Adaptive thermostats for noisy gradient systems

Leimkuhler, Benedict; Shang, Xiaocheng

DOI:
10.1137/15m102318x

Document Version
Peer reviewed version

Citation for published version (Harvard):
Leimkuhler, B & Shang, X 2016, 'Adaptive thermostats for noisy gradient systems', *SIAM Journal on Scientific Computing*, vol. 38, no. 2, pp. A712-A736. https://doi.org/10.1137/15m102318x

ADAPTIVE THERMOSTATS FOR NOISY GRADIENT SYSTEMS

BENEDICT LEIMKUHLER[†§] AND XIAOCHENG SHANG[‡§]

Abstract. We study numerical methods for sampling probability measures in high dimension where the underlying model is only approximately identified with a gradient system. Extended stochastic dynamical methods are discussed which have application to multiscale models, nonequilibrium molecular dynamics and Bayesian sampling techniques arising in emerging machine learning applications. In addition to providing a more comprehensive discussion of the foundations of these methods, we propose a new numerical method for the Adaptive Langevin/Stochastic Gradient Nosé-Hoover Thermostat that achieves a dramatic improvement in numerical efficiency over the most popular stochastic gradient methods reported in the literature. We also demonstrate that the newly-established method inherits a superconvergence property (fourth order convergence to the invariant measure for configurational quantities) recently demonstrated in the setting of Langevin dynamics. Our findings are verified by numerical experiments.

Key words. stochastic differential equations, adaptive thermostat, Bayesian sampling, machine learning, invariant measure, ergodicity

AMS subject classifications. 65C30, 60H35, 37M25

1. Introduction.

Stochastic thermostats [37, 55, 56] are powerful tools for sampling probability measures on high dimensional spaces. These methods combine an extended dynamics with degenerate stochastic perturbation to ensure ergodicity. The traditional use of thermostats in molecular dynamics is to sample a well specified equilibrium system involving a known force field which is the gradient of a potential energy function. Recently, however, these techniques have become increasingly popular for problems of more general form, including the following:

- multiscale models in which the forces are obtained by approximate sampling in another scale regime [17, 21, 40, 44, 48, 49];
- nonequilibrium physical models in which the potential energy function is either evolving or does not completely specify the system [27, 41, 45, 47, 58, 59];
- Bayesian machine learning applications in which a dataset defines an objective function which leads to an effective force law [3, 11, 14, 46, 57, 62, 63].

In this article, we consider thermostats and numerical methods for sampling an underlying probability measure in the presence of error, under the assumption that the errors are random with a simple distributional form and unknown, but constant or slowly varying, parameters. In the cases considered, these methods are simple to implement, robust, and efficient.

1.1. Thermostats.

The main tool that we employ in this article is the general concept of a thermostat as a (stochastic) distributional control for a dynamical system. These methods originate in molecular dynamics and it is simplest to explain them in that context. Classical molecular dynamics (MD) tracks the motion of individual atoms determined by Newton's law in the microcanonical (NVE) ensemble, where energy (i.e. the Hamiltonian of the system) is always conserved [4, 18, 20, 38]. However constant energy is not the appropriate setting of a real-world laboratory environment. In most cases, one wishes instead to sample the canonical (NVT) ensemble, where temperature, as an intensive variable, is conserved, by using thermostat techniques [18, 25].

The idea of a thermostat is to modify dynamics so that a prescribed invariant measure is sampled. There are competing aims in this type of work. For example, one may wish to perturb the underlying Newtonian dynamics minimally, so that temporal correlations are preserved, or one may be interested in sampling rare events in a system with metastable states, thus a variety of methods have been developed. The most obvious proposals, and also the oldest, are Brownian and Langevin dynamics.

[†]School of Mathematics and Maxwell Institute for Mathematical Sciences, James Clerk Maxwell Building, The King's Buildings, University of Edinburgh, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, UK (b.leimkuhler@ed.ac.uk).

[‡]Corresponding author. This author's work was supported by the University of Edinburgh and China Scholarship Council (x.shang@ed.ac.uk).

[§]Address in common for the first and second authors.

In Brownian (sometimes called "overdamped Langevin") dynamics, the system is

$$\mathrm{d}\mathbf{q} = -\lambda \nabla U(\mathbf{q})\mathrm{d}t + \sqrt{2\beta^{-1}\lambda}\mathrm{d}\mathbf{W}\,, \tag{1.1}$$

where $\mathbf{q}$ represents a $3N$-dimensional vector of time-dependent random variables, $\mathrm{d}\mathbf{W}$ represents a vector of infinitesimal Wiener increments, β is a positive parameter (proportional to the reciprocal temperature), U is the potential energy function and λ is a free parameter which represents a time-rescaling. It can be shown [9] that this system (1.1) ergodically samples the Gibbs-Boltzmann probability distribution $\bar{\rho}_\beta \propto \exp(-\beta U)$. For simplicity, we assume that the configurations $\mathbf{q}$ are restricted to a compact and simply connected domain $\Omega_{\mathbf{q}}$. In molecular dynamics applications, the starting point is the potential energy function which is usually assumed to be a semi-empirical formula constructed from primitive functions via an a priori parameter fitting procedure. Alternatively, one may assume that it is the probability distribution that is specified and that the potential energy is constructed from it via

$$U = -\beta^{-1}\ln\rho\,,$$

which of course requires that $\rho > 0$. In many applications it is found that the use of a first order dynamics such as (1.1) is inefficient or introduces unphysical dynamical properties and one employs, instead, the Langevin dynamics method:

$$\mathrm{d}\mathbf{q} = \mathbf{M}^{-1}\mathbf{p}\mathrm{d}t\,, \tag{1.2}$$

$$\mathrm{d}\mathbf{p} = -\nabla U(\mathbf{q})\mathrm{d}t - \gamma\mathbf{p}\mathrm{d}t + \sqrt{2\beta^{-1}\gamma}\mathbf{M}^{1/2}\mathrm{d}\mathbf{W}\,. \tag{1.3}$$

Again, γ in these equations is a free parameter, termed the "friction constant". It is related to the timescale on which the variables of the system interact with particles of a fictitious extended "bath", but it cannot be associated with a simple time-rescaling of the equations of motion and is thus different from λ in (1.1). It is a little more involved to show that (1.2)-(1.3) ergodically [42] samples the distribution with density $\rho_\beta \propto \exp(-\beta H(\mathbf{q},\mathbf{p}))$, where $H(\mathbf{q},\mathbf{p}) = \mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}/2 + U(\mathbf{q})$. In molecular dynamics, the $3N \times 3N$ matrix $\mathbf{M}$ is typically diagonal and contains the masses of atoms, $\mathbf{p}$ represents the momentum vector, and H is the Hamiltonian or energy function. In more general settings, the masses and friction coefficient may be treated as free parameters and by computing long trajectories of (1.2)-(1.3) one may obtain averages with respect to $\bar{\rho}_\beta(\mathbf{q})$, i.e., if $\{(\mathbf{q}(\tau),\mathbf{p}(\tau)) : \tau \geq 0\}$ is a path generated by solving the SDE system (1.2)-(1.3), one has, for suitable test functions $\phi(\mathbf{q})$, and under certain conditions on the potential energy function U [42],

$$\lim_{\tau\to\infty} \tau^{-1}\int_0^\tau \phi(\mathbf{q}(\tau))\,\mathrm{d}\tau = \int_{\Omega_{\mathbf{q}}} \phi(\mathbf{q})\bar{\rho}_\beta(\mathbf{q})\,\mathrm{d}\omega_{\mathbf{q}}\,,$$

where $\mathrm{d}\omega_{\mathbf{q}} = \mathrm{d}\mathbf{q}_1\mathrm{d}\mathbf{q}_2\ldots\mathrm{d}\mathbf{q}_N$. In other words, the projected path defines a sampler for the density $\bar{\rho}_\beta$.

Langevin dynamics can thus be seen as an extended system which allows sampling to be performed in a reduced cross section of phase space by marginalization over long trajectories; this is the essential property of a thermostat. Other types of thermostats include Nosé-Hoover-Langevin (NHL) dynamics [37,55] and various generalized schemes (see e.g. [32]). In these methods, one adds additional auxiliary variables which are meant to control the dynamics (via a negative feedback loop), and the auxiliary variables are then further coupled to stochastic processes of Ornstein-Uhlenbeck type which can provide ergodicity [37]. (Note that the use of purely deterministic approaches, such as Nosé-Hoover, results in ergodicity issues [30, 31].) The use of auxiliary variables can provide a degree of flexibility in the design of the thermostat, for example allowing the treatment of systems arising in fluid dynamics [16], or imposing an isokinetic constraint [33]. Very recently, we have further generalized the NHL method to obtain Pairwise Nosé-Hoover-Langevin (PNHL), which is a momentum-conserving thermostat and thus applicable to the simulation of hydrodynamic behavior in complex fluids and polymers in mesoscales [39].

1.2. Noisy Gradients. The gradient (or Hamiltonian) structure is essential to the nature of all the methods described above since it is only by use of this feature that the underlying Fokker-Planck equation can be shown to have the desired steady state solution. However, in many applications, in particular multiscale modelling, the force is corrupted by significant approximation error and cannot be viewed as the gradient of a single global potential function. One imagines a large extended system involving configurational variables $\mathbf{q}$ and $\mathbf{y}$, with $(\mathbf{q},\mathbf{y}) \in \Omega_{\mathbf{q}} \times \Omega_{\mathbf{y}}$ (compact), and an overall distribution described by a Gibbs-Boltzmann density

$$\tilde{\rho}(\mathbf{q},\mathbf{y}) = Z^{-1}\exp\left(-\beta\tilde{U}(\mathbf{q},\mathbf{y})\right),$$

where Z is a normalizing constant so that $\tilde{\rho}$ is a probability density. One calculates the mean force acting on $\mathbf{q}$, $\hat{f}(\mathbf{q})$, by averaging the forces in the extended Gibbsian system, $\tilde{f}(\mathbf{q},\mathbf{y})$, as

$$\hat{f}(\mathbf{q}) = \int_{\Omega_{\mathbf{y}}} \tilde{f}(\mathbf{q},\mathbf{y})\tilde{\rho}(\mathbf{q},\mathbf{y})\,\mathrm{d}\omega_{\mathbf{y}}\,.$$

If, as would typically be assumed, $\tilde{f}(\mathbf{q},\mathbf{y}) = -\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q},\mathbf{y})$, i.e. the force in the extended system is conservative, then we may interpret $\hat{f}$ as a conservative force, as well, specifically the gradient of the potential of mean force, which is given by

$$\hat{U}(\mathbf{q}) = -\beta^{-1}\ln\int_{\Omega_{\mathbf{y}}}\exp\left(-\beta\tilde{U}(\mathbf{q},\mathbf{y})\right)\,\mathrm{d}\omega_{\mathbf{y}}\,.$$

The challenge arises when this integral must be approximated. For example, if this is done by Monte Carlo integration, for fixed $\mathbf{q}$, one generates samples $\mathbf{y}^1,\mathbf{y}^2,\ldots,\mathbf{y}^k$ from the distribution with density $\tilde{\rho}(\mathbf{q},\mathbf{y})$ and thus approximates the mean force by

$$\bar{f}^k(\mathbf{q}) = k^{-1}\sum_{i=1}^{k}\tilde{f}\left(\mathbf{q},\mathbf{y}^i\right)\,.$$

In practice most systems constructed in this way, for example those arising in mixed quantum and classical molecular models [7], will admit very substantial errors in the forces, that is

$$\bar{f}^k(\mathbf{q}) = \hat{f}(\mathbf{q}) + \Delta^k(\mathbf{q})\,.$$

Depending on the method of computation, it may be reasonable to assume that the errors Δ^k are normally distributed with zero mean, which is justified by the central limit theorem [5], but the variance of the errors is generally not known and will be dependent on the location $\mathbf{q}$ where they are computed, thus we would expect

$$\Delta^k(\mathbf{q}) \sim \mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}^k(\mathbf{q})\right)\,, \tag{1.4}$$

where $\boldsymbol{\Sigma}^k(\mathbf{q})$ is an unknown covariance matrix. It should be noted that the assumption of the errors being Gaussian distributed is also often adopted in Bayesian inverse problems [12] and elsewhere.

