numerical integrators we consider (apart from for rOABAO, which is an approximate Monte Carlo estimation) for a range of stepsizes h and friction parameters γ. BAOAB, OBABO, rOABAO fail to approximate the true kinetic Langevin dynamics for large stepsizes, but still have low bias in the invariant measure as they act like overdamped Langevin dynamics. The γ-limit convergent property is reflected in Figure 2 for large γ as BAOAB, OBABO and rOABAO have large contraction rates for large values of the stepsize and no longer scale with $1/\gamma$, like the other schemes.

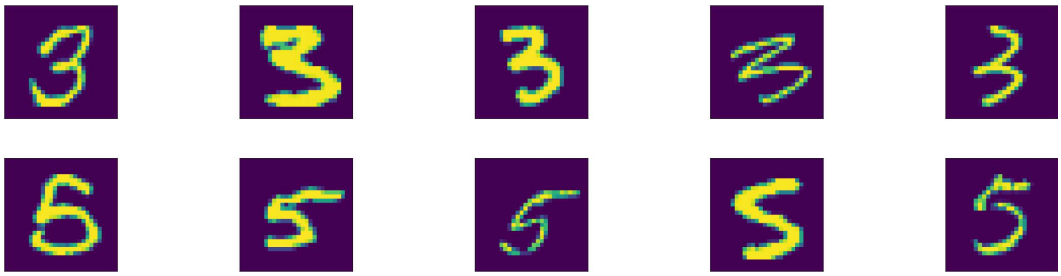

FIGURE 3. MNIST 3 and 5 digits.

SVV, SPV and BBK remain stable, but have convergence rates which scale with $1/\gamma$, indicated by the parallel contour lines in Figure 2. The SES and EM methods have large regions of instability, SES being unstable for small values of the friction parameter when h scales larger than γ.

We have only illustrated convergence results towards the invariant measure, there has been work which provides Wasserstein bias estimates for a few of the numerical methods explored (see [35,47,58]). Although the focus of this article is to provide convergence rate estimates, we will provide a comparative numerical study of the bias of each of these numerical methods for some choices of the friction parameter for an application in the following section.

9.2. Bayesian logistic regression on MNIST

We next consider a more involved example, which has a ∇-Lipschitz and convex potential. This is a Bayesian posterior sampling application in multinomial logistic regression using the MNIST machine learning data set [39]. The data set contains 60 000 training data points and 10 000 test data points. The images are of size 28 by 28 pixels and hence can be represented in $\mathbb{R}^{784}$. However, we will consider the reduced problem of classifying digits 3 and 5. Sample images are shown in Figure 3.

We use a i.i.d. Gaussian prior p_0 with mean 0 and variance $\sigma^2 = 0.001$. The likelihood function for logistic regression is

$$p(y^j \mid x^j, \mathbf{q}) = \frac{\exp\left(y^j \langle x^j, \mathbf{q}\rangle\right)}{1 + \exp\left(y^j \langle x^j, \mathbf{q}\rangle\right)},$$

where there are 2 classes (*i.e.* y^j can take values 0 and 1, with 1 corresponding to digit 5, and 0 corresponding to digit 3) and $(x^j, y^j)_{j=1}^N$ are the respective training points and labels for a data set of size N (there are $N = 11552$ training images of 3 or 5). We then define the posterior potential by

$$U(\mathbf{q}) = -\log\left(p_0(\mathbf{q})\right) - \sum_{i=1}^{N} \log\left(p(y^j \mid x^j, \mathbf{q})\right). \tag{9.1}$$

A commonly used method in machine learning and other fields is rely on a stochastic gradient approximation, an unbiased estimator of the gradient of the potential defined in (9.1). This is typically obtained based on a sub-sample of size B of a data set of size N, where $B \ll N$, *i.e.* for a random selection $I_B \subset [N] := \{1, \dots, N\}$ one would consider the gradient of

$$\widehat{\nabla_{\mathbf{q}} U(\mathbf{q})}_{SG} = -\nabla_{\mathbf{q}} \log\left(p_0(\mathbf{q})\right) - \frac{N}{B} \sum_{i \in I_B} \nabla_{\mathbf{q}} \log\left(p(y^j \mid x^j, \mathbf{q})\right), \tag{9.2}$$

where the sub-samples are chosen i.i.d at each gradient evaluation (or iteration) of the algorithm. Let $W = I_B$ as defined above, and $\mathcal{G}(\mathbf{q}, W) = \widehat{\nabla_{\mathbf{q}} U(\mathbf{q})}_{SG}$, then it is easy to see that the conditions of Definition 7.1 hold.

TABLE 4. Bias for potential function, $\gamma = \sqrt{M}$.

