




OPTIMAL FRICTION MATRIX FOR UNDERDAMPED LANGEVIN SAMPLING

MARTIN CHAK^{1,*}, NIKOLAS KANTAS¹, TONY LELIÈVRE^{2,3} AND GRIGORIOS A. PAVLIOTIS¹

Abstract. We propose a procedure for optimising the friction matrix of underdamped Langevin dynamics when used for continuous time Markov Chain Monte Carlo. Starting from a central limit theorem for the ergodic average, we present a new expression of the gradient of the asymptotic variance with respect to friction matrix. In addition, we present an approximation method that uses simulations of the associated first variation/tangent process. Our algorithm is applied to a variety of numerical examples such as toy problems with tractable asymptotic variance, diffusion bridge sampling and Bayesian inference problems for high dimensional logistic regression.

Mathematics Subject Classification. 60J25, 60J60.

Received August 25, 2022. Accepted October 3, 2023.

1. INTRODUCTION

Let π be a probability measure on \mathbb{R}^n with smooth positive bounded density with respect to the Lebesgue measure, also denoted π and such that $\pi \propto e^{-U}$, where $U : \mathbb{R}^n \rightarrow \mathbb{R}$ is the associated smooth potential or negative log density. In a range of applications including statistics, molecular dynamics, engineering, finance, machine learning to name a few, an important quantity of interest is the expectation of $f \in L^2(\pi)$ with respect to π ,

$$\pi(f) := \int f \, d\pi,$$

which is in most cases intractable and numerically approximated most commonly by Markov Chain Monte Carlo (MCMC) methods. Here, $f \in L^2(\pi)$ is referred to as an observable or a test function. In this paper, we consider the setting where the Markov process used in MCMC is an approximation of underdamped Langevin dynamics and we focus our analysis in the continuous time setting. Denoting \mathbb{S}_{++}^n as the set of real symmetric $n \times n$ positive definite matrices, the underdamped Langevin dynamics with mass $M \in \mathbb{S}_{++}^n$ and friction matrix $\Gamma \in \mathbb{S}_{++}^n$ is given by the \mathbb{R}^{2n} -valued solution (q_t, p_t) to

$$dq_t = M^{-1}p_t \, dt \tag{1.1a}$$

$$dp_t = -\nabla U(q_t) - \Gamma M^{-1}p_t \, dt + \sqrt{2\Gamma}dW_t, \tag{1.1b}$$

Keywords and phrases. Asymptotic variance, self-tuning algorithm, Langevin dynamics, variance reduction, Poisson equation.

¹ Department of Mathematics, Imperial College London, London, UK.

² CERMICS, École des Ponts, Champs-sur-Marne, France.

³ MATHERIALS, Inria Paris, Paris, France.

*Corresponding author: martin.chak@sorbonne-universite.fr

where W_t denotes a standard Wiener process on \mathbb{R}^n and $\sqrt{\Gamma} \in \mathbb{R}^{n \times n}$ is any matrix satisfying $\sqrt{\Gamma}\sqrt{\Gamma}^\top = \Gamma$. Under general conditions, the probability distribution of the solution to underdamped Langevin dynamics (1.1) converges to a unique invariant probability measure given by

$$\tilde{\pi}(\mathrm{d}q, \mathrm{d}p) = Z^{-1} e^{-U(q) - \frac{p^\top M^{-1} p}{2}} \mathrm{d}q \mathrm{d}p, \quad (1.2)$$

where $Z = \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} e^{-U(q) - \frac{p^\top M^{-1} p}{2}} \mathrm{d}q \mathrm{d}p$ is the normalising constant. Since the marginal distribution of $\tilde{\pi}$ in the q variable is π , the expression

$$\pi_T(f) := \frac{1}{T} \int_0^T f(q_t) \mathrm{d}t \quad (1.3)$$

is used to approximate $\pi(f)$.

In this paper, we will focus mainly on the effect of the choice of friction matrix Γ on the efficiency of the estimator $\pi_T(f)$. In the literature, authors have used crude methods to tune Γ . For example, multiple chains with different values of scalar $\Gamma > 0$ have been numerically simulated and subsequently compared in terms of the empirical correlations or other observation criteria, see *e.g.* [2, 11, 44]. This is an expensive and cumbersome process, which calls for systematic approaches to design Γ . However, in order to design Γ , it is important to fix a criterion to optimise. Possible choices of criterion include the spectral gap [53], rate of convergence in Wasserstein distance [19], autocorrelation or integrated autocorrelation time (IACT) [41, 67], or the asymptotic variance [23, 24]. Different criteria may lead to different optimal friction matrices Γ ; we show this in Section 1.2 below by looking at cases where π is Gaussian. First, we present our approach in the following.

Out of the possible criteria, we aim to optimise Γ with respect to the *asymptotic variance* in the convergence of $\pi_T(f)$ to $\pi(f)$ as $T \rightarrow \infty$ for any particular observable(s) f in a wide class of observables that depend only on q . In what follows, we will concentrate on the asymptotic variance for any given observable, but this can be extended trivially to multiple observables using for example the sum of asymptotic variances for the different observables. The main ideas for the procedure to optimise Γ are given in Section 1.1. To the best of our knowledge, the present work constitutes the first systematic procedure for choosing the friction in an optimal manner and the first result on optimal Γ in the space of matrices. Finally, we note that for practical implementation, one needs to discretise the dynamics in time. There are different discretisation methods, the convergence properties of which have been studied actively in numerous recent works, see *e.g.* [16, 19, 26, 29, 38, 48, 54, 64]. Since the main question in the present work is on the choice of Γ , we will focus solely on the continuous time dynamics and leave development and investigations to discretisations as future work.

1.1. Outline of approach

We proceed with a informal description of our approach, precise statements can be found in Theorems 3.2 and 3.3. It is known that under suitable assumptions on U and f , a central limit theorem

$$\frac{1}{\sqrt{T}} \int_0^T (f(q_t) - \pi(f)) \mathrm{d}t \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2) \quad \text{as } T \rightarrow \infty \quad (1.4)$$

holds (see for example [12]) and that σ^2 , the asymptotic variance [23], has the form

$$\sigma^2 = 2 \int \phi(f - \pi(f)) \mathrm{d}\tilde{\pi}, \quad (1.5)$$

where ϕ is a solution to the Poisson equation

$$-L\phi = f - \pi(f) \quad (1.6)$$

and L denotes the infinitesimal generator associated to (1.1) (precise statements based on [9, 63] will follow below). Two key observations are then made. The first and main observation is an explicit formula for the

direction derivative (a definition will be presented in Sect. 3.1) of σ^2 with respect to any direction in Γ . For any $\Gamma \in \mathbb{S}_{++}^n$ and direction $\delta\Gamma \in \mathbb{R}^{n \times n}$, the derivative of σ^2 at Γ in the direction $\delta\Gamma$, denoted $d\sigma^2.\delta\Gamma$, is shown to be given as follows

$$d\sigma^2.\delta\Gamma = -2 \int (\nabla_p \phi)^\top \delta\Gamma \nabla_p \tilde{\phi} d\tilde{\pi}, \quad (1.7)$$

where ϕ is the solution to (1.6) at Γ and $\tilde{\phi}$ is given by

$$\tilde{\phi}(q, p) = \phi(q, -p). \quad (1.8)$$

A direction that guarantees a decrease in σ^2 is then $\delta\Gamma = \Delta\Gamma$ defined by

$$\Delta\Gamma := \int \nabla_p \phi \otimes \nabla_p \tilde{\phi} d\tilde{\pi}, \quad (1.9)$$

where \otimes denotes the outer product. Note in addition, one could also take either $\delta\Gamma$ to be the diagonal elements of (1.9) or $\delta\Gamma = I_n \int \nabla_p \phi \cdot \nabla_p \tilde{\phi} d\tilde{\pi}$, since both cases give a negative variation in asymptotic variance.