The most straightforward approach to the problem is to first treat the estimation problem for $\boldsymbol{\Sigma}^k$ separately, by some means, and then to use this within a standard Brownian or Langevin dynamics algorithm. The difficulty is that this requires a high level of local accuracy in the calculations, which is likely to be burdensome and involve redundant computation. What we would prefer to do is to resolve the correct target distribution by a global calculation.

This problem has recently been encountered in the data science community where it has attracted considerable attention [3, 11, 14, 46, 57, 62, 63]. To illustrate, we consider the problem of Bayesian sampling [8,51], where one is interested in correctly drawing states from a posterior probability density defined as

$$\pi(\boldsymbol{\theta}|\mathbf{X}) \propto \pi(\mathbf{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,, \tag{1.5}$$

where $\boldsymbol{\theta}$ is the parameter vector of interest, $\mathbf{X}$ represents the entire dataset, and, $\pi(\mathbf{X}|\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ represent the likelihood and prior distributions, respectively. In these applications, the distribution parameters are interpreted as the configuration variables ($\boldsymbol{\theta} \equiv \mathbf{q}$). We introduce a potential energy $U(\boldsymbol{\theta})$ by defining $\pi(\boldsymbol{\theta}|\mathbf{X}) \propto \exp(-\beta U(\boldsymbol{\theta}))$, thus taking the logarithm of (1.5) gives

$$U(\boldsymbol{\theta}) = -\log \pi(\mathbf{X}|\boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta})\,. \tag{1.6}$$

Assuming the data are independent and identically distributed (i.i.d.), the logarithm of the likelihood distribution can then be calculated as

$$\log \pi(\mathbf{X}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \pi(\mathbf{x}_i|\boldsymbol{\theta})\,, \tag{1.7}$$

where N is the size of the entire dataset.

However, in machine learning applications, one often finds that directly sampling with the entire large-scale dataset is computationally infeasible. For instance, standard Markov Chain Monte Carlo (MCMC) methods [43] require the calculation of the acceptance probability and the creation of informed proposals based on the whole dataset, while the gradient is evaluated through the whole dataset in the Hybrid Monte Carlo (HMC) method [8, 15, 23], again resulting in severe computational complexity.

In order to improve the efficiency of simulation, the so-called Stochastic Gradient Langevin Dynamics (SGLD) was recently proposed [63] based on using a random (and much smaller, $\tilde{n} \ll N$) subset to approximate the likelihood of the dataset for given parameters,

$$\log \pi(\mathbf{X}|\boldsymbol{\theta}) \approx \frac{N}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \log \pi(\mathbf{x}_{r_i}|\boldsymbol{\theta})\,, \tag{1.8}$$

where $\{\mathbf{x}_{r_i}\}_{i=1}^{\tilde{n}}$ represents a random subset of $\mathbf{X}$. Overall, the "noisy" potential energy now can be written as

$$\tilde{U}(\boldsymbol{\theta}) = -\frac{N}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \log \pi(\mathbf{x}_{r_i}|\boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta})\,, \tag{1.9}$$

with "noisy" force $\tilde{\mathbf{F}}(\boldsymbol{\theta}) = -\nabla \tilde{U}(\boldsymbol{\theta})$.

1.3. Sampling Methods for Noisy Gradients. The challenge is to identify a method to compute samples distributed according to the Gibbs distribution $\rho(\mathbf{q}) = Z^{-1} \exp(-\beta U(\mathbf{q}))$ where the only available information is a stochastically perturbed force $\tilde{\mathbf{F}}(\mathbf{q})$ defined in the previous section.

In the original SGLD method, samples are generated by Brownian dynamics,

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \Delta t_n \tilde{\mathbf{F}}(\mathbf{q}_n) + \sqrt{2\beta^{-1}\Delta t_n}\mathbf{R}_n\,, \tag{1.10}$$

where $\mathbf{R}_n$ is a vector of i.i.d. standard normal random variables. It should be emphasized that Δt_n is a sequence of stepsizes decreasing to zero [63]. Although a central limit theorem associated with the decreasing stepsize sequence was established by Teh et al. [61], a fixed stepsize is often preferred in practice which is the choice in this article as in Vollmer et al. [62], where a modified SGLD (mSGLD) is introduced:

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \Delta t \tilde{\mathbf{F}}(\mathbf{q}_n) + \sqrt{2\beta^{-1}\Delta t}\left(\mathbf{I} - \frac{\Delta t}{4}\mathrm{Cov}\tilde{\mathbf{F}}(\mathbf{q}_n)\right)\mathbf{R}_n\,, \tag{1.11}$$

where

$$\mathrm{Cov}\tilde{\mathbf{F}}_{ij} = \mathbb{E}\left[\left(\tilde{\mathbf{F}}_i - \mathbb{E}(\tilde{\mathbf{F}}_i)\right)\left(\tilde{\mathbf{F}}_j - \mathbb{E}(\tilde{\mathbf{F}}_j)\right)^T\right] \tag{1.12}$$

is the covariance matrix of the noisy force.

A Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) method was also proposed very recently by Chen et al. [11], which incorporates a parameter-dependent diffusion matrix $\mathbf{\Sigma}(\mathbf{q})$ (i.e. the covariance matrix of the noisy force). $\mathbf{\Sigma}(\mathbf{q})$ is intended to effectively offset the stochastic perturbation of the gradient. However, it is very difficult to accommodate $\mathbf{\Sigma}(\mathbf{q})$ in practice, moreover, as pointed out in [14] poor estimation of it may have a significant adverse influence in correctly sampling the target distribution unless the stepsize is small enough.

These problems challenge the conventional mechanism of thermostats. An article of Jones and Leimkuhler [26] provides an alternative means of tackling this problem by showing that Nosé-Hoover dynamics is able to adaptively dissipate excess heat pumped into the system while maintaining the Gibbs (canonical) distribution. In the setting of systems involving a driving stochastic perturbation, the Nosé-Hoover method is referred to as Ad-NH, with similar generalizations of Nosé-Hoover-Langevin (Ad-NHL) and Langevin dynamics (Ad-Langevin) available. An idea equivalent to Ad-Langevin was very recently applied in the setting of Bayesian sampling for use in data science calculations by Ding et al. [14], which they referred to as the Stochastic Gradient Nosé-Hoover Thermostat (SGNHT). It showed significant advantages over alternative techniques such as Stochastic Gradient Hamiltonian Monte Carlo [11]. However, the numerical method used by Ding et al. [14] is not optimal, either in terms of its accuracy (measured per unit work) or its stability (measured by the largest usable stepsize).

Although extended systems have been increasingly popular in molecular simulations, the mathematical analysis of the order of convergence, specifically in terms of the bias in averaged quantities computed using numerical trajectories, is not fully understood. Using a splitting approach, we propose in this article an alternative numerical method for Adaptive Langevin simulation that substantially improves on the existing schemes in the literature in terms of accuracy, robustness and overall numerical efficiency.

The rest of the article is organized as follows. In Section 2, we describe the construction of adaptive formulations for noisy gradients including the Adaptive Langevin (SGNHT) method. Section 3 considers the construction of numerical methods for solving the stochastic differential equations. Numerical experiments are performed in Section 4. Our experiments are of a more limited nature in comparison with those of Ding et al. [14], but we believe them to be representative of performance on a significant class of problems. Finally, we summarize our findings in Section 5.

2. Adaptive Thermostats for Noisy Gradients. In this section, we discuss the construction of thermostats to approximate samples with respect to the target measure (i.e. the correct marginalized Gibbs density), if the covariance matrix of the noisy force is constant, i.e. $\mathbf{\Sigma}(\mathbf{q}) = \sigma^2\mathbf{I}$ (σ is a constant positive quantity). The procedure was outlined in the paper of Jones and Leimkuhler [26] and relies on the fact that a fixed amplitude noise perturbation engenders a shift of the auxiliary variable in the extended stationary distribution associated with the Nosé-Hoover thermostat.

If the system is not coming from a Newtonian dynamics model, then it is unclear that we need to rely on second order dynamics for this purpose. To see why this is the case, we explain what goes wrong if we try to use first order dynamics. In what follows, we assume that the covariance matrix of the noisy force is constant, although we intend ultimately to apply the method more generally (see recent work on a novel covariance-controlled adaptive Langevin thermostat that can handle parameter-dependent noise in [57]). Even in the constant σ case it is a nontrivial problem to extract statistics related to a particular target temperature, since we do not assume that σ is known.

For σ constant, let us first consider the SDE

$$\mathrm{d}\mathbf{q} = -\xi\nabla U(\mathbf{q})\mathrm{d}t + \sigma\mathrm{d}\mathbf{W}\,, \tag{2.1}$$

$$\mathrm{d}\xi = \chi(\mathbf{q})\mathrm{d}t\,, \tag{2.2}$$

and seek $\chi(\cdot)$ so that an extended Gibbs distribution with density of the form $\psi(\mathbf{q},\xi) = \bar{\rho}_\beta(\mathbf{q})\varphi(\xi)$ is (ergodically) preserved. The variable ξ is an auxiliary variable. We do not generally care what its distribution is, but it is crucial that

(i) the overall density is in product form, and
(ii) $\varphi(\xi) \geq 0$ is normalizable and of a simple, easily sampled form.

These conditions ensure that we can easily average out over the auxiliary variable to compute the averages of functions of $\mathbf{q}$ which are of greatest interest.

Proposition 1: Let $\chi(\mathbf{q}) = -\beta^{-1}\Delta U(\mathbf{q}) + \|\nabla U(\mathbf{q})\|^2$, then (2.1)-(2.2) preserves the modified Gibbs distribution

$$\tilde{\rho}(\mathbf{q}, \xi) = \bar{\rho}_\beta(\mathbf{q}) e^{-\beta(\xi - \hat{\gamma})^2/2}\,,$$

where $\hat{\gamma} = \beta\sigma^2/2$.
Proof: The Fokker-Planck equation corresponding to (2.1)-(2.2) is

$$\rho_t = \mathcal{L}^\dagger \rho := \xi \nabla \cdot (\nabla U(\mathbf{q})\rho(\mathbf{q}, \xi)) + \frac{\sigma^2}{2}\Delta\rho - \frac{\partial}{\partial \xi}(\chi(\mathbf{q})\rho)\,.$$

Just insert $\tilde{\rho}$ into the operator $\mathcal{L}^\dagger$ to see that it vanishes. □.

Proposition 1 tells us that if we can solve the system (2.1)-(2.2), under an assumption of ergodicity, we can compute averages with respect to the target Gibbs distribution without actually knowing the value of σ. σ could be observed retrospectively by simply averaging ξ during simulation, since $\langle \xi \rangle = \beta\sigma^2/2$.

The problem is that the dynamics (2.1)-(2.2) is not quite what we want. A typical numerical method for this system might be constructed based on modification of the Euler-Maruyama method:

$$\mathbf{q}_{n+1} = \mathbf{q}_n - \Delta t \xi_n \nabla U(\mathbf{q}_n) + \sigma\sqrt{\Delta t}\mathbf{R}_n\,, \tag{2.3}$$

$$\xi_{n+1} = \xi_n + \Delta t \chi(\mathbf{q}_n)\,, \tag{2.4}$$

however, observe that this method requires separate knowledge of $\nabla U(\mathbf{q})$ and σ, which is generally impossible a priori, as we assume that the force is polluted by unknown noise. The form of the equations means that we evaluate the product of ξ and the deterministic force on the one hand, and the random perturbation separately, and these contributions are independently scaled by Δt and $\sqrt{\Delta t}$, respectively.