Algorithm	$h = 2/\sqrt{M}$ $\gamma = \sqrt{M}$	$h = 1/\sqrt{M}$ $\gamma = \sqrt{M}$	$h = 1/(2\sqrt{M})$ $\gamma = \sqrt{M}$	$h = 1/(4\sqrt{M})$ $\gamma = \sqrt{M}$
EM	4.2(±0.089)	1.5(±0.13)	0.79(±0.18)	0.28(±0.23)
BBK	2.7(±0.061)	0.67(±0.099)	0.016(±0.14)	−0.18(±0.2)
SPV	123(±0.079)	32.1(±0.091)	8.19(±0.13)	2.07(±0.18)
SVV	126(±0.097)	32.8(±0.091)	8.17(±0.13)	2.03(±0.17)
BAOAB	−0.043(±0.049)	−0.002(±0.058)	0.13(±0.086)	−0.055(±0.12)
BAOAB VRSG	0.47(±0.043)	0.23(±0.066)	0.035(±0.087)	0.036(±0.12)
OBABO	2.7(±0.056)	0.67(±0.076)	0.22(±0.13)	0.17(±0.19)
rOABAO	−2.6(±0.062)	−0.61(±0.094)	0.025(±0.13)	−0.16(±0.19)
SES/EB	2.6(±0.072)	1.2(±0.094)	0.71(±0.11)	0.2(±0.18)

One more accurate estimator for the gradient is the variance reduced stochastic gradient [37], also called the control variate method in the context of MCMC (see [2, 52]). This estimator uses the minimizer (or an approximation) $\mathbf{q}_{\min}$, and estimates the gradient as

$$\begin{aligned}\widehat{\nabla_{\mathbf{q}} U(\mathbf{q})}_{\mathrm{VRSG}} = &-\nabla_{\mathbf{q}} \log\left(p_0(\mathbf{q})\right) - \nabla_{\mathbf{q}_{\min}} \sum_{i=1}^{N} \log\left(p(y^j \mid x^j, \mathbf{q}_{\min})\right) \\ &- \frac{N}{B} \sum_{i \in I_B} \left[\nabla_{\mathbf{q}} \log\left(p(y^j \mid x^j, \mathbf{q})\right) - \nabla_{\mathbf{q}_{\min}} \log\left(p(y^j \mid x^j, \mathbf{q}_{\min})\right)\right].\end{aligned} \tag{9.3}$$

This can be also shown to satisfy Definition 7.1 with $W = I_B$ and $\mathcal{G}(\mathbf{q}, W) = \widehat{\nabla_{\mathbf{q}} U(\mathbf{q})}_{\mathrm{VRSG}}$. Both (9.2) and (9.3) are unbiased estimators of the gradient. In situations where the distribution is concentrated near the minimizer (as the sample size is large compared to the number of parameters, or the prior is sufficiently strong), the (9.3) approximation has much smaller variance, and we found that this reduces the bias of sampling algorithms. In the following numerics we first consider full gradients for each scheme. We also implemented variance reduced stochastic gradients for BAOAB, based on (9.3) with batch size $B = 100$.

We minimized the potential based on the BFGS algorithm, and computed the smallest and largest eigenvalues of the Hessian at the minimizer, which were $m = 10^3$ and $M = 1.7342 \cdot 10^5$. Note that computing the upper and lower bounds on the Hessian globally is not easy for this problem, so we used these eigenvalues at the minimizer instead for setting the parameters in our simulations. We tried two different friction parameters: $\gamma = \sqrt{M}$ (the lowest value of γ for which our theory works) and $\gamma = \sqrt{m}$ (a good choice based on the contraction rates for Gaussians shown on Fig. 2). In terms of stepsize, we tried $h \in \{2/\sqrt{M}, 1/\sqrt{M}, 1/(2\sqrt{M}), 1/(4\sqrt{M})\}$. The stepsize $h = 2/\sqrt{M}$ is near the anticipated stability threshold of these methods, this is confirmed by the fact that a larger stepsize ($h = 4/\sqrt{M}$) resulted in unstable behaviour and biases above 10^3 for all methods.

We used the potential U as a test function, which is often a good choice for examining convergence of Markov chains. The ground truth posterior mean of U was established based on running a well-tuned HMC with accept/reject steps (400 parallel runs, 440 million gradient evaluations in total, with 10% burn-in), this had a standard deviation of 0.023. The posterior standard deviation of U was also estimated based on these samples, it was found to be 19.82.

All tested methods were run in parallel 80 times for 120 000 iterations per run (20 000 burn-in, 100 000 samples), initiated from the minimum of the potential. We computed effective sample sizes based on the approach of [64], using the Matlab package `https://github.com/lacerbi/multiESS`. All methods were implemented in Matlab on a desktop computer using GPU acceleration (Tabs. 4–7).

TABLE 5. Bias for potential function, $\gamma = \sqrt{m}$.