The second observation is on the form of $\nabla_p \phi$. The solution of the Poisson equation in (1.6) admits the representation (see for example [51], Cor. 2.21)

$$\phi(q, p) = \int_0^\infty \mathbb{E}[f(q_t) - \pi(f)] dt, \quad (1.10)$$

where (q_t, p_t) solves (1.1) with initial condition $(q_0, p_0) = (q, p)$. Under convexity of the potential U and other suitable assumptions, we have

$$\nabla_p \phi(q, p) = \int_0^\infty \mathbb{E}[\nabla f(q_t)^\top D_p q_t] dt, \quad (1.11)$$

where $D_p q_t$ denotes the $\mathbb{R}^{n \times n}$ -matrix made of partial derivatives of q_t with respect to the initial velocity p_0 . The process $D_p q_t$ is in fact known to satisfy a stochastic differential equation (SDE). The SDE satisfied by $D_p q_t$ is obtained by simultaneously taking partial derivatives in all of the terms of (1.1); the form of the SDE is given below in (3.3). Moreover, this SDE may be simulated numerically, which allows us to approximate (1.11) and subsequently (1.9) by using any suitable discretisation of the SDE. From a numerical viewpoint, the advantage of considering (1.11) rather than a finite difference derivative of the right-hand side of (1.10) is that we may use theoretical properties of $D_p q_t$ to inform us of the behaviour of numerical approximations to (1.11). In particular, we will show that $D_p q_t$ decays to zero exponentially quickly in time given convexity assumptions on U , so that the infinite time integral (1.11) can be accurately approximated with a finite time integral using only short simulations of $D_p q_t$. More details are given in Section 3.2.

Since the SDE satisfied by $D_p q_t$ may be simulated alongside (1.1), the expressions (1.10) and (1.11) enable the direction $\Delta\Gamma$ in (1.9) to be estimated “on the fly” in parallel to (1.1) and leads to an adaptive MCMC algorithm (see [3] for a review) where Γ is adaptively updated over time in (1.1). Note that in order to save on computational cost, it is also possible to approximate (1.11) and (1.9) for only an initial period, then to use the resulting Γ to simulate just (1.1) in the remaining computational effort. We will apply this algorithm to different simulation problems and show that in some cases this leads to a significant performance improvement. In particular, for posterior mean estimation in logistic regression for two datasets of hundreds of dimensions, improved friction matrices were found to reduce the Monte Carlo variance by almost an order of magnitude compared to the default choice of $\Gamma = I_n$; details will be presented in Section 5, see in particular Tables 2 and 3.

1.2. On the choice of criterion

We illustrate here on a simple example the difference of choosing σ^2 compared to autocorrelation time or other equivalent performance measures considered in [2, 11, 19, 41, 44, 67] and note that they can be conflicting goals. Following the early work [41], consider the problem of optimal friction for scalar Gaussian distribution π

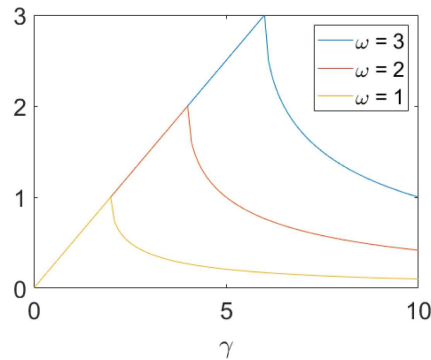


FIGURE 1. $\min_i(|\operatorname{Re}(\lambda_i)|)$ for different values of γ , where λ_i are the eigenvalues of the matrix appearing in (1.12), also the spectral gap for the generator of (1.1) with $n = 1$, $\omega > 0$, $U(q) = \frac{1}{2}\omega^2 q^2$, $M = 1$ and $\Gamma = \gamma$. Critical values of γ are given by 2ω .

in case that the autocorrelation time is used as criterion. Let $n = 1$, $\omega, \gamma > 0$, $U(q) = \frac{1}{2}\omega^2 q^2$, $M = 1$, $\Gamma = \gamma$, then the autocorrelation functions for (1.1) satisfy

$$\frac{d}{dt} \begin{pmatrix} \mathbb{E}(q_t q_0) \\ \mathbb{E}(p_t q_0) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\omega^2 & -\gamma \end{pmatrix} \begin{pmatrix} \mathbb{E}(q_t q_0) \\ \mathbb{E}(p_t q_0) \end{pmatrix}. \quad (1.12)$$

By considering the eigenvalues of the 2-by-2 matrix appearing on the right-hand side of (1.12), the conclusion in [42] is that the optimal γ for minimising the magnitude of $\mathbb{E}(q_t q_0)$ is given by the critical damping $\gamma = 2\omega$, see Figure 1. A similar conclusion can be made when considering the spectral gap [60]. On the other hand, for the observable f given by $f(q) = q$, a formal calculation gives $\sigma^2 = 2 \int_0^\infty \mathbb{E}(q_t q_0) dt d\tilde{\pi}(q_0, p_0)$ due to (1.5) and (1.10). Despite the appearance of $\mathbb{E}(q_t q_0)$ as before, our results assert that $\gamma = 0$ is optimal for the asymptotic variance (see Cor. 4.9 and discussion below for more precise statements). This highlights the discrepancy between prioritising the asymptotic variance for observables and prioritising the autocorrelation time (similarly, the spectral gap). Of course, the central limit theorem criterion is dependent on the test function f , but we emphasise that multiple asymptotic variances can be used to construct a composite objective function to minimise, so that Γ can be optimised with respect to several observables of interest simultaneously. In particular, Remark B.1 describes the implementation for a linear combination of asymptotic variances at no extra cost in terms of evaluations of π or its gradients.

1.3. Outline of the paper

The rest of the paper is organised as follows. In Section 2, we provide a mathematical setting in which the underdamped Langevin dynamics with a friction matrix and in particular (1.1) has a well-defined solution and satisfies the central limit theorem for suitable observables, together with notations used throughout the paper. In Section 3, prerequisite results and the main formulae (1.7) and (1.11) are precisely stated. The main results of this section and of this paper are Theorems 3.2 and 3.3. In Section 4, we study mainly the case where U is quadratic and f is polynomial up to fourth order. In Section 5, numerical methods in approximating (1.9) together with an adaptive algorithm in Algorithm 1 resulting from (1.7) and (1.11) is outlined, alongside examples of U and f where improvements in variance are observed. In Section 6, deferred proofs are given. In Section 7, we conclude and discuss future work.

2. SETTING

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, $(\mathcal{F}_t)_{t \geq 0}$ be a normal (satisfying the usual conditions) filtration with $(W_t)_{t \geq 0}$ a standard Wiener process on \mathbb{R}^n with respect to $(\mathcal{F}_t)_{t \geq 0}$, $\tilde{\pi}$ be a probability measure given by (1.2) for some potential function $U : \mathbb{R}^n \rightarrow \mathbb{R}$ and mass matrix $M \in \mathbb{S}_{++}^n$.

The set of smooth compactly supported functions is denoted C_c^∞ . Following the notation of [27], we denote the infinitesimal generator (see (A.5) for a definition) associated to (1.1) as L , which is given formally by its differential operator form, denoted \mathcal{L} , when acting on the subset $C_c^\infty(\mathbb{R}^{2n})$,

$$\mathcal{L} = p^\top M^{-1} \nabla_q - \nabla U(q)^\top \nabla_p - p^\top M^{-1} \Gamma \nabla_p + \nabla_p^\top \Gamma \nabla_p. \quad (2.1)$$

Its formal $L^2(\mathbb{R}^n)$ -adjoint \mathcal{L}^\top satisfies

$$\mathcal{L}^\top \tilde{\pi} = 0, \quad (2.2)$$

so that $\tilde{\pi}$ (see (1.2)) is an invariant probability measure for (1.1) for a normalisation constant Z . Let

$$L_0^2(\pi) := \left\{ g \in L^2(\pi) : \int g \, d\pi = 0 \right\}$$

and similar for $L_0^2(\tilde{\pi})$. The notation D^2U will be used for the Hessian matrix of U . As in the introduction, $I_n \in \mathbb{R}^{n \times n}$ denotes the identity matrix. For a matrix A , $|A|$ denotes the operator norm associated with the Euclidean norm. The notation e_i is used to denote the i th Euclidean basis vector. For $A, B \in \mathbb{R}^{n \times n}$, $A : B := \sum_{i,j} A_{ij} B_{ij}$ and $A_S = \frac{1}{2}(A + A^\top)$. Finally, $\langle \cdot, \cdot \rangle_{\tilde{\pi}}$ denotes the inner product in $L^2(\tilde{\pi})$ and similar for π .