2.1. The Adaptive Langevin Thermostat (Ad-Langevin). To adaptively control the invariant distribution, we consider the following second order formulation, which was first introduced in the paper of Jones and Leimkuhler [26]:

$$\begin{aligned} \mathrm{d}\mathbf{q} &= \mathbf{M}^{-1}\mathbf{p}\mathrm{d}t\,, \\ \mathrm{d}\mathbf{p} &= \tilde{\mathbf{F}}(\mathbf{q})\mathrm{d}t - \xi\mathbf{p}\mathrm{d}t + \sigma_{\mathrm{A}}\mathbf{M}^{1/2}\mathrm{d}\mathbf{W}_{\mathrm{A}}\,, \\ \mathrm{d}\xi &= \mu^{-1}\left[\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} - N_{\mathrm{d}}k_{\mathrm{B}}T\right]\mathrm{d}t\,, \end{aligned} \tag{2.5}$$

In these equations, $\tilde{\mathbf{F}}(\mathbf{q})$ is meant to represent a noisy gradient which may be thought of as being defined by the relation

$$\tilde{\mathbf{F}}(\mathbf{q}) = -\nabla U(\mathbf{q}) + \sigma\mathbf{M}^{1/2}\mathbf{R}\,, \tag{2.6}$$

where $\mathbf{R} = \mathbf{R}(t)$ is a collection of independent Gaussian white noise processes, i.e. $\langle \mathrm{R}_i(t)\mathrm{R}_j(s)\rangle = \delta_{ij}\delta(t-s)$ where δ_{ij} is the Kronecker delta and $\delta(t-s)$ is the Dirac delta function. $\sigma_{\mathrm{A}}\mathbf{M}^{1/2}\mathrm{d}\mathbf{W}_{\mathrm{A}}$ indicates the artificial noise added into the system to enhance the ergodicity, i.e. the constant σ_{A} is known a priori. All the components of the Wiener process $\mathbf{W}_{\mathrm{A}}(t)$ are assumed to be independent. N_{d} denotes the number of degrees of freedom of the system. μ is a coupling parameter which is referred to as the "thermal mass". k_{B} and T, satisfying the relation $\beta^{-1} = k_{\mathrm{B}}T$, represent the Boltzmann constant and system temperature, respectively.

A similar system was used by Ding et al. [14] (SGNHT), who also explored its application to three examples from machine learning. These experiments demonstrated that Ad-Langevin has superior performance compared to SGHMC in various applications, confirming the importance of adaptively dissipating additional noise in sampling. However, there remain two important issues that we wish to address in this article: (1) the underlying dynamics of the Ad-Langevin method is not clear due to the presence of the stochastically perturbed gradient; (2) little attention has been paid to the design of optimal numerical methods for implementing Ad-Langevin with attention to stability and numerical efficiency.

One may wonder why the artificial noise is needed (i.e. $\sigma_{\mathrm{A}} \neq 0$), since we are assuming the presence of noise in the gradient itself. The reason is as follows: in defining a numerical method for the noisy gradient system, the force (including the random perturbation) will in general be multiplied by Δt, where Δt is the timestep. On the other hand, the Itō rule implies that the scaling of random perturbations in an SDE should be by a factor proportional to $\sqrt{\Delta t}$, thus effectively, if we are to relate the thermostatted method to a standard SDE, the standard deviation of the noise is reduced by multiplication by the factor $\sqrt{\Delta t}$. The noise perturbation introduced at each timestep (and the effective diffusion) is thus reduced for small stepsizes and it is therefore important to inject an additional artificial noise in order to stabilize the invariant distribution. A rewriting of the Ad-Langevin system as a standard Itō SDE system makes clear the relation between the different terms

$$\begin{aligned} \mathrm{d}\mathbf{q} &= \mathbf{M}^{-1}\mathbf{p}\mathrm{d}t\,, \\ \mathrm{d}\mathbf{p} &= -\nabla U(\mathbf{q})\mathrm{d}t + \sigma\sqrt{\Delta t}\mathbf{M}^{1/2}\mathrm{d}\mathbf{W} - \xi\mathbf{p}\mathrm{d}t + \sigma_{\mathrm{A}}\mathbf{M}^{1/2}\mathrm{d}\mathbf{W}_{\mathrm{A}}\,, \\ \mathrm{d}\xi &= \mu^{-1}\left[\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} - N_{\mathrm{d}}k_{\mathrm{B}}T\right]\mathrm{d}t\,, \end{aligned} \tag{2.7}$$

where $\mathbf{W} = \mathbf{W}(t)$ is an additional vector of standard Wiener processes.

Let us note the main features of the dynamics (2.7):

(i) The equations are a combination of Langevin dynamics and Nosé-Hoover dynamics. If ξ is constant in the equation for the momentum, then the system reduces to Langevin dynamics. In the absence of noise $\sigma_{\mathrm{A}} = 0$ (and $\sigma = 0$) then the system reduces to Nosé-Hoover. The system (2.7) may be regarded as a sort of Langevin dynamics where the friction coefficient, rather than being fixed a priori, is automatically and adaptively determined in order to achieve the desired temperature (which is specified in the control law defining the evolution of ξ).

(ii) The invariant distribution for the given system may be directly obtained by study of its Fokker-Planck equation. Following [26], it is straightforward to show that (2.7) has the following invariant distribution,

$$\tilde{\rho}_\beta(\mathbf{q},\mathbf{p},\xi) = \frac{1}{Z}\exp\left(-\beta H(\mathbf{q},\mathbf{p})\right)\exp\left(-\frac{\beta\mu}{2}(\xi-\hat{\gamma})^2\right)\,, \tag{2.8}$$

where Z is the normalizing constant and

$$\hat{\gamma} = \frac{\beta\left(\sigma_{\mathrm{F}}^2 + \sigma_{\mathrm{A}}^2\right)}{2}\,, \tag{2.9}$$

where $\sigma_{\mathrm{F}} = \sigma\sqrt{\Delta t}$. Observe that this means that if $\sigma_{\mathrm{A}} = 0$, then, as $\lim_{\Delta t\to 0}\sigma_{\mathrm{F}} = 0$, we find that ξ tends to a variable which is normally distributed with mean zero. Alternatively, if $\sigma_{\mathrm{A}} \neq 0$, one would obtain

$$\xi \overset{\mathscr{L}}{\to} \mathcal{N}\left(\frac{\beta\sigma_{\mathrm{A}}^2}{2}, \beta^{-1}\mu^{-1}\right), \qquad t\to\infty\,, \qquad \Delta t\to 0\,,$$

where $\beta^{-1}\mu^{-1}$ is the variance and the symbol $\overset{\mathscr{L}}{\to}$ indicates that ξ converges in probability law to a normally distributed random variable with the indicated parameters. The order of the limits here is important: $t\to\infty$ first (to reach the invariant distribution), then $\Delta t\to 0$.

(iii) The ergodicity of (2.7) with respect to the distribution indicated above can easily be demonstrated by reference to Hörmander's condition for hypoellipticity following the method in [42], as for Langevin dynamics. The only additional step is to verify that the noise propagates into the ξ variable, which follows due to its strong coupling to the momenta.

(iv) This dynamics is a bit unusual in that it must be viewed as stepsize-dependent, although we mention that such mixed systems are used in the study of backward error analysis [38]. One simply thinks of the characteristics of stochastic paths associated with (2.7) as being stepsize dependent. Although (2.7) takes on the appearance of a standard Itō SDE system, we must bear in mind that in discretizing these equations the conservative force $\mathbf{F}(\mathbf{q})$ and the associated noise term $\sigma\sqrt{\Delta t}\mathbf{M}^{1/2}\mathrm{d}\mathbf{W}$ must be evaluated together at every stage, since the formulation (2.7) is a notational device to make clear the properties of the system.

3. Numerical Methods for Adaptive Thermostats. Since stochastic systems in most of the cases cannot be solved "exactly", splitting methods are often adopted in practice. For instance here, the vector field of the Ad-Langevin/SGNHT (2.5) can be split into four pieces which are denoted as "A", "B", "O" and "D", in such a way that each piece can be solved "exactly",

$$\mathrm{d}\begin{bmatrix}\mathbf{q}\\ \mathbf{p}\\ \xi\end{bmatrix} = \underbrace{\begin{bmatrix}\mathbf{M}^{-1}\mathbf{p}\\ \mathbf{0}\\ 0\end{bmatrix}\mathrm{d}t}_{\mathrm{A}} + \underbrace{\begin{bmatrix}\mathbf{0}\\ -\nabla U(\mathbf{q})+\sigma\mathbf{M}^{1/2}\mathbf{R}\\ 0\end{bmatrix}\mathrm{d}t}_{\mathrm{B}} + \underbrace{\begin{bmatrix}\mathbf{0}\\ -\xi\mathbf{p}\mathrm{d}t+\sigma_{\mathrm{A}}\mathbf{M}^{1/2}\mathrm{d}\mathbf{W}_{\mathrm{A}}\\ 0\end{bmatrix}}_{\mathrm{O}} + \underbrace{\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ G(\mathbf{p})\end{bmatrix}\mathrm{d}t}_{\mathrm{D}},$$

where $G(\mathbf{p}) = \mu^{-1}\left[\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} - N_{\mathrm{d}}k_{\mathrm{B}}T\right]$.

Clearly parts "A" and "D" can be solved "exactly". As mentioned previously, the underlying dynamics for "B" is

$$\mathrm{d}\mathbf{p} = -\nabla U(\mathbf{q})\mathrm{d}t + \sigma_{\mathrm{F}}\mathbf{M}^{1/2}\mathrm{d}\mathbf{W}\,, \tag{3.1}$$

where $\mathbf{q}$ is fixed and $\sigma_{\mathrm{F}} = \sigma\sqrt{\Delta t}$. Integrating (3.1) from 0 to Δt gives the exact solution in distribution of this part as:

$$\mathbf{p}(\Delta t) = \mathbf{p}(0) - \Delta t\nabla U(\mathbf{q}) + \sqrt{\Delta t}\sigma_{\mathrm{F}}\mathbf{M}^{1/2}\mathbf{R}\,, \tag{3.2}$$

$$= \mathbf{p}(0) + \Delta t[-\nabla U(\mathbf{q}) + \sigma\mathbf{M}^{1/2}\mathbf{R}] = \mathbf{p}(0) + \Delta t\tilde{\mathbf{F}}(\mathbf{q})\,, \tag{3.3}$$

where $\mathbf{R}$ is a vector of i.i.d. standard normal random variables. It should be noted that applying the Euler-Maruyama method to (3.1) gives the same result, thus, for constant force, Euler-Maruyama is "exact".

The "O" or "Ornstein-Uhlenbeck" part is usually stated with ξ a positive constant, in which case the solution is found to be [29]

$$\mathbf{p}(\Delta t) = e^{-\xi\Delta t}\mathbf{p}(0) + \sigma_{\mathrm{A}}\sqrt{\frac{1-e^{-2\xi\Delta t}}{2\xi}}\mathbf{M}^{1/2}\mathbf{R}\,, \tag{3.4}$$

where $\mathbf{p}(0)$ is the initial value of the variable and $\mathbf{R}$ is a vector of i.i.d. standard normal random variables. However, the same formula (3.4) is easily seen to be valid for $\xi < 0$, since the quantity $(1-e^{-2\xi\Delta t})/(2\xi)$ is strictly greater than zero unless $\xi = 0$. (The proof is obtained by following the standard procedure [29].) When $\xi = 0$, one can simply replace $(1-e^{-2\xi\Delta t})/(2\xi)$ by its well-defined asymptotic limit,

$$\mathbf{p}(\Delta t) = \mathbf{p}(0) + \sqrt{\Delta t}\sigma_{\mathrm{A}}\mathbf{M}^{1/2}\mathbf{R}\,. \tag{3.5}$$

The generators associated with each piece are defined, respectively, as

$$\begin{aligned}
\mathcal{L}_{\mathrm{A}} &= \mathbf{M}^{-1}\mathbf{p}\cdot\nabla_{\mathbf{q}}\,,\\
\mathcal{L}_{\mathrm{B}} &= -\nabla U(\mathbf{q})\cdot\nabla_{\mathbf{p}} + \frac{\sigma_{\mathrm{F}}^2}{2}\mathrm{Tr}\left(\mathbf{M}\nabla_{\mathbf{p}}^2\right)\,,\\
\mathcal{L}_{\mathrm{O}} &= -\xi\mathbf{p}\cdot\nabla_{\mathbf{p}} + \frac{\sigma_{\mathrm{A}}^2}{2}\mathrm{Tr}\left(\mathbf{M}\nabla_{\mathbf{p}}^2\right)\,,\\
\mathcal{L}_{\mathrm{D}} &= G(\mathbf{p})\frac{\partial}{\partial\xi}\,,
\end{aligned}$$

where $\sigma_{\mathrm{F}} = \sigma\sqrt{\Delta t}$ in part "B" is stepsize dependent.