Algorithm	$h = 2/\sqrt{M}$ $\gamma = \sqrt{m}$	$h = 1/\sqrt{M}$ $\gamma = \sqrt{m}$	$h = 1/(2\sqrt{M})$ $\gamma = \sqrt{m}$	$h = 1/(4\sqrt{M})$ $\gamma = \sqrt{m}$
EM	$6.4 \cdot 10^4(\pm 0.82)$	$1.5 \cdot 10^4(\pm 0.72)$	$1.1 \cdot 10^3(\pm 0.73)$	$4.9(\pm 0.11)$
BBK	$2.8(\pm 0.034)$	$0.68(\pm 0.041)$	$0.1(\pm 0.05)$	$0.0038(\pm 0.066)$
SPV	$0.72(\pm 0.036)$	$0.14(\pm 0.043)$	$0.06(\pm 0.054)$	$-0.014(\pm 0.073)$
SVV	$3.5(\pm 0.036)$	$0.81(\pm 0.043)$	$0.26(\pm 0.061)$	$0.05(\pm 0.089)$
BAOAB	$0.03(\pm 0.038)$	$-0.011(\pm 0.049)$	$-0.046(\pm 0.062)$	$0.043(\pm 0.074)$
BAOAB VRSG	$6.4(\pm 0.04)$	$2.4(\pm 0.051)$	$1.1(\pm 0.063)$	$0.55(\pm 0.075)$
OBABO	$2.7(\pm 0.032)$	$0.65(\pm 0.041)$	$0.22(\pm 0.052)$	$0.11(\pm 0.071)$
rOABAO	$-1.7(\pm 0.041)$	$-0.55(\pm 0.041)$	$-0.2(\pm 0.054)$	$-0.033(\pm 0.081)$
SES/EB	$6.0 \cdot 10^4(\pm 0.61)$	$1.5 \cdot 10^4(\pm 0.48)$	$1.1 \cdot 10^3(\pm 0.59)$	$4.7(\pm 0.068)$

TABLE 6. Gradient evaluations/ESS (potential function), $\gamma = \sqrt{M}$.

Algorithm	$h = 2/\sqrt{M}$ $\gamma = \sqrt{M}$	$h = 1/\sqrt{M}$ $\gamma = \sqrt{M}$	$h = 1/(2\sqrt{M})$ $\gamma = \sqrt{M}$	$h = 1/(4\sqrt{M})$ $\gamma = \sqrt{M}$
EM	$146(\pm 0.7)$	$221(\pm 0.998)$	$282(\pm 0.822)$	$327(\pm 0.581)$
BBK	$85(\pm 0.535)$	$148(\pm 0.726)$	$221(\pm 0.969)$	$285(\pm 0.933)$
SPV	$86.7(\pm 0.554)$	$148(\pm 0.775)$	$221(\pm 0.887)$	$284(\pm 0.992)$
SVV	$86.5(\pm 0.645)$	$147(\pm 0.801)$	$222(\pm 0.916)$	$283(\pm 0.825)$
BAOAB	$44.3(\pm 0.304)$	$88.7(\pm 0.585)$	$152(\pm 0.812)$	$228(\pm 0.822)$
BAOAB VRSG	$44.6(\pm 0.332)$	$86.8(\pm 0.578)$	$152(\pm 0.915)$	$226(\pm 0.934)$
OBABO	$68.6(\pm 0.491)$	$140(\pm 0.84)$	$218(\pm 0.942)$	$282(\pm 0.809)$
rOABAO	$68.5(\pm 0.507)$	$140(\pm 0.692)$	$219(\pm 0.781)$	$283(\pm 0.862)$
SES/EB	$87.4(\pm 0.593)$	$149(\pm 0.663)$	$220(\pm 0.831)$	$284(\pm 0.809)$

TABLE 7. Gradient evaluations/ESS (potential function), $\gamma = \sqrt{m}$. N.A. indicates that the method did not converge for the given stepsize.

Algorithm	$h = 2/\sqrt{M}$ $\gamma = \sqrt{m}$	$h = 1/\sqrt{M}$ $\gamma = \sqrt{m}$	$h = 1/(2\sqrt{M})$ $\gamma = \sqrt{m}$	$h = 1/(4\sqrt{M})$ $\gamma = \sqrt{m}$
EM	N.A.	N.A.	N.A.	$189(\pm 0.955)$
BBK	$15(\pm 0.124)$	$30.1(\pm 0.233)$	$57.5(\pm 0.352)$	$108(\pm 0.717)$
SPV	$15.1(\pm 0.106)$	$29.7(\pm 0.209)$	$57.4(\pm 0.408)$	$109(\pm 0.725)$
SVV	$15(\pm 0.121)$	$29.9(\pm 0.222)$	$57.5(\pm 0.341)$	$108(\pm 0.628)$
BAOAB	$18.8(\pm 0.128)$	$36.4(\pm 0.288)$	$66.4(\pm 0.461)$	$116(\pm 0.849)$
BAOAB VRSG	$19.7(\pm 0.169)$	$36.4(\pm 0.242)$	$67.8(\pm 0.447)$	$114(\pm 0.662)$
OBABO	$15(\pm 0.118)$	$30(\pm 0.204)$	$57.5(\pm 0.471)$	$108(\pm 0.711)$
rOABAO	$16.5(\pm 0.236)$	$29.7(\pm 0.218)$	$58.2(\pm 0.356)$	$109(\pm 0.669)$
SES/EB	N.A.	N.A.	N.A.	$108(\pm 0.652)$