2.1. Semigroup bound, Poisson equation and central limit theorem

In this section, a central limit theorem for the solution to (1.1) is established, where the resulting asymptotic variance will be used as a cost function for the optimisation of Γ . Specifically, it will be shown that under some weighted L^∞ bound on the observable $f \in L^2(\pi)$, the estimator $\pi_T(f)$ (defined in (1.3)) converges to $\pi(f)$ as $T \rightarrow \infty$ such that (1.4) holds with (1.5).

It is well known that the asymptotic variance can be expressed in terms of the solution to the Poisson equation (1.6) using for example the Kipnis–Varadhan framework, see Chapter 2 in [46], Section 3.1.3 in [51], [12] and references therein. We will first present the exponential decay of the semigroup (see (A.4)). This will allow us to show that the expression (1.10) makes sense as an element in $L^2(\tilde{\pi})$ with zero mean and also that it solves the Poisson equation (1.6). Furthermore, in Theorem 2.2 below, we establish convergence in law to the invariant measure for the Langevin dynamics (1.1).

We will pose the following assumptions on U :

Assumption 2.1. *The function $U \in C^\infty(\mathbb{R}^n)$ satisfies $U \geq 0$. Moreover, there exist constants $\beta_1, \beta_2 > 0$ and $\alpha \in \mathbb{R}$ such that*

$$\forall q \in \mathbb{R}^n, \langle q, \nabla_q U(q) \rangle \geq \beta_1 U(q) + \beta_2 |q|^2 + \alpha. \quad (2.3)$$

The following Lyapunov function $\mathcal{K}_l : \mathbb{R}^{2n} \rightarrow \mathbb{R}$ for all $l \in \mathbb{N}$ will be used:

$$\mathcal{K}_l(z) = \mathcal{K}_l(q, p) = \left(cU(q) + a|q|^2 + b\langle q, p \rangle + \frac{c}{2}|p|^2 + 1 \right)^l \quad (2.4)$$

for constants $a, b, c > 0$. The well-posedness of equation (1.1) is stated in the appendix in Theorem A.1.

Theorem 2.2. *Under Assumption 2.1, $\tilde{\pi}$ is the unique invariant probability measure for the SDE (1.1) and for all $l \in \mathbb{N}$, there exist constants $\kappa_l, C_l > 0$ depending on l and constants $a, b, c > 0$ independent of l such that the solution $z_t^z = (q_t, p_t)$ to (1.1) with initial condition z satisfies*

$$|\mathbb{E}[\varphi(z_t^z)] - \tilde{\pi}(\varphi)| \leq C_l e^{-t\kappa_l} \mathcal{K}_l(z) \left\| \frac{\varphi - \tilde{\pi}(\varphi)}{\mathcal{K}_l} \right\|_{L^\infty} \quad (2.5)$$

for Lebesgue almost all initial $z \in \mathbb{R}^{2n}$, $\mathcal{K}_l \geq 1$ given by (2.4) and all Lebesgue measurable φ satisfying

$$\frac{\varphi}{\mathcal{K}_l} \in L^\infty \quad (2.6)$$

Moreover for any $l \in \mathbb{N}$, \mathcal{K}_l satisfies

$$\int \mathcal{K}_l d\tilde{\pi} < \infty \quad (2.7)$$

and

$$\mathcal{L}\mathcal{K}_l \leq -a_l \mathcal{K}_l + b_l \quad (2.8)$$

for some constants $a_l, b_l > 0$.

For the sake of brevity we omit the proof. The fact that $\tilde{\pi}$ is invariant is thanks to (2.2). For the rest of the statements, the proof is contained in Theorem 3 from [63]. In the latter the setting is more general than (1.1) in that the friction matrix is dependent on q and the drift is not necessarily conservative, *i.e.* the forcing term is not the gradient of a scalar function and the fluctuation-dissipation theorem (see Eq. (6.2) in [60]) does not hold, but of course, Theorem 3 of [63] applies in particular to our setting.

Remark 2.3. Inequality (2.5) holds for all initial $z \in \mathbb{R}^{2n}$, as opposed to almost all z , given any bounded measurable φ . This is a consequence of combining (2.5) together with the strong Feller property given by Theorem 4.2 in [22].

The following corollary holds by taking φ as indicator functions and Remark 2.3.

Corollary 2.4. Under Assumption 2.1, for all initial $z \in \mathbb{R}^{2n}$, the transition probability ρ_t^z of (1.1), given by $\rho_t^z(A) = \mathbb{P}(z_t^z \in A)$, satisfies

$$\|\rho_t^z - \tilde{\pi}\|_{TV} \rightarrow 0 \quad \text{as } t \rightarrow \infty$$

where $\|\cdot\|_{TV}$ denotes the total variation norm.

The solution to the Poisson equation is given next following [12].

Theorem 2.5. Under Assumption 2.1, if $f \in L_0^2(\tilde{\pi})$ satisfies $\frac{f}{\mathcal{K}_l} \in L^\infty$ for some $l \in \mathbb{N}$, then there exists a unique solution $\phi \in L_0^2(\tilde{\pi})$ to the Poisson equation (1.6). Moreover, the solution is given for almost all $z \in \mathbb{R}^{2n}$ by

$$\phi(z) = \int_0^\infty (P_t(f))(z) dt. \quad (2.9)$$

Theorem 2.5 can be proved using the strategy of Corollary 3.2 from [12]. For the reader's convenience, a complete proof of Theorem 2.5 is given in Appendix A.

We proceed to state the central limit theorem for the solution to (1.1).

Theorem 2.6. Under Assumption 2.1, if $f \in L^2(\tilde{\pi})$ satisfies $\frac{f}{\mathcal{K}_l} \in L^\infty$ for some $l \in \mathbb{N}$, then the random variable $\frac{1}{\sqrt{t}} \int_0^t (f(z_s) - \pi(f)) ds$ converges in distribution to $\mathcal{N}(0, \sigma_f^2)$ as $t \rightarrow \infty$ for any initial distribution of z_t , where

$$\sigma_f^2 = 2 \int \phi(z)(f(z) - \pi(f)) \tilde{\pi}(dz) \quad (2.10)$$

and $\phi \in L_0^2(\tilde{\pi})$ is the solution to (1.6).

Proof. By Corollary 2.4 and Theorem 2.5, the result follows by Theorem 2.6 in [9]. Note that the joint measurability assumption in [9] of the transition probability is verified in Theorem A.1. See also Theorem 3.1 in [12]. \square

3. DIRECTIONAL DERIVATIVE OF σ^2

In this section, we give a number of natural preliminary results that pave the path for the main result in Theorem 3.2, in which a formula for the derivative (1.7) of σ^2 with respect to Γ is provided.

3.1. Main derivative formula

In order for the integral in a formula like (1.7) to be finite, control on the derivatives in p is required. This will also be used in the proof of Theorem 3.2 and it is given by the following lemma.

Lemma 3.1. *Under Assumption 2.1, if $f \in L_0^2(\tilde{\pi})$ satisfies $\frac{f}{\kappa_l} \in L^\infty$ for some $l \in \mathbb{N}$, then the weak derivative in p of the solution ϕ to $-L\phi = f$ satisfies $\int |\nabla_p \phi|^2 d\tilde{\pi} < \infty$.*

The proof of Lemma 3.1 is stated in Appendix A for completeness.

The main result of this section is the expression for the directional derivative of the asymptotic variance and is given next. Since Lemma 3.1 is available only for observable functions of position q , the formula for the derivative is given for such observables. The directional derivative of $E : \mathbb{S}_{++}^n \rightarrow \mathbb{R}$ at $\Gamma \in \mathbb{S}_{++}^n$ in a symmetric matrix direction $\delta\Gamma \in \mathbb{R}^{n \times n}$ is denoted by $dE(\Gamma) \cdot \delta\Gamma = \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} (E(\Gamma + \epsilon \delta\Gamma) - E(\Gamma))$ whenever the limit exists. The dependence on Γ is omitted in the notation when no confusion is possible. The proof of Theorem 3.2 is deferred to Section 6.