Overall, the generator of the Ad-Langevin/SGNHT (2.5) system can be written as

$$\mathcal{L} = \mathcal{L}_{\mathrm{A}} + \mathcal{L}_{\mathrm{B}} + \mathcal{L}_{\mathrm{O}} + \mathcal{L}_{\mathrm{D}}\,. \tag{3.6}$$

The flow map (or phase space propagator) of the system can be written in the shorthand notation

$$\mathcal{F}_t = e^{t\mathcal{L}}\,,$$

where the exponential map here denotes the solution operator. Approximations of $\mathcal{F}_t$ can be obtained as products (taken in different arrangements) of exponentials of the splitting terms. For example, the phase space propagation of the method proposed by Ding et al. [14] for the Ad-Langevin/SGNHT (2.5) system (denoted as "SGNHT-N") can be written as

$$\exp\left(\Delta t\hat{\mathcal{L}}_{\mathrm{SGNHT-N}}\right) = \exp\left(\Delta t\mathcal{L}_{\mathrm{P}}\right)\exp\left(\Delta t\mathcal{L}_{\mathrm{A}}\right)\exp\left(\Delta t\mathcal{L}_{\mathrm{D}}\right)\,, \tag{3.7}$$

where

$$\mathcal{L}_{\mathrm{P}} = \mathcal{L}_{\mathrm{B}} + \mathcal{L}_{\mathrm{O}}\,, \tag{3.8}$$

and $\exp\left(\Delta t\mathcal{L}_f\right)$ represents the phase space propagator associated with the corresponding vector field f. Because of its nonsymmetric structure, one anticipates first order convergence to the invariant measure (for any choice of σ). Due to the naming of the component parts, the SGNHT-N method may be denoted by "PAD".

Overall, the SGNHT-N/PAD integration method is as follows:

$$\begin{aligned}
\mathbf{p}_{n+1} &= \mathbf{p}_n + \Delta t\left(-\nabla U(\mathbf{q}_n) + \sigma\mathbf{M}^{1/2}\mathbf{R}_n'\right) - \Delta t\xi_n\mathbf{p}_n + \sqrt{\Delta t}\sigma_{\mathrm{A}}\mathbf{M}^{1/2}\mathbf{R}_n\,,\\
\mathbf{q}_{n+1} &= \mathbf{q}_n + \Delta t\mathbf{M}^{-1}\mathbf{p}_{n+1}\,,\\
\xi_{n+1} &= \xi_n + \Delta t\mu^{-1}\left(\mathbf{p}_{n+1}^T\mathbf{M}^{-1}\mathbf{p}_{n+1} - N_{\mathrm{d}}k_{\mathrm{B}}T\right)\,,
\end{aligned}$$

where $\mathbf{R}_n'$ and $\mathbf{R}_n$, respectively, are vectors of i.i.d. standard normal random variables.

We propose symmetric alternative methods, such as the following symmetric Ad-Langevin/SGNHT (SGNHT-S) splitting method

$$e^{\Delta t\hat{\mathcal{L}}_{\mathrm{SGNHT-S}}} = e^{\frac{\Delta t}{2}\mathcal{L}_{\mathrm{B}}}e^{\frac{\Delta t}{2}\mathcal{L}_{\mathrm{A}}}e^{\frac{\Delta t}{2}\mathcal{L}_{\mathrm{D}}}e^{\Delta t\mathcal{L}_{\mathrm{O}}}e^{\frac{\Delta t}{2}\mathcal{L}_{\mathrm{D}}}e^{\frac{\Delta t}{2}\mathcal{L}_{\mathrm{A}}}e^{\frac{\Delta t}{2}\mathcal{L}_{\mathrm{B}}}\,, \tag{3.9}$$

where exact solvers for parts "B" and "O" derived above are applied. The SGNHT-S method may be referred to as "BADODAB" where it should be noted that the various operations are symmetrically applied and the steplengths are uniform and span the interval Δt. Other symmetric splittings are considered below.

The SGNHT-S numerical integration method may be written:

$$
\begin{aligned}
\mathbf{p}_{n+1/3} &= \mathbf{p}_n + (\Delta t/2)\left(-\nabla U(\mathbf{q}_n) + \sigma \mathbf{M}^{1/2}\mathbf{R}'_n\right)\,,\\
\mathbf{q}_{n+1/2} &= \mathbf{q}_n + (\Delta t/2)\mathbf{M}^{-1}\mathbf{p}_{n+1/3}\,,\\
\xi_{n+1/2} &= \xi_n + (\Delta t/2)\mu^{-1}\left(\mathbf{p}_{n+1/3}^T\mathbf{M}^{-1}\mathbf{p}_{n+1/3} - N_{\mathrm{d}}k_{\mathrm{B}}T\right)\,,\\
\text{if } (\xi_{n+1/2}\neq 0):\ \mathbf{p}_{n+2/3} &= e^{-\xi_{n+1/2}\Delta t}\mathbf{p}_{n+1/3} + \sigma_{\mathrm{A}}\sqrt{(1-e^{-2\xi_{n+1/2}\Delta t})/(2\xi_{n+1/2})}\mathbf{M}^{1/2}\mathbf{R}_n\,,\\
\text{else}:\ \mathbf{p}_{n+2/3} &= \mathbf{p}_{n+1/3} + \sqrt{\Delta t}\sigma_{\mathrm{A}}\mathbf{M}^{1/2}\mathbf{R}_n\,,\\
\xi_{n+1} &= \xi_{n+1/2} + (\Delta t/2)\mu^{-1}\left(\mathbf{p}_{n+2/3}^T\mathbf{M}^{-1}\mathbf{p}_{n+2/3} - N_{\mathrm{d}}k_{\mathrm{B}}T\right)\,,\\
\mathbf{q}_{n+1} &= \mathbf{q}_{n+1/2} + (\Delta t/2)\mathbf{M}^{-1}\mathbf{p}_{n+2/3}\,,\\
\mathbf{p}_{n+1} &= \mathbf{p}_{n+2/3} + (\Delta t/2)\left(-\nabla U(\mathbf{q}_{n+1}) + \sigma \mathbf{M}^{1/2}\mathbf{R}'_{n+1}\right)\,.
\end{aligned}
$$

The force computed at the end of each timestep can be reused at the start of the next step, thus only one force calculation is needed in SGNHT-S at each timestep, the same as for SGNHT-N. In practice, one could replace the exponential and square root operations in the exact solver of the "O" part by their respective well-defined asymptotic expansions to reduce the computational cost.

3.1. Order of Convergence of Ad-Langevin/SGNHT. The analysis of the accuracy of ergodic averages (averages with respect to the invariant measure) in stochastic numerical methods can be performed using the framework of long-time Talay-Tubaro expansion, as developed in [1,2,13,34–36,60]. In what follows we compare the order of convergence of the two Ad-Langevin/SGNHT methods with a clean gradient.

For a splitting method described by $\mathcal{L} = \mathcal{L}_\alpha + \mathcal{L}_\beta + \ldots + \mathcal{L}_\zeta$, we define the effective operator $\hat{\mathcal{L}}^\dagger$ associated with the perturbed system obtained using the numerical method with stepsize Δt by the relation

$$
\exp\left(\Delta t\hat{\mathcal{L}}^\dagger\right) = \exp\left(\Delta t\mathcal{L}_\alpha^\dagger\right)\exp\left(\Delta t\mathcal{L}_\beta^\dagger\right)\cdots\exp\left(\Delta t\mathcal{L}_\zeta^\dagger\right)\,.
$$

This operator can be computed using the Baker-Campbell-Hausdorff (BCH) expansion and can thus be viewed as a perturbation of the exact Fokker-Planck operator $\mathcal{L}^\dagger$

$$
\hat{\mathcal{L}}^\dagger = \mathcal{L}^\dagger + \Delta t\mathcal{L}_1^\dagger + \Delta t^2\mathcal{L}_2^\dagger + O(\Delta t^3)\,, \tag{3.10}
$$

for some perturbation operators $\mathcal{L}_i^\dagger$.

We also define the invariant distribution $\hat{\rho}$ associated with the numerical method as an approximation of the target invariant distribution $\tilde{\rho}_\beta$

$$
\hat{\rho} = \tilde{\rho}_\beta\left[1 + \Delta t f_1 + \Delta t^2 f_2 + \Delta t^3 f_3 + O(\Delta t^4)\right]\,, \tag{3.11}
$$

for some correction functions f_i satisfying $\langle f_i\rangle = 0$.

Substituting $\hat{\mathcal{L}}^\dagger$ and $\hat{\rho}$ into the stationary Fokker-Planck equation

$$
\hat{\mathcal{L}}^\dagger\hat{\rho} = 0
$$

yields

$$
\left(\mathcal{L}^\dagger + \Delta t\mathcal{L}_1^\dagger + \Delta t^2\mathcal{L}_2^\dagger + O(\Delta t^3)\right)\left(\tilde{\rho}_\beta\left[1 + \Delta t f_1 + \Delta t^2 f_2 + \Delta t^3 f_3 + O(\Delta t^4)\right]\right) = 0\,.
$$

Since the exact Fokker-Planck operator preserves the invariant canonical distribution, i.e. $\mathcal{L}^\dagger\tilde{\rho}_\beta = 0$, we obtain

$$
\mathcal{L}^\dagger(\tilde{\rho}_\beta f_1) = -\mathcal{L}_1^\dagger\tilde{\rho}_\beta\,, \tag{3.12}
$$

by equating first order terms in Δt.

For any particular integration scheme it is possible to find the perturbation operator $\mathcal{L}_1^\dagger$ by using the BCH expansion. Then we can calculate its action on $\tilde{\rho}_\beta$. The last step, namely obtaining the leading correction function f_1, requires the solution of the above PDE (see examples in Langevin dynamics [34]). In general solving for f_1 in closed form is difficult, and it does not get simpler as we consider, as here, more complicated formulations than Langevin dynamics and more complicated splittings.

According to the BCH expansion, for (noncommutative) linear operators X and Y, we have

$$\exp(\Delta t X)\exp(\Delta t Y) = \exp(\Delta t Z_1)\,,$$

where

$$Z_1 = X + Y + \frac{\Delta t}{2}[X,Y] + \frac{\Delta t^2}{12}([X,[X,Y]] - [Y,[X,Y]]) + O(\Delta t^3)\,, \tag{3.13}$$

and subsequently

$$\exp\left(\frac{\Delta t}{2}X\right)\exp(\Delta t Y)\exp\left(\frac{\Delta t}{2}X\right) = \exp(\Delta t Z_2)\,,$$

where

$$Z_2 = X + Y + \frac{\Delta t^2}{12}\left([Y,[Y,X]] - \frac{1}{2}[X,[X,Y]]\right) + O(\Delta t^4)\,. \tag{3.14}$$

The notation $[X,Y] = XY - YX$ denotes the commutator of operators X and Y.

These equations demonstrate that for nonsymmetric splitting methods, there typically exists a nonzero term $\mathcal{L}_1^\dagger \propto [X,Y] \neq 0$, while the condition $\mathcal{L}_1^\dagger = 0$, implying $f_1 = 0$, is automatically satisfied for symmetric splitting methods, thus, for observables $\phi(\mathbf{q},\mathbf{p},\xi)$, assuming the asymptotic expansion holds, the computed average would be of order two

$$\langle\phi\rangle_{\Delta t} = \langle\phi\rangle + \Delta t\langle\phi f_1\rangle + \Delta t^2\langle\phi f_2\rangle + \ldots = \langle\phi\rangle + O(\Delta t^2)\,,$$

where $\langle\cdot\rangle$ denote the average with respect to the target invariant distribution. Therefore, the SGNHT-S method (3.9) would have second order convergence for all the observables.

We can work out the leading operator $\mathcal{L}_1^\dagger$ associated with the nonsymmetric SGNHT-N ("PAD") method (3.7) of Ding et al. [14],

$$\mathcal{L}_{1,\mathrm{PAD}}^\dagger = \frac{1}{2}\left(\left[\mathcal{L}_\mathrm{D}^\dagger,\mathcal{L}_\mathrm{A}^\dagger\right] + \left[\mathcal{L}_\mathrm{D}^\dagger,\mathcal{L}_\mathrm{P}^\dagger\right] + \left[\mathcal{L}_\mathrm{A}^\dagger,\mathcal{L}_\mathrm{P}^\dagger\right]\right)\,. \tag{3.15}$$

It is clear that the leading term $f_{1,\mathrm{PAD}}$ in the perturbed distribution (3.11) is in general nonzero. Therefore the nonsymmetric SGNHT-N ("PAD") method would be expected to exhibit first order convergence to the invariant measure. It should be noted that, if certain conditions are satisfied, higher order convergence to the invariant measure would be possible as demonstrated by Abdulle et al. [1, 2]. However, it can be easily demonstrated that it is not the case here for the SGNHT-N ("PAD") method. In the presence of a noisy gradient, the Ad-Langevin/SGNHT methods, despite the stepsize-dependency (2.7), would similarly (and generally) be expected to be first order with respect to the invariant distribution.