Firstly, when changing from $\gamma = \sqrt{M}$ to $\gamma = \sqrt{m}$, we can see that the changes in bias are not significant for BAOAB, OBABO, rOABAO, and BBK, the bias increases significantly for EM and SES (instability issues) and somewhat for BAOAB VRSG, and the bias decreases significantly for SPV and SVV. In terms of gradient evaluations per ESS, the choice $\gamma = \sqrt{m}$ is more efficient by a factor of 2–6 for all methods except EM and SES. This is in line with the recent research in accelerated convergence rates for underdamped Langevin dynamics [15, 68].

We can see that BAOAB has impressively low bias for the potential test function even at the largest stepsize $2/\sqrt{M}$, and it also has a competitive computational cost in terms of gradient evaluations/effective sample size (ESS). The VRSG variant of BAOAB has somewhat larger bias (especially at lower frictions), but it requires a similar number of iterations per ESS, with much lower computational cost per iteration compared to using full gradients. The rOABAO scheme based on randomized midpoints has a relatively low bias at all stepsizes, and requires a rather small number of gradient evaluations per iteration. It is beyond the scope of this paper, but we think that more significant differences could arise between these schemes for less smooth potentials.

10. Conclusion

In this article we have extended the results of [44] to further integration schemes. By building stepsize-dependent norms we achieve convergence rates which hold on a large interval of stepsize, in many cases the same as the stability threshold of the numerical method (up to a constant factor). We further considered the case of stochastic gradients, where we allow a flexible choice of unbiased gradient estimator under the assumption that the expected variance of the Jacobian of the estimator is bounded. We show that this results in a reduced convergence rate based on the variance of the Jacobian of the estimator, which coincides with what we have observed numerically for small batch sizes in a subsampled stochastic gradient. Most previous results in the literature (see *e.g.* [22]) require the mean square error of the stochastic gradients $\mathbb{E}(\|G_k - \nabla U(x_k)\|^2)$ to be uniformly bounded, which can be easily violated even for strongly convex and gradient-Lipschitz potentials U_i when using standard subsampling schemes. We do not need such a stringent requirement; our conditions on the stochastic gradients stated in Assumption 7.2 are applicable for subsampling-based estimators as long as each U_i is gradient-Lipschitz.

We have provided numerical results comparing the bias of each of the numerical methods based on choices of the friction parameter which are optimal according to our theory or the optimal choice for the Gaussian distribution, where we solved for the convergence rates exactly. We compared the errors of the integrators in a Bayesian logistic regression application and have seen that some of the integrators performed well with large stepsizes, even in the presence of stochastic gradients.

Our theoretical and numerical results indicate that using stochastic gradients with advanced numerical integrators can perform well and have significant computational advantages compared to full-gradient methods. In the case of sufficiently large batch sizes, there is little change in terms of the stability threshold and convergence rate compared to the full-gradient version.

Appendix A. Stochastic gradient kinetic Langevin dynamics integrators

For the Euler–Maruyama, stochastic Euler scheme, rOABAO, stochastic position Verlet only one force evaluation is used in each iteration, so every gradient evaluation is taken as a stochastic gradient estimate. The complete algorithms are stated below in Algorithms 1–4.

Algorithm 1: Stochastic gradient Euler–Maruyama (EM).

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- for $k = 1, 2, \ldots, K$ do
  - Sample $W_k \sim \rho$
  - $G_{k-1} \to \mathcal{G}(x_{k-1}, W_k)$
  - Sample $\xi_k \sim \mathcal{N}(0_d, I_d)$
  - $x_k \to x_{k-1} + h v_{k-1}$
  - $v_k \to v_{k-1} - h G_{k-1} - h\gamma v_{k-1} + \sqrt{2\gamma h}\xi_k$
- Output: Samples $(x_k)_{k=0}^K$.

Algorithm 2: Stochastic gradient stochastic Euler scheme (SES/EB).