Theorem 3.2. *Under Assumption 2.1, if $f = f(q) \in L_0^2(\pi)$ is continuous, satisfies $\frac{f}{\kappa_l} \in L^\infty$ for some $l \in \mathbb{N}$ and there exists $\epsilon' > 0$ such that $\Gamma, \Gamma + \epsilon \delta\Gamma \in \mathbb{S}_{++}^n$ for $|\epsilon| \leq \epsilon'$, then the directional derivative of the asymptotic variance σ^2 at Γ in the direction $\delta\Gamma$ has the form*

$$d\sigma^2(\Gamma) \cdot \delta\Gamma = -2 \int (\nabla_p \phi)^\top \delta\Gamma \nabla_p \tilde{\phi} d\tilde{\pi}, \quad (3.1)$$

where ϕ is the solution (2.9) to the Poisson equation (1.6) with L associated with Γ and $\tilde{\phi}$ is given by (1.8).

As mentioned in the introduction, from (3.1), the direction (1.9) guarantees a decrease in asymptotic variance at the infinitesimal level; similarly the scalar change in Γ given by (1.9) where the outer product is replaced by a dot product guarantees a decrease in σ^2 . This is easily verified by substituting such an expression for $\delta\Gamma = \Delta\Gamma$ into the right-hand side of (3.1).

3.2. A formula using a tangent process

The directional derivative $d\sigma^2 \cdot \delta\Gamma$ of the asymptotic variance can be written in a more useful form than in (3.1) for analysis and for simulation based approximation. In this section, the derivatives $(D_p q_t, D_p p_t)$ with respect to the initial conditions, also known as the first variation process of (1.1), will be shown to approximate $\nabla_p \phi$ according to (1.11), which is in turn used to approximate $d\sigma^2 \cdot \delta\Gamma$. The main methodology in the numerical sections will then be to approximate $(D_p q_t, D_p p_t)$ by simulating the SDE it satisfies.

This alternative formula given in Theorem 3.3 provides a way to avoid using a direct finite difference Monte Carlo estimate of the derivative of an expectation. It also allows us to give assumptions on U such that the approximated quantities have desirable properties. We elaborate on the advantage over the finite difference approach for a moment. The expression $\nabla_p \phi$ may be written, from the expression (1.10), as

$$\nabla_p \phi = \nabla_p \int_0^\infty \mathbb{E}[f(q_t) - \pi(f)] dt, \quad (3.2)$$

where the derivative is with respect to the initial velocity p_0 of the trajectory (q_t, p_t) . The right-hand side of (3.2) can be approximated by simulating realisations of the Langevin process at any given initial value, then perturbing the initial velocities and repeating each simulation. In this approach, one can choose how the Brownian realisations in the trajectories with perturbed initial velocities are correlated to those without.

However, if for example the realisations are made independent, then the Monte Carlo variance is magnified for small perturbations in initial velocity, unless increased computational cost is spent corresponding to the smallness of the perturbation in initial velocity. The other natural choice is to synchronise the Brownian realisations (that is, to use a common random number approach for variance reduction). In numerical experiments, this approach seems to work well and is similar compared to our approach using (1.11). However, a simple calculation using the fundamental theorem of calculus shows that, for small perturbations in initial velocity, the resulting procedure is closely related to an approximation of (1.11) by simulating the SDE satisfied by $(D_p q_t, D_p p_t)$, because $D_p q_t, D_p p_t$ are partial derivatives with respect to fixed $\omega \in \Omega$.

In both of the described approaches that respectively use (3.2) and (1.11), we would like to know if a finite time integral approximation of either expression is a good approximation. The results in this section provides justification in this direction by proving (1.10) and providing exponential decay estimates on $(D_p q_t, D_p p_t)$ under suitable conditions on U . In the numerical experiments of Section 5, we choose then to also approximate $(D_p q_t, D_p p_t)$ by simulating directly the SDE it satisfies, rather than use the finite difference procedure based on (3.2), since that aligns more directly with the analysis in this section. An alternative approach to the analysis here is to work with bounds on the derivatives of the associated semigroups, see *e.g.* [4, 18]. In particular, such bounds would also justify a finite time integral approximation of $\nabla_p \phi$, but they provide less direct information on trajectories of the numerical approximations. Finally, note that the approach of passing the derivative under the expectation to analyse derivatives of semigroups is far from new, see for example [17, 47].

We proceed with the main result of this section. For simplicity, we set $M = I_n$ here. The first variation process with respect to the initial momenta p associated to (1.1), denoted by $(D_p q_t, D_p p_t) \in \mathbb{R}^{n \times 2n}$ for $t \geq 0$, is defined as the matrix-valued solution to

$$\partial_t \begin{pmatrix} D_p q_t \\ D_p p_t \end{pmatrix} = \begin{pmatrix} 0 & I_n \\ -D^2 U(q_t) & -\Gamma \end{pmatrix} \begin{pmatrix} D_p q_t \\ D_p p_t \end{pmatrix} \quad (3.3)$$

with the initial condition $D_p q_0 = 0, D_p p_0 = I_n$. By Theorem V.39 in [62], the partial derivatives of (q_t, p_t) with respect to the initial values in p is the unique solution to (3.3) and $(D_p q_t, D_p p_t)$ is continuous with respect to those initial values. We omit the notational dependence of (q_t, p_t) on its initial condition $(q_0, p_0) = (q, p) = z$ whenever convenient in the following.

Theorem 3.3. *Let Assumption 2.1 and $M = I_n$ hold. If in addition,*

- *there exist $U_0 > 0$ and $Q \in \mathbb{S}_{++}^n$ such that for all $q \in \mathbb{R}^n, v \in \mathbb{R}^n$,*

$$v^\top D^2 U(q) v \geq U_0 |v|^2, \quad D^2 U(q) = Q + F(q),$$

where $F : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ is small enough everywhere in the following sense:

$$|F(q)| \leq \hat{\lambda} := \min \left(\frac{\lambda_m}{2}, \frac{\lambda_m U_0^2}{8 \lambda_M^2}, \frac{\lambda_m U_0}{16}, \frac{U_0}{8} \sqrt{\sigma_{\min}(Q)} \right), \quad (3.4)$$

where $\lambda_m, \lambda_M > 0$ are respectively the smallest and largest eigenvalue of Γ and $\sigma_{\min}(Q)$ denotes the smallest eigenvalue of Q ;

- *$f = f(q) \in L_0^2(\pi) \cap C^1(\mathbb{R}^n)$ and satisfies $\frac{|f| + |\nabla f|}{K_l} \in L^\infty$ for some $l \in \mathbb{N}$,*

then the weak derivative $\nabla_p \phi$ has the form (1.11), where q_t solves (1.1) with initial condition $(q_0, p_0) = (q, p)$ and $(D_p q_t, D_p p_t)$ solves (3.3), the latter satisfying

$$|D_p q_t|^2 + |D_p p_t|^2 \leq C' e^{-Ct} \quad (3.5)$$

almost surely for some constants $C, C' > 0$ independent of (q_0, p_0) and $\omega \in \Omega$.

The additional assumptions on U are made in order to ensure that the process $(D_p q_t, D_p p_t)$ converges to zero exponentially quickly so that the integral in (1.11) is finite. In particular, (3.4) requires U to be close to a quadratic function $q^\top Q q$; see also [10] for a situation where a similar assumption is made for the long time behaviour for the Vlasov–Fokker–Planck equation.

Remark 3.4. It will become evident in the proof that exponential decay of the first variation process is not necessary for the derivation of equation (1.11). On the other hand, Proposition 1 in [19] and Proposition 4 in [55] explores more detailed conditions under which exponential contractivity holds and does not hold for scalar friction $\Gamma = \gamma > 0$; our result places the focus on conditions on U such that contractivity holds for all $\Gamma \in S_{++}^n$. The exponential decay (3.5) may also be viewed within the context of Lyapunov exponents for random dynamical systems.