3.2. Superconvergence Property. Recently, it has been demonstrated in the setting of Langevin dynamics that a particular symmetric splitting method ("BAOAB"), which requires only one force calculation per step, is fourth order for configurational quantities in the ergodic limit and in the limit of large friction [34, 36].

In what follows we demonstrate that the newly-proposed SGNHT-S ("BADODAB") method (3.9) effectively inherits the superconvergence property of BAOAB in the setting of Ad-Langevin/SGNHT

system (2.7) with a clean gradient, in case the parameters σ_{A} and μ are both taken to infinity in a suitable way. For simplicity we consider here a one-dimensional model $H = p^2/2 + U(q)$, but the analysis could easily be extended to higher dimensions.

Following the standard procedure described in Section 3.1, we obtain the following PDE associated with the BADODAB method,

$$\mathcal{L}^\dagger(\tilde{\rho}_\beta f_2) = -\mathcal{L}_2^\dagger \tilde{\rho}_\beta \,, \tag{3.16}$$

where $\mathcal{L}^\dagger$ is the exact Fokker-Planck operator

$$\mathcal{L}^\dagger = -p\partial_q + U'(q)\partial_p + \xi\partial_p(p\cdot) + \frac{\hat{\gamma}}{\beta}\partial_{pp} - \frac{1}{\mu}(p^2 - \beta^{-1})\partial_\xi \tag{3.17}$$

with invariant measure

$$\tilde{\rho}_\beta(q,p,\xi) = \frac{1}{Z}\exp\left(-\beta H(q,p)\right)\exp\left(-\frac{\beta\mu}{2}(\xi-\hat{\gamma})^2\right) , \tag{3.18}$$

where $\hat{\gamma} = \langle\xi\rangle = \beta\sigma_{\mathrm{A}}^2/2$ and $\mathcal{L}_2^\dagger$ can be calculated by using BCH expansion

$$\begin{aligned}\mathcal{L}_2^\dagger = &\frac{1}{12}\left(\left[\mathcal{L}_{\mathrm{O}}^\dagger,\left[\mathcal{L}_{\mathrm{O}}^\dagger,\mathcal{L}_{\mathrm{D}}^\dagger\right]\right] + \left[\mathcal{L}_{\mathrm{D}}^\dagger + \mathcal{L}_{\mathrm{O}}^\dagger,\left[\mathcal{L}_{\mathrm{D}}^\dagger + \mathcal{L}_{\mathrm{O}}^\dagger,\mathcal{L}_{\mathrm{A}}^\dagger\right]\right] + \left[\mathcal{L}_{\mathrm{A}}^\dagger + \mathcal{L}_{\mathrm{D}}^\dagger + \mathcal{L}_{\mathrm{O}}^\dagger,\left[\mathcal{L}_{\mathrm{A}}^\dagger + \mathcal{L}_{\mathrm{D}}^\dagger + \mathcal{L}_{\mathrm{O}}^\dagger,\mathcal{L}_{\mathrm{B}}^\dagger\right]\right]\right) \\ &- \frac{1}{24}\left(\left[\mathcal{L}_{\mathrm{D}}^\dagger,\left[\mathcal{L}_{\mathrm{D}}^\dagger,\mathcal{L}_{\mathrm{O}}^\dagger\right]\right] + \left[\mathcal{L}_{\mathrm{A}}^\dagger,\left[\mathcal{L}_{\mathrm{A}}^\dagger,\mathcal{L}_{\mathrm{D}}^\dagger + \mathcal{L}_{\mathrm{O}}^\dagger\right]\right] + \left[\mathcal{L}_{\mathrm{B}}^\dagger,\left[\mathcal{L}_{\mathrm{B}}^\dagger,\mathcal{L}_{\mathrm{A}}^\dagger + \mathcal{L}_{\mathrm{D}}^\dagger + \mathcal{L}_{\mathrm{O}}^\dagger\right]\right]\right) ,\end{aligned}$$

whose action on the extended invariant measure reads

$$\begin{aligned}\mathcal{L}_2^\dagger\tilde{\rho}_\beta = &\frac{1}{12}\left[-\beta p^3 U'''(q) + 4\beta p^2\xi^3 + 3\beta\xi p^2 U''(q) + 3\beta p U'(q)U''(q) + \frac{6\xi p^2}{\mu}\left(1-\beta p^2\right)\right]\tilde{\rho}_\beta \\ &+ \frac{\hat{\gamma}}{12}\left[3U''(q) + 4\xi^2 - 16\beta p^2\xi^2 - 6\beta U''(q)p^2 + \frac{6}{\mu}\left(2\beta p^4 - 5p^2 + \beta^{-1}\right)\right]\tilde{\rho}_\beta \\ &+ \hat{\gamma}^2\xi\left(2\beta p^2 - 1\right)\tilde{\rho}_\beta + \hat{\gamma}^3\left(\frac{2}{3} - \beta p^2\right)\tilde{\rho}_\beta \,.\end{aligned}$$

The equation is very complicated and we have no direct means of solving it. However, the additional variable ξ has mean $\hat{\gamma}$. If we suppose that μ is large, then the variance of ξ will be small. In this case we can consider the approximation obtained by replacing functions of ξ in the PDE (3.16) by their corresponding averages

$$\langle\xi\rangle = \hat{\gamma}\,, \qquad \langle\xi^2\rangle = \frac{1}{\beta\mu} + \hat{\gamma}^2\,, \qquad \langle\xi^3\rangle = \frac{3\hat{\gamma}}{\beta\mu} + \hat{\gamma}^3\,. \tag{3.19}$$

We use this as part of an averaging of the stationary Fokker-Planck equation with respect to the auxiliary variable. That is, we project the Fokker-Planck equation and its solution by integrating with respect to the Gaussian distribution of ξ in the ergodic limit. We can think of this is as defining a sort of "subspace projection"; it is related to the Galerkin method that is widely used in solving high dimensional linear systems and PDEs, including Fokker-Planck equations [10, 50]. In this case, we apply the projection operator [19]

$$\mathcal{P}\nu(q,p,\xi) := \frac{\int_{\Omega_\xi}\tilde{\rho}_\beta(q,p,\xi)\nu(q,p,\xi)\,\mathrm{d}\xi}{\int_{\Omega_\xi}\tilde{\rho}_\beta(q,p,\xi)\,\mathrm{d}\xi}\,, \tag{3.20}$$

where ν is an arbitrary function, to the PDE (3.16). Effectively, this results in the reduced equation

$$\check{\mathcal{L}}^\dagger(\rho_\beta\hat{f}_2) = -\rho_\beta\mathcal{P}\frac{\mathcal{L}_2^\dagger\tilde{\rho}_\beta}{\tilde{\rho}_\beta}\,, \tag{3.21}$$

where the operator $\check{\mathcal{L}}^\dagger$ is just the operator $\mathcal{L}^\dagger$ reduced by the action of the projection, and which acts on functions of q and p; this is nothing other than the corresponding adjoint generator of Langevin dynamics. Likewise $\hat{f}_2$ is now a function of q and p only. The right hand side simplifies to

$$\rho_\beta \mathcal{P}\frac{\mathcal{L}_2^\dagger \tilde{\rho}_\beta}{\tilde{\rho}_\beta} = \left(\frac{\beta}{12}\left[3pU'(q)U''(q) - p^3U'''(q)\right] + \frac{\hat{\gamma}}{12}\left[3U''(q) - 3\beta p^2 U''(q) + \frac{1}{\mu}\left(6\beta p^4 - 28p^2 + 10\beta^{-1}\right)\right]\right)\rho_\beta\,.$$

where ρ_β is the Gibbs (canonical) density ($\exp(-\beta H(q,p))$).

We consider the high friction limit ($\hat{\gamma}\to\infty$) and expand $\hat{f}_2$ in a series involving the reciprocal friction $\varepsilon = 1/\hat{\gamma}$,

$$\hat{f}_2(q,p) = \hat{f}_{2,0}(q,p) + \varepsilon \hat{f}_{2,1}(q,p) + \varepsilon^2 \hat{f}_{2,2}(q,p) + \dots \tag{3.22}$$

with each function $\hat{f}_{2,i}$ satisfying $\langle \hat{f}_{2,i}\rangle = 0$. Dividing (3.16) by the friction coefficient $\hat{\gamma}$, we obtain

$$\left(\bar{\mathcal{L}}_{\mathrm{O}}^\dagger + \varepsilon \mathcal{L}_{\mathrm{H}}^\dagger\right)\left(\hat{f}_{2,0} + \varepsilon \hat{f}_{2,1} + O(\varepsilon^2)\right)\rho_\beta = -\varepsilon \rho_\beta \mathcal{P}\frac{\mathcal{L}_2^\dagger \tilde{\rho}_\beta}{\tilde{\rho}_\beta}\,, \tag{3.23}$$

where

$$\bar{\mathcal{L}}_{\mathrm{O}}^\dagger = \partial_p(p\cdot) + \beta^{-1}\partial_{pp}\,, \qquad \mathcal{L}_{\mathrm{H}}^\dagger = -p\partial_q + U'(q)\partial_p\,. \tag{3.24}$$

We take the high thermal mass limit ($\mu\to\infty$) in such a way that $\varepsilon = 1/\mu = 1/\hat{\gamma}$. The use of this limit yields the following terms of the expansion of the right hand side in powers of ε. Defining

$$-\varepsilon\rho_\beta\mathcal{P}\frac{\mathcal{L}_2^\dagger\tilde{\rho}_\beta}{\tilde{\rho}_\beta} \equiv g = (g_0 + \varepsilon g_1)\,\rho_\beta\,,$$

we have

$$g_0 = -\frac{1}{4}\left[U''(q) - \beta p^2 U''(q)\right]\,, \tag{3.25}$$

$$g_1 = -\frac{1}{12}\left[3\beta p U'(q)U''(q) - \beta p^3 U'''(q) + 6\beta p^4 - 28p^2 + 10\beta^{-1}\right]\,. \tag{3.26}$$

Furthermore, by equating powers of the reciprocal friction ε, we can solve a sequence of equations

$$\begin{aligned}
\bar{\mathcal{L}}_{\mathrm{O}}^\dagger(\rho_\beta \hat{f}_{2,0}) &= g_0\rho_\beta\,,\\
\mathcal{L}_{\mathrm{H}}^\dagger(\rho_\beta \hat{f}_{2,0}) + \bar{\mathcal{L}}_{\mathrm{O}}^\dagger(\rho_\beta \hat{f}_{2,1}) &= g_1\rho_\beta\,,\\
\mathcal{L}_{\mathrm{H}}^\dagger(\rho_\beta \hat{f}_{2,1}) + \bar{\mathcal{L}}_{\mathrm{O}}^\dagger(\rho_\beta \hat{f}_{2,2}) &= 0\,,\\
&\vdots
\end{aligned}$$

to obtain the leading term $\hat{f}_{2,0}$,

$$\hat{f}_{2,0} \equiv \hat{f}_{2,0}^{\mathrm{BADODAB}} = \frac{1}{8}\left(U''(q) - \beta p^2 U''(q)\right)\,. \tag{3.27}$$

Moreover, it can be easily shown that the marginal average of $\hat{f}_{2,0}^{\mathrm{BADODAB}}$ with respect to momentum is zero,

$$\int_{\Omega_p} \hat{f}_{2,0}^{\mathrm{BADODAB}}(q,p)\rho_\beta\,\mathrm{d}\omega_p = 0\,, \tag{3.28}$$

which leads to the average of configurational observables $\phi(q)$ with respect to the invariant measure as

$$\langle\phi(q)\rangle_{\mathrm{BADODAB}} = \langle\phi(q)\rangle + \Delta t^2\langle\phi(q)\hat{f}_{2,0}^{\mathrm{BADODAB}}\rangle + O(\varepsilon\Delta t^2 + \Delta t^4)\,.$$

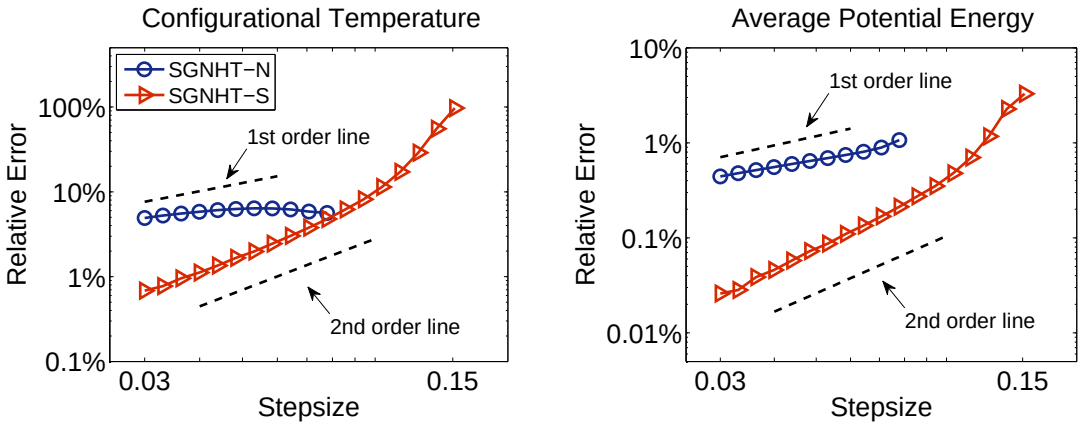

Fig. 1: Log-log plot of the relative error in computed configurational temperature (left) and average potential energy (right) against stepsize by using two Ad-Langevin/SGNHT methods (with a clean gradient). The system ($\sigma_{\mathrm{A}} = 3$) was simulated for 5000 reduced time units but only the last 80% of the data were collected to calculate the quantity to make sure the system was well equilibrated. Ten different runs were averaged to further reduce the sampling errors. The stepsizes tested began at $\Delta t = 0.03$ and were increased incrementally by 10% until both methods showed significant relative error (SGNHT-N became unstable at around $\Delta t = 0.08$).