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- for $k = 1, 2, \ldots, K$ do
  - Sample $(\zeta_k, \omega_k) \sim \mathcal{N}(0_d, \Sigma)$, where Σ is given in (4.4).
  - Sample $W_k \sim \rho$
  - $G_{k-1} \to \mathcal{G}(x_{k-1}, W_k)$
  - $x_k \to x_{k-1} + \frac{1-\eta}{\gamma} v_{k-1} - \frac{\gamma h + \eta - 1}{\gamma^2} G_{k-1} + \zeta_k$
  - $v_k \to \eta v_{k-1} - \frac{1-\eta}{\gamma} G + \omega_k$
- Output: Samples $(x_k)_{k=0}^K$.

Algorithm 3: Stochastic gradient rOABAO.

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- for $k = 1, 2, \ldots, K$ do
  - Sample $u_k \sim \mathcal{U}(0, h)$
  - Sample $W_k \sim \rho$
  - $G_{k-1} \to \mathcal{G}(x_{k-1} + u_k v_{k-1}, W_k)$
  - Sample $\xi_k \sim \mathcal{N}(0_d, I_d)$
  - (O) $v \to \eta^{1/2} v_{k-1} + \sqrt{1-\eta}\xi_k$
  - $x_k \to x_{k-1} + h v_{k-1} - \frac{h^2}{2} G_{k-1}$
  - $v_k \to v_{k-1} - h G_{k-1}$
  - Sample $\xi_k' \sim \mathcal{N}(0_d, I_d)$
  - (O) $v_k \to \eta^{1/2} v + \sqrt{1-\eta}\xi_k'$
- Output: Samples $(x_k)_{k=0}^K$.

Algorithm 4: Stochastic gradient stochastic velocity Verlet (SVV).

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- for $k = 1, 2, \ldots, K$ do
  - (A) $x \to x_{k-1} + \frac{h}{2} v_{k-1}$
  - Sample $W_k \sim \rho$
  - $G_{k-1} \to \mathcal{G}(x, W_k)$
  - Sample $\xi_k \sim \mathcal{N}(0_d, I_d)$
  - ($\mathcal{V}$) $v_k \to \eta v_{k-1} - \frac{1-\eta}{\gamma} G_{k-1} + \sqrt{1-\eta^2}\xi_k$
  - (A) $x_k \to x + \frac{h}{2} v_k$
- Output: Samples $(x_k)_{k=0}^K$.

For BAOAB the first and last $\mathcal{B}$ of each iteration share a stochastic gradient evaluation to make the algorithm roughly one gradient evaluation per step. The complete algorithm is given in Algorithm 5.

Algorithm 5: Stochastic gradient BAOAB.

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- Sample $W_1 \sim \rho$
- $G_0 \to \mathcal{G}(x_0, W_1)$
- for $k = 1, 2, \ldots, K$ do
  - (B) $v \to v_{k-1} - \frac{h}{2} G_{k-1}$
  - (A) $x \to x_{k-1} + \frac{h}{2} v$
    Sample $\xi_k \sim \mathcal{N}(0_d, I_d)$
  - (O) $v \to \eta v + \sqrt{1-\eta^2}\xi_k$
  - (A) $x_k \to x + \frac{h}{2} v$
    Sample $W_{k+1} \sim \rho$
    $G_k \to \mathcal{G}(x_k, W_{k+1})$
  - (B) $v_k \to v - \frac{h}{2} G_k$
- Output: Samples $(x_k)_{k=0}^K$.

Similarly for OBABO the first and last $\mathcal{B}$ of each iteration share a stochastic gradient evaluation to make the algorithm roughly one gradient evaluation per step. The complete algorithm is given in Algorithm 6.

Algorithm 6: Stochastic gradient OBABO.

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- Sample $W_1 \sim \rho$
- $G_0 \to \mathcal{G}(x_0, W_1)$
- for $k = 1, 2, \ldots, K$ do
    Sample $\xi_k \sim \mathcal{N}(0_d, I_d)$
  - (O) $v \to \eta^{1/2} v_{k-1} + \sqrt{1-\eta}\xi_k$
  - (B) $x \to x_{k-1} - \frac{h}{2} G_{k-1}$
  - (A) $x \to x + hv$
    Sample $W_{k+1} \sim \rho$
    $G_k \to \mathcal{G}(x, W_{k+1})$
  - (B) $x_k \to x - \frac{h}{2} G_k$
    Sample $\xi'_k \sim \mathcal{N}(0_d, I_d)$
  - (O) $v_k \to \eta^{1/2} v + \sqrt{1-\eta}\xi'_k$
- Output: Samples $(x_k)_{k=0}^K$.

If we express each iteration of the BBK methods as $\Phi_{\mathrm{B}_2} \circ \Phi_{\mathrm{A}} \circ \Phi_{\mathrm{B}_1}$ as in Section 6, then the last Φ_{B_2} step of each iteration and the first Φ_{B_1} of the next iteration share the same stochastic gradient evaluation to make the algorithm roughly one gradient evaluation per step. The complete algorithm is given in Algorithm 7.