Proof. Let $b > 0$ be the constant

$$b = \min \left(\frac{\lambda_m U_0}{2\lambda_M^2}, \frac{\lambda_m}{4}, \frac{1}{2} \sqrt{\sigma_{\min}(Q)} \right) \quad (3.6)$$

so that $\hat{\lambda}$ reduces to $\hat{\lambda} = \min \left(\frac{\lambda_m}{2}, b \frac{U_0}{4} \right)$ and, since $b \leq \frac{1}{2} \sqrt{\sigma_{\min}(Q)}$, the matrix $\begin{pmatrix} Q & bI_n \\ bI_n & I_n \end{pmatrix}$ is positive definite. We have the following bound.

$$\begin{aligned} \frac{1}{2} \partial_t \left[e_i^\top \begin{pmatrix} D_p q_t \\ D_p p_t \end{pmatrix}^\top \begin{pmatrix} Q & bI_n \\ bI_n & I_n \end{pmatrix} \begin{pmatrix} D_p q_t \\ D_p p_t \end{pmatrix} e_i \right] &= e_i^\top D_p q_t^\top Q D_p p_t e_i + b |D_p p_t e_i|^2 \\ &\quad - e_i^\top (b D_p q_t + D_p p_t)^\top (D^2 U(q) D_p q_t + \Gamma D_p p_t) e_i \\ &= -b e_i^\top D_p q_t^\top D^2 U(q_t) D_p q_t e_i + e_i^\top D_p q_t^\top (-b\Gamma - F(q_t)) D_p p_t e_i \\ &\quad - e_i^\top D_p p_t^\top (\Gamma - bI_n) D_p p_t e_i \\ &\leq \left(-bU_0 + \frac{bU_0}{2} + \frac{\hat{\lambda}}{2} \right) |D_p q_t e_i|^2 + \left(-\lambda_m + b + \frac{b\lambda_M^2}{2U_0} + \frac{\hat{\lambda}}{2} \right) |D_p p_t e_i|^2 \\ &\leq -\frac{bU_0}{4} |D_p q_t e_i|^2 - \frac{\lambda_m}{4} |D_p p_t e_i|^2 \\ &\leq -C e_i^\top \begin{pmatrix} D_p q_t \\ D_p p_t \end{pmatrix}^\top \begin{pmatrix} Q & bI_n \\ bI_n & I_n \end{pmatrix} \begin{pmatrix} D_p q_t \\ D_p p_t \end{pmatrix} e_i \end{aligned} \quad (3.7)$$

for some generic constant $C > 0$ independent of the initial values (q_0, p_0) and $\omega \in \Omega$. Consequently, it holds for any $1 \leq i \leq n$ that

$$\left[e_i^\top \begin{pmatrix} D_p q_t \\ D_p p_t \end{pmatrix}^\top \begin{pmatrix} Q & bI_n \\ bI_n & I_n \end{pmatrix} \begin{pmatrix} D_p q_t \\ D_p p_t \end{pmatrix} e_i \right] \leq e^{-2Ct},$$

which implies (3.5) and, using the (weighted) boundedness assumption on $|\nabla f|$,

$$\begin{aligned} |(\nabla f(q_t)^\top D_p q_t)_i| &\leq C' e^{-Ct} |\nabla f(q_t)| \\ &\leq C' e^{-Ct} (|\nabla f(q_t)| - \pi(|\nabla f|)) + C' e^{-Ct} \end{aligned} \quad (3.8)$$

for a generic $C' > 0$ independent of (q_0, p_0) and $\omega \in \Omega$. Due to (3.8) together with Fubini's theorem, it holds for $T > 0$ and a test function $g \in C_c^\infty(\mathbb{R}^{2n})$ that

$$\int \int_0^T \mathbb{E}[f(q_t^z)] dt \nabla_p g(z) dz = \int_0^T \mathbb{E} \left[\int f(q_t^z) \nabla_p g(z) dz \right] dt$$

$$\begin{aligned}
&= - \int_0^T \mathbb{E} \left[\int \nabla f(q_t^z)^\top D_p q_t^z g(z) dz \right] dt \\
&= - \int_0^T \int_0^T \mathbb{E} \left[\nabla f(q_t^z)^\top D_p q_t^z \right] dt g(z) dz.
\end{aligned}$$

Using Theorem 2.2, (3.8) again and dominated convergence to take $T \rightarrow \infty$ on both sides concludes the proof. \square

The following is a brief discussion about how equation (1.11) can be used in practice to approximate the gradient direction $\int \nabla_p \phi \otimes \nabla_p \tilde{\phi} d\tilde{\pi}$ from realisations (q_t, p_t) of (1.1). The underlying idea is to approximate $\nabla_p \phi(q, p) \otimes \nabla_p \tilde{\phi}(q, p)$ for points (q, p) along the trajectory of (q_t, p_t) , so that (q, p) is approximately distributed according to $\tilde{\pi}$ due to the ergodicity of (q_t, p_t) . We have in mind at first setting for example $(q, p) = (q_0, p_0)$, where (q_0, p_0) is the initial condition from equation (1.1), so that equation (1.11) implies $\nabla_p \phi(q, p) \otimes \nabla_p \tilde{\phi}(q, p)$ can be approximated using

$$\delta\Gamma = \int_0^T \nabla f(q_s^{(q,p)})^\top D_p q_s^{(q,p)} ds \otimes \int_0^T \nabla f(q_s^{(q,-p)})^\top D_p q_s^{(q,-p)} ds, \quad (3.9)$$

where $(q_s^{(q,p)}, p_s^{(q,p)})$ solves (1.1) with initial condition (q, p) , and $(q_s^{(q,-p)}, p_s^{(q,-p)})$ denotes a parallel solution of (1.1) with initial condition $(q, -p)$ and independent realisations for W , moreover, $(D_p q_s)_{0 \leq s \leq T}$ denotes corresponding solutions to (3.3) for both initial conditions. At time T , we then update Γ with $\Gamma + \hat{\delta}\delta\Gamma$ for a stepsize $\hat{\delta} > 0$ using equation (3.9). After the update in Γ , in order to proceed with further updates, we set $(q, p) = (q_T, p_T)$ and repeat the described procedure above using (3.9) with the updated $(q, p) = (q_T, p_T)$, then with $(q, p) = (q_{2T}, p_{2T})$ and so on. Note that we expect (q_{jT}, p_{jT}) to be distributed approximately as $\tilde{\pi}$ for nonzero $j \in \mathbb{N}$, so that the approximations (3.9) of the integrand $\nabla_p \phi(q, p) \otimes \nabla_p \tilde{\phi}(q, p)$ with $(q, p) = (q_{jT}, p_{jT})$ indeed makes for an approximation of the integral $\int \nabla_p \phi \otimes \nabla_p \tilde{\phi} d\tilde{\pi}$. There can be sources of bias from not being at stationarity and using finite T , in the sense that the points (q, p) along the trajectory of (q_t, p_t) are clearly not exactly distributed according to $\tilde{\pi}$ and the finite time integrals in (3.9) are not equal to the infinite time integral as required. However, both of these factors may be mitigated in practice. For example, we may choose the first point (q, p) above to be a point in the trajectory of (q_{t^*}, p_{t^*}) for some burn-in time t^* instead of $(q, p) = (q_0, p_0)$. We may also choose T large enough so that $(D_p q_T, D_p p_T)$ has converged to zero with satisfaction (which will happen for large enough T thanks to Thm. 3.3), so that the finite time integral is a good approximation of the infinite time one. The overall approach is summarised in Algorithm 1 and given with more detail in Algorithm 2.

The next result is that the estimator (3.9) has finite variance.

Theorem 3.5. *Let the assumptions of Theorem 3.3 and $M = I_n$ hold. For Lebesgue almost-all $(q, p) \in \mathbb{R}^{2n}$, each entry of $\delta\Gamma$ defined in (3.9) has finite variance.*

Proof. It suffices to show that (3.9) has finite second moment, for which it suffices to show that each element in the vector of time integrals $\int_0^T \nabla f(q_t^{(q,p)})^\top D_p q_t^{(q,p)} ds$ has finite second moments by independence. For each index i , using (3.5),

$$\begin{aligned}
|(\nabla f(q_t)^\top D_p q_t)_i|^2 &\leq C'^2 e^{-2Ct} |\nabla f(q_t)|^2 \\
&\leq C'^2 e^{-2Ct} \left(|\nabla f(q_t)|^2 - \pi(|\nabla f|^2) \right) + C'^2 e^{-2Ct} \pi(|\nabla f|^2),
\end{aligned}$$

so that using the (weighted) boundedness assumption on $|\nabla f|$ together with Theorem 2.2 and Fubini's theorem, the proof concludes. \square

4. GAUSSIAN CASES

Throughout this section, the target measure π is assumed to be Gaussian, when π is mean zero this is $\pi \propto \exp(-\frac{1}{2}q^\top \Sigma^{-1}q)$ for $\Sigma \in \mathbb{S}_{++}^n$, in other words, the potential is quadratic, $U(q) = \frac{1}{2}q^\top \Sigma^{-1}q$. For polynomial function observables, we look for solutions to the Poisson equation (1.6) by using a polynomial ansatz and comparing coefficients in order to obtain an explicit expression for the asymptotic variance. The results provide benchmarks to test the performance of the algorithms that arise from using the gradient in Theorem 3.2 as well as intuition for how Γ can be improved in concrete cases. We will consider the following cases for the observables:

- (1) Quadratic $f(q) = \frac{1}{2}q^\top U_0 q$ under the assumption of commutativity between U_0 and Σ (Prop. 4.6), also $f(q) = \frac{1}{2}U_0 q^2 + lq$ in one dimension (Prop. 4.7);
- (2) Odd polynomial f , where the asymptotic variance will be shown to decrease to zero as $\Gamma \rightarrow 0$ (Props. 4.8, 4.10 and Cor. 4.9);
- (3) Quartic f in one dimension, in which case the situation is similar to quadratic f (Prop. 4.11).