Thus, for configurational observables the BADODAB method has fourth order convergence to the invariant measure in the large friction and thermal mass limits (i.e. $\varepsilon \to 0$),

$$\lim_{\varepsilon \to 0} \langle \phi(q) \rangle_{\mathrm{BADODAB}} = \langle \phi(q) \rangle + O(\Delta t^4)\,.$$

It should be emphasized here that only the BADODAB and BAODOAB methods appear to have the superconvergence property among a number of different splitting methods investigated in the Ad-Langevin/SGNHT system (2.7) with a clean gradient. The superconvergence property suggests to use relatively large σ_{A} and $\mu \propto \sigma_{\mathrm{A}}^2$ in the BADODAB (SGNHT-S) method in order to enhance sampling accuracy. In fact we expect that larger values of μ than this bound will not diminish the sampling accuracy, but the effect of large values of μ is to reduce the responsiveness of the thermostat device.

4. Numerical Experiments. In this section, we conduct a variety of numerical experiments to compare the performance of the different schemes presented in this article.

4.1. Molecular Systems. Before we compare various methods in machine learning applications (i.e. with a noisy gradient), we first demonstrate the order of convergence of various splitting methods with a clean gradient.

A popular model of an N-body system with pair interactions based on a spring with rest length (i.e. pendulum) was used, a standard if simplified model of molecular dynamics. The total potential energy of the system is defined as

$$U(\mathbf{q}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \varphi(r_{ij})\,, \tag{4.1}$$

where $r_{ij} = \|\mathbf{q}_i - \mathbf{q}_j\|$ denotes the distance between two particles i and j, and $\varphi(r_{ij})$ represents the pair potential energy

$$\varphi(r_{ij}) = \begin{cases} \dfrac{k}{2}\left(r_{ij} - r_{\mathrm{c}}\right)^2\,, & r_{ij} < r_{\mathrm{c}}\,; \\ 0\,, & r_{ij} \geq r_{\mathrm{c}}\,, \end{cases} \tag{4.2}$$

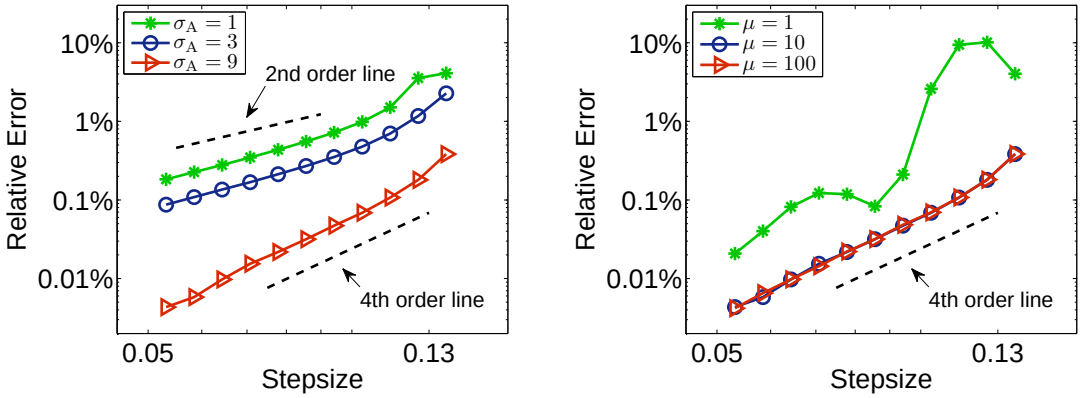

Fig. 2: Log-log plot of the relative error in computed average potential energy against stepsize by using the SGNHT-S/BADODAB method with (left) different values of σ_A ($\mu = 10$) and (right) different values of μ ($\sigma_A = 9$). The format of the plots is the same as in Fig. 1 except fifty different runs were used to reduce the sampling errors in high accuracy regime.

where k and r_c represent the spring constant and the cutoff radius, respectively.

A system consisting of $N = 500$ identical particles (i.e. unit mass) was simulated in a cubic box with periodic boundary conditions [4]. The positions of the particles were initialized on a cubic grid with equidistant grid spacing, while the initial momenta were i.i.d. random variables with mean zero and variance $k_B T$, which was set to be unity. The thermal mass μ was chosen to be 10 unless otherwise stated. Particle density $\rho_d = 4$ was used with spring constant $k = 25$ and cutoff radius $r_c = 1$.

We first compare the two SGNHT methods on controlling two configurational quantities: configurational temperature and average potential energy. The configurational temperature [22], which in principle, as the kinetic temperature, should be equal to the target temperature, can be defined as

$$k_B T = \frac{\sum_i \langle \|\nabla_i U\|^2 \rangle}{\sum_i \langle \nabla_i^2 U \rangle},$$

where the angle brackets denote the averages, $\nabla_i U$ and $\nabla_i^2 U$ represent the gradient and Laplacian of the potential energy U with respect to the position of particle i, respectively (see more discussions in [39]).

As shown in Figure 1, with the help of the dashed order lines, we can see that SGNHT-N and SGNHT-S show first and second order convergence, respectively, as expected. It is clear that not only does SGNHT-S have at least one order of magnitude improvement in accuracy in both observables, but also much greater robustness over the SGNHT-N method, which becomes completely unstable at around $\Delta t = 0.08$. The results on the configurational temperature and average potential energy are rather similar, therefore in what follows we only present average potential energy results.

We also investigate the effect of changing the value of σ_A in the SGNHT-S/BADODAB scheme proposed in this article. As can be seen from Figure 2, the SGNHT-S method displays second order convergence to the invariant measure when σ_A is relatively small, while a fourth order convergence is observed in the high "friction" limit ($\sigma_A = 9$) as anticipated from the analysis of the previous section. It should be emphasized here that the superconvergence property was only observed in the BADODAB and BAODOAB methods, which both reduce to the BAOAB method [34, 36] in Langevin dynamics.

Figure 2 also compares the effect of varying the value of the thermal mass μ when σ_A is fixed. It can be seen that the BADODAB method displays a clear fourth order convergence when μ is relatively large, while when μ is small, not only is the smooth discretization error dependence on stepsize lost but also significantly larger relative error is observed. This reinforces the choice of a relatively large

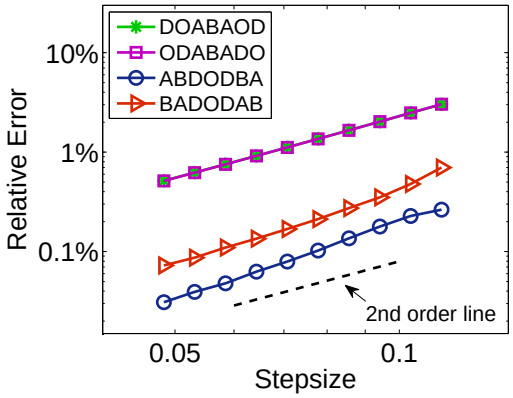

Fig. 3: Log-log plot of the relative error in computed average potential energy against stepsize by using various splitting methods of the Ad-Langevin/SGNHT system ($\sigma_A = 3$). The format of the plot is the same as in Fig. 1.

value of μ. It is worth pointing out that $\mu = 10$ works as well as $\mu = 100$, therefore $\mu = 10$ is used throughout this article since a relatively smaller μ corresponds to a tighter interaction between the thermostat and the system and thus it can fluctuate more rapidly to accommodate changes in the noise and adapt more easily.

We also explore the performance of various splitting methods of the Ad-Langevin/SGNHT system (2.7) with fixed values of σ_A and μ in Figure 3. All the methods clearly show second order convergence with ABDODBA and BADODAB methods achieving one order of magnitude improvement in accuracy compared to the other methods. This again illustrates the importance of optimal design of numerical methods. The ABDODBA method seems to be slightly better that the BADODAB method in the regime of $\sigma_A = 3$, however, as demonstrated in Figure 2 the BADODAB method achieves a dramatic improvement in accuracy when σ_A is relatively large (e.g. $\sigma_A = 9$) while other schemes remain the same except the BAODOAB method.

4.2. Bayesian Inference. In this subsection we compare methods in a classical Bayesian inference model in one dimension, i.e. to estimate the mean of a normal distribution with known variance [14]. More precisely, given N i.i.d. samples from a normal distribution, $x_i \sim \mathcal{N}(\check{\mu}, \hat{\sigma}^2)$ where it should be noted that $\check{\mu}$ is the true mean when we draw samples, with known $\hat{\sigma}^2$, and a uniform prior distribution ranging from $-N/2$ to $N/2$, we are able to calculate the posterior distribution of the mean in a closed form

$$\hat{\mu} \sim \mathcal{N}\left(\hat{x}, \frac{\hat{\sigma}^2}{N}\right), \tag{4.3}$$

where $\hat{x} = \sum_{i=1}^{N} x_i/N$. In the context of stochastic gradient approximation, we have

$$
\begin{aligned}
\pi(\hat{\mu}|\mathbf{X}) \propto \pi(\mathbf{X}|\hat{\mu})\pi(\hat{\mu}) &\approx \left(\prod_{i=1}^{\tilde{n}} \pi(\mathbf{x}_{r_i}|\hat{\mu})\right)^{\frac{N}{\tilde{n}}} \pi(\hat{\mu})\,, \\
&= \left(\frac{1}{\sqrt{2\pi}\hat{\sigma}}\right)^{N} \left[\prod_{i=1}^{\tilde{n}} \exp\left(-\frac{(x_i-\hat{\mu})^2}{2\hat{\sigma}^2}\right)\right]^{\frac{N}{\tilde{n}}} \frac{1}{N}\,, \\
&= \left(\frac{1}{\sqrt{2\pi}\hat{\sigma}}\right)^{N} \exp\left(-\frac{N}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \frac{(x_i-\hat{\mu})^2}{2\hat{\sigma}^2}\right) \frac{1}{N}\,, \\
&\propto \exp\left(-\frac{N}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \frac{(x_i-\hat{\mu})^2}{2\hat{\sigma}^2}\right)\,, \\
&= \exp\left[-\frac{1}{2\hat{\sigma}^2}\frac{N}{\tilde{n}}\left(\sum_{i=1}^{\tilde{n}}(x_i-\bar{x})^2 + \tilde{n}(\bar{x}-\hat{\mu})^2\right)\right]\,, \\
&\propto \exp\left(-\frac{N}{2\hat{\sigma}^2}(\bar{x}-\hat{\mu})^2\right)\,,
\end{aligned}
\tag{4.4}
$$

where $\bar{x} = \sum_{i=1}^{\tilde{n}} x_i/\tilde{n}$. It clearly recovers the true distribution (4.3) when $\tilde{n} = N$. Taking the logarithm and differentiating the posterior distribution obtained at the end of (4.4) with respect to $\hat{\mu}$ gives the noisy force

$$
\tilde{F}(\hat{\mu}) = \frac{N}{\hat{\sigma}^2}\left(\hat{\mu} - \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} x_i\right)\,. \tag{4.5}
$$

In this simple case, the noise of the stochastic gradient is independent of $\hat{\mu}$ and is a constant given $\tilde{n}$. Moreover, we are able to obtain its mean and variance with respect to the stochastic gradient [24,62]

$$
\begin{aligned}
\mathbb{E}\tilde{F}(\hat{\mu}) = F(\hat{\mu}) &= \frac{N}{\hat{\sigma}^2}\left(\hat{\mu} - \frac{1}{N}\sum_{i=1}^{N} x_i\right)\,, \\
\mathrm{Var}\tilde{F}(\hat{\mu}) &= \frac{1}{\hat{\sigma}^4}\frac{N(N-1)}{\tilde{n}}\mathrm{Var}\mathbf{X}\,,
\end{aligned}
\tag{4.6}
$$

where $\mathrm{Var}\mathbf{X}$ is the variance of the dataset. Thus, it is straightforward to verify that the noise is normally distributed according to the central limit theorem.