Algorithm 7: Stochastic gradient Brunger–Brooks–Karplus (BBK).

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- Sample $W_1 \sim \rho$
- $G_0 \to \mathcal{G}(x_0, W_1)$
- for $k = 1, 2, \ldots, K$ do
    Sample $\xi_k \sim \mathcal{N}(0_d, I_d)$
  - (B_1) $v \to v_{k-1} + \frac{h}{2}\left(-G_{k-1} - \gamma v_{k-1} + \frac{\sqrt{2\gamma}}{\sqrt{h}}\xi_k\right)$
  - (A) $x_k \to x_{k-1} + hv$
    Sample $W_{k+1} \sim \rho$
    $G_k \to \mathcal{G}(x_k, W_{k+1})$
    Sample $\xi'_k \sim \mathcal{N}(0_d, I_d)$
  - (B_2) $v_k \to \left(1 + \frac{h}{2}\gamma\right)^{-1}\left(v - \frac{h}{2} G_k + \frac{\sqrt{2\gamma}}{\sqrt{h}}\xi'_k\right)$
- Output: Samples $(x_k)_{k=0}^K$.

Finally for SVV the last $\mathcal{V}$ step of each iteration and the first $\mathcal{V}$ of the next iteration share the same stochastic gradient evaluation. The complete algorithm is given in Algorithm 8.

Algorithm 8: Stochastic gradient stochastic velocity Verlet (SVV).

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- Sample $W_1 \sim \rho$
- $G_0 \to \mathcal{G}(x_0, W_1)$
- for $k = 1, 2, \ldots, K$ do
  - Sample $\xi_k \sim \mathcal{N}(0_d, I_d)$
  - $(\mathcal{V})$ $v \to \eta^{1/2} v_{k-1} - \frac{1-\eta^{1/2}}{\gamma} G_{k-1} + \sqrt{1-\eta}\xi_k$
  - (A) $x_k \to x_{k-1} + hv$
    $G_k \to \mathcal{G}(x_k, W_{k+1})$
    Sample $\xi'_k \sim \mathcal{N}(0_d, I_d)$
  - $(\mathcal{V})$ $v_k \to \eta^{1/2} v - \frac{1-\eta^{1/2}}{\gamma} G_k + \sqrt{1-\eta}\xi'_k$
- Output: Samples $(x_k)_{k=0}^K$.

Acknowledgements

The authors acknowledge the support of the Engineering and Physical Sciences Research Council Grant EP/S023291/1 (MAC-MIGS Centre for Doctoral Training). The authors thank Sinho Chewi for the valuable discussion about acceleration in kinetic Langevin dynamics.

References

[1] A. Abdulle, G. Vilmart and K.C. Zygalakis, Long time accuracy of Lie–Trotter splitting methods for Langevin dynamics. *SIAM J. Numer. Anal.* **53** (2015) 1–16.

[2] J. Baker, P. Fearnhead, E.B. Fox and C. Nemeth, Control variates for stochastic gradient MCMC. *Stat. Comput.* **29** (2019) 599–615.

[3] J. Besag, Discussion: Markov chains for exploring posterior distributions. *Ann. Stat.* **22** (1994) 1734–1741.

[4] J. Bierkens, P. Fearnhead and G. Roberts, The zig–zag process and super-efficient sampling for Bayesian analysis of big data. *Ann. Stat.* **47** (2019) 1288–1320.

[5] S.D. Bond and B.J. Leimkuhler, Molecular dynamics and the accuracy of numerically computed averages. *Acta Numer.* **16** (2007) 1–65.

[6] N. Bou-Rabee and A. Eberle, Couplings for Andersen dynamics, in Annales de l'Institut Henri Poincare (B) Probabilites et statistiques. Vol. 58. Institut Henri Poincaré (2022) 916–944.

[7] N. Bou-Rabee and A. Eberle, Mixing time guarantees for unadjusted Hamiltonian Monte Carlo. *Bernoulli* **29** (2023) 75–104.

[8] N. Bou-Rabee and M. Marsden, Unadjusted Hamiltonian MCMC with stratified Monte Carlo time integration. Preprint arXiv:2211.11003 (2022).

[9] N. Bou-Rabee, A. Eberle and R. Zimmer, Coupling and convergence for Hamiltonian Monte Carlo. *Ann. Appl. Probab.* **30** (2020) 1209–1250.

[10] A. Bouchard-Côté, S.J. Vollmer and A. Doucet, The bouncy particle sampler: a nonreversible rejection-free Markov chain Monte Carlo method. *J. Am. Stat. Assoc.* **113** (2018) 855–867.

[11] S. Boyd, S.P. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press (2004).