All of the proofs and derivations for the results in this section can be found in Section 6, except the short proofs of Proposition 4.4, 4.6 and 4.7 that are stated in this section for clarity of presentation. We proceed with stating in more detail the general situation of this section.

Let $\Sigma \in \mathbb{S}_{++}^n$, $U_0 \in \mathbb{S}_{++}^n$ and $l \in \mathbb{R}^n$. The Gaussian invariant measure $\tilde{\pi}$ and the observable $f : \mathbb{R}^{2n} \rightarrow \mathbb{R}$ are given by

$$\tilde{\pi} \propto \exp\left(-\frac{1}{2}q \cdot \Sigma^{-1}q - \frac{1}{2}p \cdot M^{-1}p\right), \quad f(q) = \frac{1}{2}q \cdot U_0 q + l \cdot q \quad (4.1)$$

and the value $\pi(f)$ becomes

$$\pi(f) = \int f d\pi = \int \frac{1}{2}q \cdot U_0 q d\pi = \frac{1}{2}U_0 : \Sigma. \quad (4.2)$$

The infinitesimal generator \mathcal{L} becomes in this case

$$\begin{aligned} \mathcal{L} &= \begin{pmatrix} 0 & M^{-1} \\ -\Sigma^{-1} & -\Gamma M^{-1} \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix} \cdot \nabla + \nabla_p \cdot \Gamma \nabla_p \\ &= M^{-1}p \cdot \nabla_q - \Sigma^{-1}q \cdot \nabla_p - \Gamma M^{-1}p \cdot \nabla_p + \nabla_p \cdot \Gamma \nabla_p. \end{aligned} \quad (4.3)$$

Consider the natural candidate solution ϕ to the Poisson equation (1.6) given by

$$\phi(q, p) = \frac{1}{2}q \cdot Gq + q \cdot Ep + \frac{1}{2}p \cdot Hp + g \cdot q + h \cdot p - \frac{1}{2}(G : \Sigma + H : M) \quad (4.4)$$

for some constant matrices $G, E, H \in \mathbb{R}^{n \times n}$ and vectors $g, h \in \mathbb{R}^n$. Note that we allow G and H not to be symmetric and specify G_S and H_S as the respective symmetric parts in order to make a clear distinction.

Lemma 4.1. *Given f and $\tilde{\pi}$ in (4.1) and \mathcal{L} of the form (4.3), it holds that ϕ given by (4.4) is a solution to the Poisson equation (1.6) if and only if*

$$\Sigma^{-1}q \cdot (E^\top q + h) - \Gamma : H_S - \frac{1}{2}q \cdot U_0 q - l \cdot q + \frac{1}{2}U_0 : \Sigma = 0, \quad (4.5)$$

$$-M^{-1}(G_S q + g) + H_S \Sigma^{-1}q + M^{-1}\Gamma(E^\top q + h) = 0, \quad (4.6)$$

$$-E^\top M^{-1} + H_S \Gamma M^{-1} = A_1, \quad (4.7)$$

for some antisymmetric $A_1 \in \mathbb{R}^{n \times n}$.

4.1. Quadratic observable

Similar calculations in this situation have appeared previously in Proposition 1 in [23], where explicit expressions analogous to G , E , H and for σ^2 are given. For our purposes of finding an optimal Γ , the approach take here is different. Instead of taking these explicit expressions, we keep unknown antisymmetric matrices (such as A_1) as they appear as an alternative to the aforementioned explicit expressions. Eventually the commutativity property between Σ and U_0 is used to show that the antisymmetric matrices are zero. We continue from (4.5) to (4.7) with finding explicit expressions for the coefficients G , E , H of ϕ .

Lemma 4.2. *Given f and $\tilde{\pi}$ in (4.1) and \mathcal{L} of the form (4.3), ϕ given by (4.4) is a solution to the Poisson equation (1.6) with (4.3) if and only if there exist antisymmetric matrices A_1, A_2 such that*

$$G_S = \frac{1}{2}M(\Sigma U_0 - \Sigma A_2 - 2A_1 M)\Gamma^{-1}\Sigma^{-1} + \frac{1}{2}\Gamma(U_0\Sigma - A_2\Sigma), \quad (4.8)$$

$$E = \frac{1}{2}U_0\Sigma + \frac{1}{2}A_2\Sigma, \quad (4.9)$$

$$H_S = \frac{1}{2}(\Sigma U_0 - \Sigma A_2 - 2A_1 M)\Gamma^{-1}, \quad (4.10)$$

$$h = \Sigma l \quad \text{and} \quad g = \Gamma \Sigma l. \quad (4.11)$$

The asymptotic variance from Theorem 2.6 can be given by a formula in terms of Σ, U_0 and the coefficients of ϕ . Before substituting the expressions from Lemma 4.2 into the formula, we give the formula itself, which is adapted from the proof of Proposition 1 in [23].

Lemma 4.3. *Let f and $\tilde{\pi}$ be given by (4.1) and \mathcal{L} be given by (4.3). If the solution ϕ to the Poisson equation (1.6) is of the form (4.4), then the asymptotic variance σ^2 given by (2.10) has the expression*

$$2\langle \phi, f - \pi(f) \rangle_{\tilde{\pi}} = \text{Tr}(G_S \Sigma U_0 \Sigma) + 2g \cdot \Sigma l. \quad (4.12)$$

From the expressions (4.8) and (4.10) for G_S and H_S respectively, it is not straightforward to check that there exist antisymmetric A_1 and A_2 such that the right hand sides are indeed symmetric at this point, which is necessary for the ansatz (4.4) for ϕ to be a valid solution. On the other hand, if Σ, U_0, Γ, M all commute, then the right hand sides of (4.8) and (4.10) are symmetric for $A_1 = A_2 = 0$ and the coefficients G and H become explicit, which allows taking derivatives of σ^2 with respect to the entries of Γ . Moreover, the explicit coefficients allow optimisation of M , which is given by the following proposition.

Proposition 4.4. *Suppose Σ, U_0 and Γ all commute. Let f be as in (4.1), $\pi(f)$ be as in (4.2), \mathcal{L} be of the form (4.3) and ϕ be the solution to the Poisson equation (1.6). It holds that*

$$\lim_{M=mI_n, m \downarrow 0} \int \phi(f - \pi(f)) d\tilde{\pi} = \inf_{M \in \mathbb{S}_{\Sigma++}} \int \phi(f - \pi(f)) d\tilde{\pi}, \quad (4.13)$$

where $\mathbb{S}_{\Sigma++}$ is the set of symmetric positive definite matrices commuting with Σ .