In our numerical experiments, σ_{A} was chosen as 1 due to the fact that large σ_{A} results in stability issues here. We generated $N = 100$ samples from $\mathcal{N}(0,1)$ and randomly select a subset of size $\tilde{n} = 10$ at each timestep to compute the noisy force (4.5). We plot the distributions of the posterior mean of the dataset obtained by using four different methods with different stepsizes in Figure 4. Clearly, two SGNHT methods completely outperformed SGLD and mSGLD methods. The latter only demonstrate good approximation of the true distribution with order of magnitude smaller stepsize compared to the former. But it should be noted that mSGLD here is slightly better than SGLD in maintaining the true distribution: the distribution of mSGLD with $\Delta t = 0.001$ is visibly much closer to the target compared to that of SGLD with the same stepsize.

Note that stepsizes for SGNHT (2nd order dynamics) and SGLD (1st order dynamics) based methods are not directly comparable—as mentioned in [34] the stepsize of a first order dynamics method like Euler-Maruyama when viewed as the limiting discretization of a Langevin integrator corresponds to $\Delta t^2/2$ where Δt is the stepsize of the Langevin method. However, in our experiments we are uninterested in the time-dynamics of the system, and only care about the invariant measure. Therefore the important relationship is the error in thermodynamic averages in comparison with the

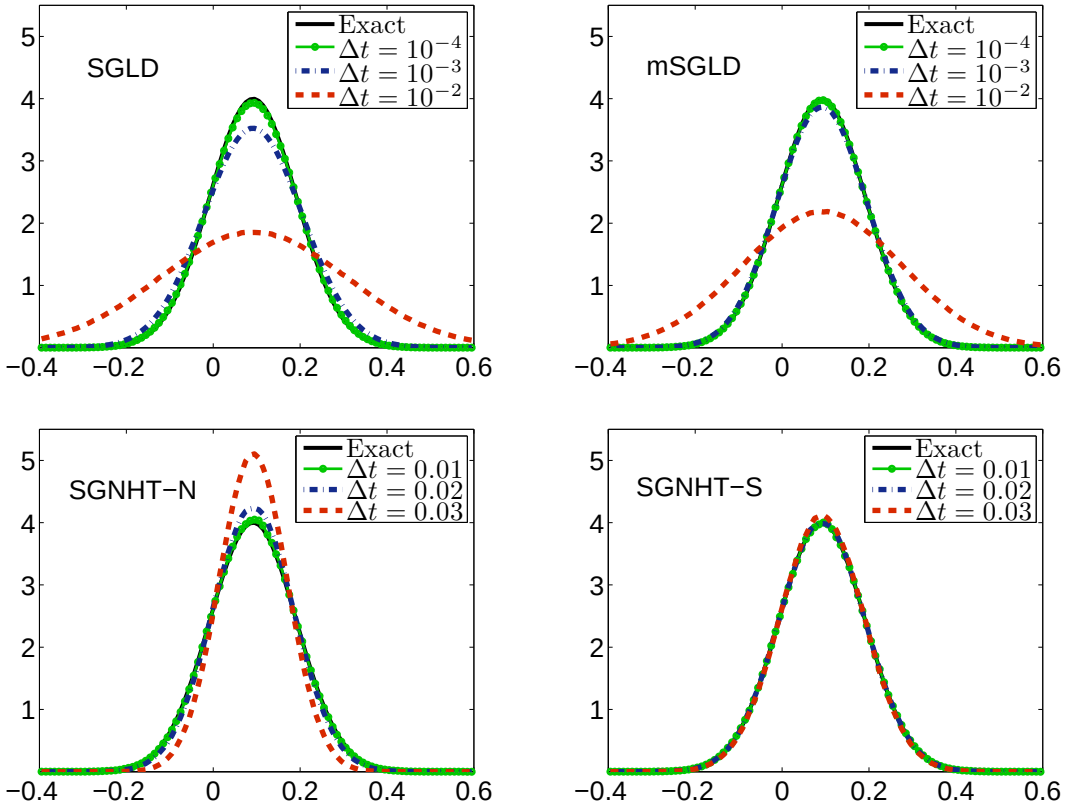

Fig. 4: Comparisons of the distribution in a one-dimensional Bayesian inference problem by using SGLD (top left), mSGLD (top right), SGNHT-N (bottom left) and SGNHT-S (bottom right) with different stepsizes indicated by different colors. The black solid line is the exact solution. Note the difference in the legends between rows.

number of timesteps (work), which quantifies the efficiency of a given method. The stepsize is just an arbitrary parameter which allows for refinement of the statistical calculation

Between the two SGNHT methods, SGNHT-S (the new scheme being proposed here) is obviously superior to SGNHT-N: the latter starts to show significant deviation to the true distribution at $\Delta t = 0.02$ while the distribution of the former still looks well matched to the true one at $\Delta t = 0.03$. Our observations are confirmed by Figure 5, where the mean absolute error (MAE) of the distribution of the two SGNHT methods were plotted. The MAE, which can be thought of as a relative error in distribution, is defined as

$$\text{MAE} = \frac{1}{\bar{N}} \sum_{i=1}^{\bar{N}} |\omega_i - \hat{\omega}_i| \,, \tag{4.7}$$

where $\bar{N}$ denotes the number of intervals, which was chosen as 100. ω_i and $\hat{\omega}_i$ represent the observed frequency in bin i and the exact expected frequency, respectively [34]. As can be seen, the stability threshold of SGNHT-N was around $\Delta t = 0.03$, beyond which the system became unstable as highlighted in the figure (in which case the system blew up, resulting in a 100% MAE). Once again, SGNHT-S not only shows an order of magnitude better accuracy but also has a much greater robustness than SGNHT-N. In particular, for defined accuracy, the SGNHT-S method is able to use double the stepsize

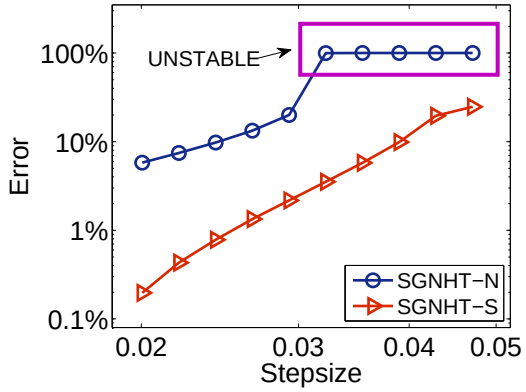

Fig. 5: Log-log plot of the mean absolute error in the distribution of the Bayesian inference model against stepsize. The box indicates that the system was unstable with corresponding stepsizes for the SGNHT-N method.

compared to SGNHT-N, which means a remarkable 50% improvement in overall numerical efficiency as defined in [39].

4.3. Bayesian Logistic Regression. Following [62], we also investigate the performance of different methods for a more complicated Bayesian logistic regression model. The data $y_i \in \{-1, 1\}$ were modelled by

$$\pi(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = f(y_i \boldsymbol{\beta}^T \mathbf{x}_i)\,,$$

where $f(z) = 1/(1+\exp(-z)) \in [0,1]$ is the logistic function and $\mathbf{x}_i \in \mathbb{R}^d$ are rows of a fixed dataset. Our goal is to estimate the posterior mean of parameter vector $\boldsymbol{\beta} \in \mathbb{R}^d$. For simplicity, a multivariate Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ was used on $\boldsymbol{\beta}$. Therefore, by using Bayes' theorem, we obtain the following posterior distribution

$$\pi(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{\beta}\|^2\right) \prod_{i=1}^{N} f(y_i \boldsymbol{\beta}^T \mathbf{x}_i)\,. \tag{4.8}$$

Following the same procedure in the Bayesian inference example (Section 4.2), we can calculate the noisy force and then plug it into different thermostats for sampling.

In our numerical experiments, we considered the $d = 3$ case with $N = 1000$ data points. We chose the dataset to be

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & 1 \\ x_{2,1} & x_{2,2} & 1 \\ \vdots & \vdots & \vdots \\ x_{1000,1} & x_{1000,2} & 1 \end{pmatrix} \tag{4.9}$$

where $x_{i,j}$ were sampled from a standard normal distribution $\mathcal{N}(0,1)$ for $i = 1, \ldots, 1000$ and $j = 1, 2$. A subset of size $\tilde{n} = 100$ was randomly chosen at each timestep to compute the noisy force.

The performance of estimating the posterior mean value of parameter vector $\boldsymbol{\beta}$ by various methods ($\sigma_{\mathrm{A}} = 6$) was tested and plotted in Figure 6. Again, SGLD and mSGLD, displaying considerable larger root mean square error (RMSE) with a fixed stepsize, were outperformed by the two SGNHT methods. In this case, the SGLD and mSGLD methods demonstrate similar control in numerical accuracy, but the latter displays much worse stability than that of the former and became unstable just above

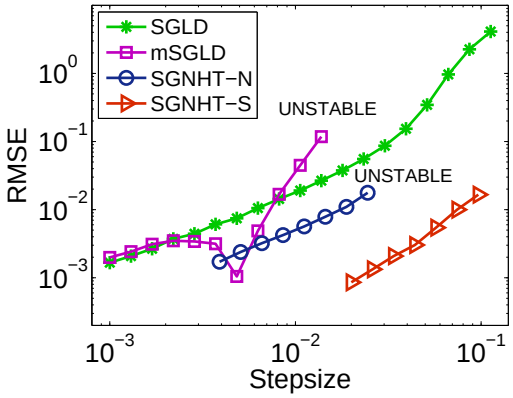

Fig. 6: Comparisons of the root mean square error (RMSE) of the posterior mean in the Bayesian logistic regression model by using various methods against stepsize. The system was simulated for 1000 reduced time units with 100,000 different runs. The stepsizes tested began at $\Delta t = 0.001$ and were increased incrementally by 30% until all methods either displayed significant error or became unstable (mSGLD and SGNHT-N).

$\Delta t = 0.01$. As reported in the original paper [62], the performance of the mSGLD method depends strongly on the size of the subset–for a larger subset, which requires higher computational cost, the bias of mSGLD can be smaller than that of SGLD.

Of the two SGNHT methods, the SGNHT-S method again shows not only at least an order of magnitude improvement on accuracy but also much better robustness than the other: SGNHT-N became unstable just above $\Delta t = 0.02$. Remarkably, the SGNHT-S method at $\Delta t = 0.1$ still achieves better accuracy than the SGLD method at $\Delta t = 0.01$. In other words, the method we propose here gives more than a 90% improvement in overall numerical efficiency compared to one of the most popular methods in the literature. For fixed accuracy, the SGNHT-S method can use almost four times the stepsize of the SGNHT-N method (i.e. an about 75% improvement in overall numerical efficiency).

5. Conclusions. We have reviewed a variety of methods in stochastic gradient systems with applications in machine learning. We have provided a theoretical discussion on the foundation (underlying dynamics) of those stochastic gradient systems, which has been lacking in the literature. We have also proposed a new symmetric splitting (at least second order) method in SGNHT (SGNHT-S/BADODAB), which substantially improves the accuracy and robustness compared to a nonsymmetric splitting (first order) method (SGNHT-N) proposed recently in the literature. Furthermore, we have demonstrated that under certain conditions the SGNHT-S/BADODAB method can inherit the superconvergence property recently discovered in integrators for Langevin dynamics, i.e. fourth order convergence to the invariant measure for configurational averages.