[12] A. Brünger, C.L. Brooks III and M. Karplus, Stochastic boundary conditions for molecular dynamics simulations of ST2 water. *Chem. Phys. Lett.* **105** (1984) 495–500.

[13] G. Bussi and M. Parrinello, Accurate sampling using Langevin dynamics. *Phys. Rev. E* **75** (2007) 056707.

[14] Y. Cao, J. Lu and L. Wang, Complexity of randomized algorithms for underdamped Langevin dynamics. *Commun. Math. Sci.* **19** (2021) 1827–1853.

[15] Y. Cao, J. Lu and L. Wang, On explicit l^2-convergence rate estimate for underdamped Langevin dynamics. *Arch. Ration. Mech. Anal.* **247** (2023) 90.

[16] S. Chandrasekhar, Stochastic problems in physics and astronomy. *Rev. Mod. Phys.* **15** (1943) 1.

[17] N. Chatterji, N. Flammarion, Y. Ma, P. Bartlett and M. Jordan, On the theory of variance reduction for stochastic gradient Monte Carlo, in International Conference on Machine Learning. PMLR (2018) 764–773.

[18] X. Cheng and P. Bartlett, Convergence of Langevin MCMC in KL-divergence, in Algorithmic Learning Theory. PMLR (2018) 186–211.

[19] X. Cheng, N.S. Chatterji, P.L. Bartlett and M.I. Jordan, Underdamped Langevin MCMC: a non-asymptotic analysis, in Conference on Learning Theory. PMLR (2018) 300–323.

[20] A. Dalalyan, Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent, in Conference on Learning Theory. PMLR (2017) 678–689.

[21] A.S. Dalalyan, Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **79** (2017) 651–676.

[22] A.S. Dalalyan and A. Karagulyan, User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Process. App.* **129** (2019) 5278–5311.

[23] A.S. Dalalyan and L. Riou-Durand, On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli* **26** (2020) 1956–1988.

[24] G. Deligiannidis, D. Paulin, A. Bouchard-Côté and A. Doucet, Randomized Hamiltonian Monte Carlo as scaling limit of the bouncy particle sampler and dimension-free convergence rates. *Ann. Appl. Probab.* **31** (2021) 2612–2662.

[25] A. Durmus and E. Moulines, Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.* **27** (2017) 1551–1587.

[26] A. Durmus and E. Moulines, High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli* **25** (2019) 2854–2882.

[27] A. Durmus, S. Majewski and B. Miasojedow, Analysis of Langevin Monte Carlo via convex optimization. *J. Mach. Learn. Res.* **20** (2019) 2666–2711.

[28] A. Durmus, A. Enfroy, É. Moulines and G. Stoltz, Uniform minorization condition and convergence bounds for discretizations of kinetic Langevin dynamics. Preprint arXiv:2107.14542 (2021).

[29] R. Dwivedi, Y. Chen, M.J. Wainwright and B. Yu, Log-concave sampling: Metropolis-hastings algorithms are fast! in Conference on Learning Theory. PMLR (2018) 793–797.

[30] A. Eberle, A. Guillin and R. Zimmer, Couplings and quantitative contraction rates for Langevin dynamics. *Ann. Probab.* **47** (2019) 1982–2010.

[31] D.L. Ermak and H. Buckholz, Numerical integration of the Langevin equation: Monte Carlo simulation. *J. Comput. Phys.* **35** (1980) 169–182.

[32] J. Finkelstein, G. Fiorin and B. Seibold, Comparison of modern Langevin integrators for simulations of coarse-grained polymer melts. *Mol. Phys.* **118** (2020) e1649493.

[33] H. Furstenberg and H. Kesten, Products of random matrices. *Ann. Math. Stat.* **31** (1960) 457–469.

[34] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari and D.B. Rubin, Bayesian Data Analysis. CRC Press (2013).

[35] N. Gouraud, P. Le Bris, A. Majka and P. Monmarché, HMC and underdamped Langevin united in the unadjusted convex smooth case. Preprint arXiv:2202.00977 (2022).

[36] D. Griffeath, A maximal coupling for Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **31** (1975) 95–106.

[37] R. Johnson and T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction. *Adv. Neural Inf. Process. Syst.* **26** (2013).

[38] V. Kargin, Products of random matrices: dimension and growth in norm. *Ann. Appl. Probab.* **23** (2010) 890–906.

[39] Y. LeCun, C. Cortes, C. Burges, *et al.*, MNIST handwritten digit database (2010).

[40] B. Leimkuhler and C. Matthews, Rational construction of stochastic numerical methods for molecular sampling. *Appl. Math./Res. eXpress* **2013** (2013) 34–56.

[41] B. Leimkuhler and C. Matthews, Molecular Dynamics. Springer (2015).

[42] B. Leimkuhler, C. Matthews and M.V. Tretyakov, On the long-time integration of stochastic gradient systems. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **470** (2014) 20140120.