Proof. Let

$$G = \frac{1}{2}M\Sigma U_0 \Gamma^{-1} \Sigma^{-1} + \frac{1}{2}\Gamma U_0 \Sigma, \quad E = \frac{1}{2}U_0 \Sigma, \quad H = \frac{1}{2}\Sigma U_0 \Gamma^{-1}, \quad (4.14a)$$

$$g = \Gamma \Sigma l, \quad h = \Sigma l \quad (4.14b)$$

so that by Lemma 4.2, ϕ given by (4.4) is the solution to the Poisson equation (1.6) and inserting G, g into (4.12) gives

$$2\langle \phi, f - \pi(f) \rangle_{\tilde{\pi}} = \frac{1}{2} \text{Tr}(M\Sigma U_0 \Gamma^{-1} U_0 \Sigma + \Gamma U_0 \Sigma^2 U_0 \Sigma) + 2l^\top \Sigma \Gamma \Sigma l. \quad (4.15)$$

The result follows since $A : B > 0$ for $A, B \in \mathbb{S}_{++}^n$. \square

Remark 4.5. The limit (4.13) in Proposition 4.4 is, together with a rescaling in the velocity space, the overdamped limit of the Langevin dynamics, see Section 2.2.4 in [52]. However, equation (4.13) does not necessarily mean overdamped dynamics are better in practice. For example when Γ is a small scalar, the overdamped limit corresponding to (4.13) results in a time speed-up inversely proportional to Γ over the overdamped limit corresponding to $\Gamma = I_n$. Consequently, any such comparison between Langevin dynamics and the overdamped limit should include constraints such as those in [33] for both sets of dynamics. We focus on the optimisation of Γ and fix $M = I_n$ in the following.

As before, we denote $\mathbb{S}_{\Sigma++}$ to be the set of symmetric positive definite matrices commuting with Σ .

Proposition 4.6. *Let Σ, U_0, l, M be such that*

$$\Sigma U_0 = U_0 \Sigma, \quad l = 0, \quad M = I_n, \quad (4.16)$$

f be as in (4.1), $\pi(f)$ be as in (4.2), \mathcal{L} be of the form (4.3) and ϕ be the solution to the Poisson equation (1.6). It holds that

$$\min_{\Gamma \in \mathbb{S}_{\Sigma++}} 2 \int \phi(f - \pi(f)) d\tilde{\pi} = \text{Tr}(U_0^2 \Sigma^{\frac{5}{2}}),$$

where the minimum is attained by $\Gamma = \Sigma^{-\frac{1}{2}}$.

Proof. Let $\Sigma = P^\top \Sigma_d P$ be the eigendecomposition of Σ for orthogonal P . Since all symmetric matrices in the set commuting with Σ share eigenvectors with Σ , it suffices to find a unique extremal point of the asymptotic variance with respect to the eigenvalues of Γ , call them $(\lambda_i)_{1 \leq i \leq n}$, $\lambda_i \geq 0$. Setting again (4.14), ϕ given by (4.4) is the solution to the Poisson equation (1.6) and the asymptotic variance σ^2 given by (2.10) becomes

$$2\langle \phi, f - \pi(f) \rangle_{\tilde{\pi}} = \frac{1}{2} \text{Tr}(\Sigma U_0 \Gamma^{-1} U_0 \Sigma + \Gamma U_0 \Sigma^2 U_0 \Sigma), \quad (4.17)$$

which reduces to a sum of functions of the form $a_i \lambda_i^{-1} + b_i \lambda_i$, $a_i, b_i > 0$ after diagonalising with P and the result follows. \square

In the scalar case, we can remove the restriction on l .

Proposition 4.7. *If $n = 1$, $U_0 \neq 0$, $l \neq 0$, $f : \mathbb{R} \rightarrow \mathbb{R}$ is given by (4.1), $\pi(f)$ is given by (4.2), \mathcal{L} is of the form (4.3) and ϕ is the solution to the Poisson equation (1.6), then $\min_{\Gamma > 0} 2 \int \phi(f - \pi(f)) d\tilde{\pi} = M^{\frac{1}{2}} \Sigma^2 U_0^2 (\Sigma + 4l^2 U_0^{-2})^{\frac{1}{2}}$ and the minimum is attained by $\Gamma = \frac{M^{\frac{1}{2}}}{(\Sigma + 4l^2 U_0^{-2})^{\frac{1}{2}}}$.*

Proof. By Lemma 4.2, the solution (4.4) to the Poisson equation (1.6) is

$$\phi = \left(\frac{U_0 \Gamma \Sigma}{4} + \frac{M U_0}{4 \Gamma} \right) q^2 + \frac{U_0 \Sigma}{2} q p + \frac{U_0 \Sigma}{4 \Gamma} p^2 + \Sigma \Gamma l q + \Sigma l p - \frac{U_0 \Gamma \Sigma^2}{4} - \frac{M U_0 \Sigma}{2 \Gamma}.$$

By Lemma 4.3, the asymptotic variance is given by

$$2 \int \phi(f - \pi(f)) d\tilde{\pi} = 2 \Sigma^2 \left(\frac{U_0^2 \Sigma}{4} + l^2 \right) \Gamma + \frac{U_0^2 \Sigma^2}{2 \Gamma},$$

which attains the stated minimum at the stated Γ . \square

4.2. Odd polynomial observable

Another special case within (4.1) where the solution ϕ can be readily identified is when $U_0 = 0$, that is, for linear observables. More generally, (almost) zero variance can be attained in the following special case.

Proposition 4.8. *Under Assumption 2.1, for a general target measure $\pi \propto e^{-U}$ on \mathbb{R}^n , if the observable f is of the form $f(q) = \alpha \cdot \nabla U$, for $\alpha = (\alpha_1, \dots, \alpha_n)$, $\alpha_i \in \mathbb{R}$, \mathcal{L} is of the general form (2.1) and ϕ is the solution to the Poisson equation (1.6), then the asymptotic variance satisfies*

$$\inf_{\Gamma \in \{\gamma I_n : \gamma > 0\}} 2 \int \phi(f - \pi(f)) d\tilde{\pi} = 0. \quad (4.18)$$

Corollary 4.9. *Given a Gaussian target measure with density $\pi \propto e^{-U}$ on \mathbb{R}^n , observable $f : \mathbb{R}^n \rightarrow \mathbb{R}$ given by $f(q) = l \cdot q$ with $l \in \mathbb{R}^n$, \mathcal{L} of the form (2.1) and ϕ the solution to the Poisson equation (1.6), then identity (4.18) holds.*

Corollary 4.9 follows from Proposition 4.8 as the special case where U is a quadratic function. Note that Corollary 4.9 is also a consequence of (4.15) in the proof of Lemma 4.4. Furthermore, the setting in Corollary 4.9 is included in that of Proposition 4.4 and the results are consistent by Remark 4.5.

We give here some intuition for the situation in Corollary 4.9. First note that the Langevin diffusion with $\Gamma = 0$ reduces to deterministic Hamiltonian dynamics and that it is the limit case for the Γ attaining arbitrarily small asymptotic variance in the proof of Proposition 4.8. The result indicates that this is optimal in the linear observable, Gaussian measure case (*i.e.* (4.1), $U_0 = 0$) and this aligns with the fact that the value (4.2) to be approximated is exactly the value at the $q = p = 0$, so that Hamiltonian dynamics starting at $q = 0$, staying there for all time, approximates the integral (4.2) with perfect accuracy. A similar idea holds for when the initial condition is not $q = p = 0$, where (4.2) is approximated exactly after any integer number of orbits in (q, p) space.

Continuing on this idea, it seems reasonable that the same statement holds more generally for any odd observable. At least, the following holds in one dimension.

Proposition 4.10. *If $n = 1$, $\hat{k} \in \mathbb{N}_0$ and $f : \mathbb{R} \rightarrow \mathbb{R}$ is an odd finite order polynomial observable given by*

$$f(q) = \sum_{i=0}^{\hat{k}} a_i q^{2i+1}, \quad (4.19)$$

$\pi(f) = 0$, \mathcal{L} is of the form (4.3) and ϕ is the solution to the Poisson equation (1.6), then the asymptotic variance satisfies (4.18).

4.3. Quartic observable

The situation in the quartic observable case, at least in one dimension, is similar to quadratic observable case.

Proposition 4.11. *If $n = 1$ and $f : \mathbb{R} \rightarrow \mathbb{R}$ is a quartic observable given by*

$$f(q) = q^4, \quad (4.20)$$

$\pi(f) = 3\Sigma^2$ for some $\Sigma > 0$, \mathcal{L} is of the form (4.3), $M = 1$ and ϕ is the solution to the Poisson equation (1.6), then there exists $\sigma_{\text{quar}} > 0$ such that $\min_{\Gamma=\gamma>0} 2 \int \phi(f - \pi(f)) d\tilde{\pi} = \sigma_{\text{quar}}$.