By conducting various numerical experiments, we have demonstrated that the two SGNHT methods outperform the popular SGLD method and its variant mSGLD. In particular, the SGNHT-S method can use up to ten times the stepsize of SGLD, which implies a remarkable more than 90% improvement in overall numerical efficiency. Between the two SGNHT methods, the SGNHT-S method can use almost four times the stepsize of SGNHT-N for defined accuracy (i.e. about 75% improvement in overall numerical efficiency).

It should be noted that, in certain cases, it may be desirable to employ a Metropolis-Hastings procedure in order to remove the discretization bias [54]. However, we emphasize that the correction is not without computational cost, particularly as the dimension is increased [6, 28, 52, 53], and the results of [34–36] and of the current article demonstrate that high accuracy with respect to the invariant distribution is often achievable using traditional numerical integration techniques, thus in many cases entirely eliminating the necessity of Metropolis-Hastings corrections (see more discussions in [36]).

Moreover, we mention that the methods of this article can in principle be combined with Metropolis-Hastings algorithms if it is necessary to completely eliminate the discretization bias.

Acknowledgements. The authors thank Ben Goddard, Charles Matthews, Tony Shardlow, Zhanxing Zhu, and Konstantinos Zygalakis for stimulating discussions and valuable suggestions. The authors further thank anonymous referees for their comments, which substantially contributed to the presentation of our results.

REFERENCES

[1] A. Abdulle, G. Vilmart, and K. C. Zygalakis, *High order numerical approximation of the invariant measure of ergodic SDEs*, SIAM Journal on Numerical Analysis, 52 (2014), pp. 1600–1622.
[2] ———, *Long time accuracy of Lie-Trotter splitting methods for Langevin dynamics*, SIAM Journal on Numerical Analysis, 53 (2015), pp. 1–16.
[3] S. Ahn, A. Korattikara, and M. Welling, *Bayesian posterior sampling via stochastic gradient Fisher scoring*, in Proceedings of the 29th International Conference on Machine Learning, 2012, pp. 1591–1598.
[4] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, 1989.
[5] R. B. Ash, *Basic Probability Theory*, Dover Publications, 2008.
[6] A. Beskos, N. Pillai, G. Roberts, J.-M. Sanz-Serna, and A. Stuart, *Optimal tuning of the hybrid Monte Carlo algorithm*, Bernoulli, 19 (2013), pp. 1501–1534.
[7] F. A. Bornemann, P. Nettesheim, and C. Schütte, *Quantum-classical molecular dynamics as an approximation to full quantum dynamics*, The Journal of Chemical Physics, 105 (1996), pp. 1074–1083.
[8] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, *Handbook of Markov Chain Monte Carlo*, CRC Press, 2011.
[9] E. Cancès, F. Legoll, and G. Stoltz, *Theoretical and numerical comparison of some sampling methods for molecular dynamics*, ESAIM: M2AN, 41 (2007), pp. 351–389.
[10] S. Chakravorty, *A homotopic Galerkin approach to the solution of the Fokker-Planck-Kolmogorov equation*, in Proceedings of the 2006 American Control Conference, IEEE, 2006.
[11] T. Chen, E. B. Fox, and C. Guestrin, *Stochastic gradient Hamiltonian Monte Carlo*, in Proceedings of the 31st International Conference on Machine Learning, 2014, pp. 1683–1691.
[12] M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, arXiv preprint arXiv:1302.6989, (2013).
[13] A. Debussche and E. Faou, *Weak backward error analysis for SDEs*, SIAM Journal on Numerical Analysis, 50 (2012), pp. 1735–1752.
[14] N. Ding, Y. Fang, R. Babbush, C. Chen, R. D. Skeel, and H. Neven, *Bayesian sampling using stochastic gradient thermostats*, in Advances in Neural Information Processing Systems 27, 2014, pp. 3203–3211.
[15] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, *Hybrid Monte Carlo*, Physics Letters B, 195 (1987), pp. 216–222.
[16] S. Dubinkina, J. Frank, and B. Leimkuhler, *Simplified modelling of a thermal bath, with application to a fluid vortex system*, Multiscale Modeling & Simulation, 8 (2010), pp. 1882–1901.
[17] D. A. Fedosov and G. E. Karniadakis, *Triple-decker: Interfacing atomistic-mesoscopic-continuum flow regimes*, Journal of Computational Physics, 228 (2009), pp. 1157–1171.
[18] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications, Second Edition*, Academic Press, 2001.
[19] D. Givon, R. Kupferman, and A. Stuart, *Extracting macroscopic dynamics: model problems and algorithms*, Nonlinearity, 17 (2004), pp. R55–R127.
[20] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer, 2006.
[21] C. Hijón, P. Español, E. Vanden-Eijnden, and R. Delgado-Buscalioni, *Mori-Zwanzig formalism as a practical computational tool*, Faraday discussions, 144 (2010), pp. 301–322.
[22] J. O. Hirschfelder, *Classical and quantum mechanical hypervirial theorems*, The Journal of Chemical Physics, 33 (1960), p. 1462.
[23] A. M. Horowitz, *A generalized guided Monte Carlo algorithm*, Physics Letters B, 268 (1991), pp. 247–252.
[24] D. G. Horvitz and D. J. Thompson, *A generalization of sampling without replacement from a finite universe*, Journal of the American Statistical Association, 47 (1952), pp. 663–685.
[25] P. H. Hünenberger, *Thermostat algorithms for molecular dynamics simulations*, Advances in Polymer Science, 173 (2005), pp. 105–149.
[26] A. Jones and B. Leimkuhler, *Adaptive stochastic methods for sampling driven molecular systems*, The Journal of Chemical Physics, 135 (2011), p. 084125.
[27] E. E. Keaveny, I. V. Pivkin, M. Maxey, and G. E. Karniadakis, *A comparative study between dissipative particle dynamics and molecular dynamics for simple-and complex-geometry flows*, The Journal of Chemical Physics, 123 (2005), p. 104107.
[28] A.D. Kennedy and B. Pendleton, *Acceptances and autocorrelations in hybrid Monte Carlo*, Nuclear Physics B - Proceedings Supplements, 20 (1991), pp. 118–121.

[29] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer, 1992.
[30] F. Legoll, M. Luskin, and R. Moeckel, *Non-ergodicity of the Nosé-Hoover thermostatted harmonic oscillator*, Arch Rational Mech Anal, 184 (2006), pp. 449–463.
[31] ———, *Non-ergodicity of Nosé-Hoover dynamics*, Nonlinearity, 22 (2009), pp. 1673–1694.
[32] B. Leimkuhler, *Generalized Bulgac-Kusnezov methods for sampling of the Gibbs-Boltzmann measure*, Physical Review E, 81 (2010), p. 026703.
[33] B. Leimkuhler, D. T. Margul, and M. E. Tuckerman, *Stochastic, resonance-free multiple time-step algorithm for molecular dynamics with very large time steps*, Molecular Physics, 111 (2013), pp. 3579–3594.
[34] B. Leimkuhler and C. Matthews, *Rational construction of stochastic numerical methods for molecular sampling*, Applied Mathematics Research eXpress, 2013 (2013), pp. 34–56.
[35] ———, *Robust and efficient configurational molecular sampling via Langevin dynamics*, The Journal of Chemical Physics, 138 (2013), p. 174102.
[36] B. Leimkuhler, C. Matthews, and G. Stoltz, *The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics*, IMA Journal of Numerical Analysis, (2015).
[37] B. Leimkuhler, E. Noorizadeh, and F. Theil, *A gentle stochastic thermostat for molecular dynamics*, Journal of Statistical Physics, 135 (2009), pp. 261–277.
[38] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge University Press, 2005.
[39] B. Leimkuhler and X. Shang, *On the numerical treatment of dissipative particle dynamics and related systems*, Journal of Computational Physics, 280 (2015), pp. 72–95.
[40] Z. Li, X. Bian, B. Caswell, and G. E. Karniadakis, *Construction of dissipative particle dynamics models for complex fluids via the Mori-Zwanzig formulation*, Soft Matter, 10 (2014), pp. 8659–8672.
[41] M. Lísal and J. K. Brennan, *Alignment of lamellar diblock copolymer phases under shear: Insight from dissipative particle dynamics simulations*, Langmuir, 23 (2007), pp. 4809–4818.
[42] J.C. Mattingly, A.M. Stuart, and D.J. Higham, *Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise*, Stochastic Processes and their Applications, 101 (2002), pp. 185–232.
[43] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of state calculations by fast computing machines*, The Journal of Chemical Physics, 21 (1953), p. 1087.
[44] L. Mones, A. Jones, A. W. Götz, T. Laino, R. C. Walker, B. Leimkuhler, G. Csányi, and N. Bernstein, *The adaptive buffered force QM/MM method in the CP2K and AMBER software packages*, Journal of Computational Chemistry, 36 (2015), pp. 633–648.
[45] C. Pastorino, T. Kreer, M. Müller, and K. Binder, *Comparison of dissipative particle dynamics and Langevin thermostats for out-of-equilibrium simulations of polymeric systems*, Phys. Rev. E, 76 (2007).
[46] S. Patterson and Y. W. Teh, *Stochastic gradient Riemannian Langevin dynamics on the probability simplex*, in Advances in Neural Information Processing Systems 26, 2013, pp. 3102–3110.
[47] B. L. Peters, A. Ramírez-Hernández, D. Q. Pike, M. Müller, and J. J. de Pablo, *Nonequilibrium simulations of lamellae forming block copolymers under steady shear: A comparison of dissipative particle dynamics and Brownian dynamics*, Macromolecules, 45 (2012), pp. 8109–8116.
[48] M. Praprotnik, L. Delle Site, and K. Kremer, *Adaptive resolution molecular-dynamics simulation: Changing the degrees of freedom on the fly*, The Journal of Chemical Physics, 123 (2005), p. 224106.
[49] ———, *Multiscale simulation of soft matter: From scale bridging to adaptive resolution*, Annu. Rev. Phys. Chem., 59 (2008), pp. 545–571.
[50] H. Risken, *Fokker-Planck Equation: Methods of Solution and Applications, Second Edition*, Springer, 1989.
[51] C. Robert and G. Casella, *Monte Carlo Statistical Methods, Second Edition*, Springer, 2004.
[52] G. O. Roberts, A. Gelman, and W. R. Gilks, *Weak convergence and optimal scaling of random walk Metropolis algorithms*, The Annals of Applied Probability, 7 (1997), pp. 110–120.
[53] G. O. Roberts and J. S. Rosenthal, *Optimal scaling of discrete approximations to Langevin diffusions*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60 (1998), pp. 255–268.
[54] G. O. Roberts and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363.
[55] A. A. Samoletov, C. P. Dettmann, and M. A. J. Chaplain, *Thermostats for "slow" configurational modes*, Journal of Statistical Physics, 128 (2007), pp. 1321–1336.
[56] ———, *Notes on configurational thermostat schemes*, The Journal of Chemical Physics, 132 (2010), p. 246101.
[57] X. Shang, Z. Zhu, B. Leimkuhler, and A. J. Storkey, *Covariance-controlled adaptive Langevin thermostat for large-scale Bayesian sampling*, in Advances in Neural Information Processing Systems 28, 2015, pp. 37–45.
[58] T. Soddemann, B. Dünweg, and K. Kremer, *Dissipative particle dynamics: A useful thermostat for equilibrium and nonequilibrium molecular dynamics simulations*, Physical Review E, 68 (2003), p. 046702.
[59] L. Stella, C. D. Lorenz, and L. Kantorovich, *Generalized Langevin equation: An efficient approach to nonequilibrium molecular dynamics of open systems*, Physical Review B, 89 (2014), p. 134303.
[60] D. Talay and L. Tubaro, *Expansion of the global error for numerical schemes solving stochastic differential equations*, Stochastic Analysis and Applications, 8 (1990), pp. 483–509.
[61] Y. W. Teh, A. Thiéry, and S. Vollmer, *Consistency and fluctuations for stochastic gradient Langevin dynamics*, arXiv preprint arXiv:1409.0578, (2014).
[62] S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh, *(Non-) asymptotic properties of stochastic gradient Langevin dynamics*, arXiv preprint arXiv:1501.00438, (2015).
[63] M. Welling and Y. W. Teh, *Bayesian learning via stochastic gradient Langevin dynamics*, in Proceedings of the

28th International Conference on Machine Learning, 2011, pp. 681–688.