[43] B. Leimkuhler, C. Matthews and G. Stoltz, The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics. *IMA J. Numer. Anal.* **36** (2016) 13–79.

[44] B. Leimkuhler, D. Paulin and P.A. Whalley, Contraction and convergence rates for discretized kinetic Langevin dynamics. *SIAM J. Numer. Anal.* **62** (2024) 1226–1258.

[45] M.B. Majka, A. Mijatović and Ł. Szpruch, Nonasymptotic bounds for sampling algorithms without log-concavity. *Ann. Appl. Probab.* **30** (2020) 1534–1581.

[46] S. Melchionna, Design of quasisymplectic propagators for Langevin dynamics. *J. Chem. Phys.* **127** (2007) 044108.

[47] P. Monmarché, High-dimensional MCMC with a standard splitting scheme for the underdamped Langevin diffusion. *Electron. J. Stat.* **15** (2021) 4117–4166.

[48] P. Monmarché, Almost sure contraction for diffusions on $\mathbb{R}^d$. Application to generalized Langevin diffusions *Stoch. Process. App.* **161** (2023) 316–349.

[49] C. Nemeth and P. Fearnhead, Stochastic gradient Markov chain Monte Carlo. *J. Am. Stat. Assoc.* **116** (2021) 433–450.

[50] G.A. Pavliotis, Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations. Vol. 60. Springer (2014).

[51] E.A.J.F. Peters and G. de With, Rejection-free Monte Carlo sampling for general potentials. *Phys. Rev. E* **85** (2012) 026703.

[52] M. Quiroz, R. Kohn, M. Villani and M.-N. Tran, Speeding up MCMC by efficient data subsampling. *J. Am. Stat. Assoc.* **114** (2019) 831–843.

[53] L. Riou-Durand and J. Vogrinc, Metropolis adjusted Langevin trajectories: a robust alternative to Hamiltonian Monte Carlo. Preprint arXiv:2202.13230 (2022).

[54] H. Robbins and S. Monro, A stochastic approximation method. *Ann. Math. Stat.* **22** (1951) 400–407.

[55] G.O. Roberts and R.L. Tweedie, Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2** (1996) 341–363.

[56] P.J. Rossky, J.D. Doll and H.L. Friedman, Brownian dynamics as smart Monte Carlo simulation. *J. Chem. Phys.* **69** (1978) 4628–4633.

[57] J.M. Sanz Serna and K.C. Zygalakis, Contractivity of Runge–Kutta methods for convex gradient systems. *SIAM J. Numer. Anal.* **58** (2020) 2079–2092.

[58] J.M. Sanz-Serna and K.C. Zygalakis, Wasserstein distance estimates for the distributions of numerical approximations to ergodic stochastic differential equations. *J. Mach. Learn. Res.* **22** (2021) 1–37.

[59] K. Schuh, Global contractivity for Langevin dynamics with distribution-dependent forces and uniform in time propagation of Chaos. Annales de l'Institut Henri Poincare (B) Probabilites et statistiques. Vol. 60. Institut Henri Poincaré (2024) 753–789.

[60] I. Sekkat and G. Stoltz, Quantifying the mini-batching error in Bayesian inference for adaptive Langevin dynamics. *J. Mach. Learn. Res.* **24** (2023) 58.

[61] R. Shen and Y.T. Lee, The randomized midpoint method for log-concave sampling. *Adv. Neural Inf. Process. Syst.* **32** (2019).

[62] R.D. Skeel and J.A. Izaguirre, An impulse integrator for Langevin dynamics. *Mol. Phys.* **100** (2002) 3885–3891.

[63] L. Vaserstein, Markovian processes on countable space product describing large systems of automata. *Probl. Peredachi Inf.* **5** (1969) 64–72.

[64] D. Vats, J.M. Flegal and G.L. Jones, Multivariate output analysis for Markov chain Monte Carlo. *Biometrika* **106** (2019) 321–337.

[65] C. Villani, Optimal Transport: Old and New. Vol. 338. Springer (2009).

[66] S.J. Vollmer, K.C. Zygalakis and Y.W. Teh, Exploration of the (non-) asymptotic bias and variance of stochastic gradient Langevin dynamics. *J. Mach. Learn. Res.* **17** (2016) 5504–5548.

[67] M. Welling and Y.W. Teh, Bayesian learning via stochastic gradient Langevin dynamics, in Proceedings of the 28th International Conference on Machine Learning (ICML-11) (2011) 681–688.

[68] M. Zhang, S. Chewi, M.B. Li, K. Balasubramanian and M.A. Erdogdu, Improved discretization analysis for underdamped Langevin Monte Carlo. The Thirty Sixth Annual Conference on Learning Theory. PMLR (2023) 36–71.