5. COMPUTATION OF THE CHANGE IN Γ

Throughout this section, the $M = I_n$ case is considered. As mentioned, the formula (3.1) gives a natural gradient descent direction (1.9) to take Γ in order to optimise σ^2 from Theorem 2.6. We focus the discussion on a Monte Carlo method to approximate $\nabla_p \phi$ and gradient directions in Γ (e.g. (1.9)) based on Theorems 3.2 and 3.3 as discussed at the beginning of Section 3.2. Note that it is also possible to use a spectral method to solve (1.6) in order to compute the change in Γ , but this is only computationally feasible in low dimensions. Algorithm 1 summarises the resulting procedure, where all expectations within (1.9) are approximated by single realisations; further justifications, alternative methods, refinements and a concrete implementation (Algorithm 2) can be found in Appendix B. We proceed with applications of the algorithm in examples.

Algorithm 1: Continuous-time outline of Γ update using (1.7) and (1.11).

Result: $\Gamma \in \mathbb{S}_{++}^n$

Start from arbitrary $(q_0, p_0) \in \mathbb{R}^{2n}$ and set $(\tilde{q}_0, \tilde{p}_0) = (q_0, -p_0)$, $Dq_0 = D\tilde{q}_0 = 0$, $Dp_0 = D\tilde{p}_0 = I_n$, $\zeta = \tilde{\zeta} = 0$, $\Gamma = I_n$, $t = t_0 = 0$;

for N epochs **do**

 simulate one time-step in (q_t, p_t) , $(\tilde{q}_t, \tilde{p}_t)$ then in $(D_p q_t, D_p p_t)$ and $(D_p \tilde{q}_t, D_p \tilde{p}_t)$;
 add to $\zeta, \tilde{\zeta}$ to approximate the row vectors

$$\zeta = \int_{t_0}^t \nabla f(q_s)^\top D_p q_s \, ds, \quad \tilde{\zeta} = \int_{t_0}^t \nabla f(\tilde{q}_s)^\top D_p \tilde{q}_s \, ds;$$

if $(D_p q_t, D_p p_t)$ is small enough in magnitude **then**

 update Γ with the gradient direction $-\zeta \otimes \tilde{\zeta}$;

 reset $(\tilde{q}_t, \tilde{p}_t) \leftarrow (q_t, -p_t)$; $(D_p q_t, D_p p_t), (D_p \tilde{q}_t, D_p \tilde{p}_t) \leftarrow (0, I_n)$; $t_0 \leftarrow t$; $\zeta, \tilde{\zeta} \leftarrow 0$;

end

$t \leftarrow t + \Delta t$.

end

Section 5.1 contains the simplest one-dimensional Gaussian case where the optimal Γ is known and it is shown that the algorithm approximates it quickly. A different Gaussian problem extracted from a diffusion bridge context is explored in Section 5.2, where the algorithm is shown to approximate a Γ matrix that exhibits an even better empirical asymptotic variance than the one given by Proposition 4.6. Finally, the algorithm is applied to finding the optimal Γ in estimating the posterior mean in a Bayesian inference problem in Section 5.3, where the situation is shown to be similar to Proposition 4.9, in the sense that the optimal Γ is close to 0; after and separately from such a finding, the empirical asymptotic variance for a small Γ is compared that for $\Gamma = I_n$, with dramatic improvement in both the full gradient and minibatch gradient cases.

5.1. One dimensional quadratic case

Here the algorithm given in Section B.1.3 is used in the simplest one dimensional

$$U(q) = \frac{V_0}{2} q^2, \quad f(q) = \frac{1}{2} q^2, \quad (5.1)$$

$V_0 > 0$, case to find the optimal constant friction. Since commutativity issues disappear in the one-dimensional case, the optimal constant friction is known analytically and is given by Proposition 4.6 to be $\Gamma = \sqrt{V_0}$, with the asymptotic variance $V_0^{-\frac{5}{2}}$. Moreover, the relationship between the asymptotic variance and Γ is explicitly given by equations (4.8) and (4.12), which reduces in this case to

$$\sigma^2(\Gamma) = \frac{1}{4V_0^2} \left(\Gamma^{-1} + \frac{1}{V_0} \Gamma \right).$$

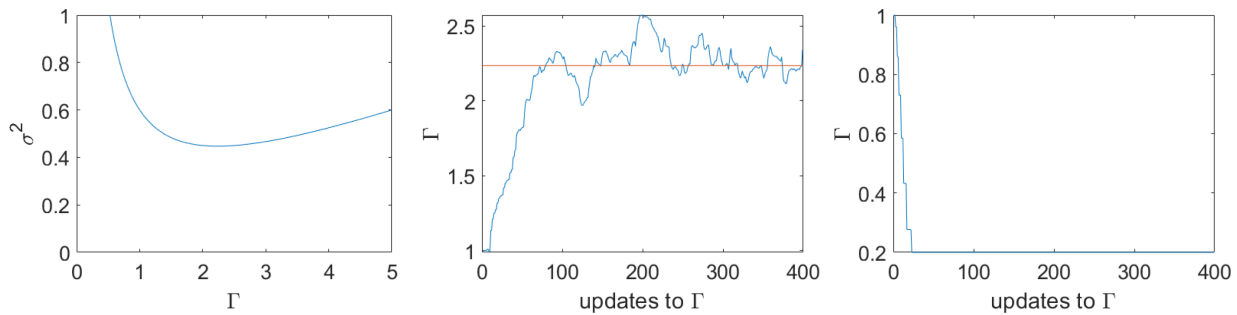


FIGURE 2. *Left*: relationship between asymptotic variance and Γ for (5.1). *Middle and right*: trajectory of Γ for (5.1) and (5.2) respectively by (B.8) with $\alpha^i = 1$, $G = 1$, $r = 0.5$ and $\mu = 0.2$. *Middle*: red line is the optimal value $\Gamma = \sqrt{5}$ given by Proposition 4.6. All plots are for $V_0 = 5$.

The case $V_0 = 5$ is illustrated in Figure 2. In the middle and right plot of Figure 2, the procedure in Section B.1.3 is used for 5×10^4 epochs, with $\Delta t = 0.08$, block-size $T = 125$, $L = 1$ and $D_{\text{conv}} = 2 \times 10^{-4}$. Changing the observable to the linear

$$f(q) = q \quad (5.2)$$

gives that the “optimal” (but unreachable in the algorithm due to the constraints) friction is 0 by Corollary 4.9. The right plot in Figure 2 shows that the procedure arrives at a similar conclusion in the sense that the Γ hits and stays at $\mu = 0.2$. The value $\mu = 0.2$ is precisely where the algorithm imposes a lower bound on the value of Γ . This is expected since it is the closest value to the “optimal” value 0 as mentioned.

5.2. Diffusion bridge sampling

The algorithm in Section B.1.3 is applied in the context of diffusion bridge sampling [35, 36] (see also for example [7, 20, 37]), where the SDE

$$dx_t = -\nabla V(x_t) dt + \sqrt{2\beta^{-1}} dW'_t \quad (5.3)$$

for a suitable $V : \mathbb{R}^d \rightarrow \mathbb{R}$, $\beta > 0$ and W'_t standard Wiener process on \mathbb{R}^d , is conditioned on the events

$$x_0 = x_- \quad \text{and} \quad x_1 = x_+ \quad (5.4)$$

for some fixed $x_0, x_+ \in \mathbb{R}^d$ and the problem setting is to sample from the path space of solutions to (5.3) conditioned on (5.4). For the derivation of the following formulations, we refer to Section 5 in [36] and Section 6.1 in [8]; here we extract a simplified potential U to apply our algorithm on after a brief description.

Let

$$V(x) = \frac{1}{2}|x|^2, \quad x_- = x_+ = 0, \quad \beta = 1, \quad d = 1.$$

Using the measure given by Brownian motion conditioned on (5.4) as the reference measure ν_0 on the path space of continuous functions $C([0, 1], \mathbb{R})$, the measure ν on $C([0, 1], \mathbb{R})$ associated to (5.3) conditioned on (5.4) satisfies $\frac{d\nu}{d\nu_0}(x) \propto \exp(-\frac{1}{4} \int_0^1 |x_t|^2 dt)$, where the left hand side denotes the Radon–Nikodym derivative, so that discretising ν on a grid in $[0, 1]$ with grid-size $\delta > 0$ gives the approximating measure $\pi(q_1, \dots, q_n) \propto e^{-U(q_1, \dots, q_n)}$

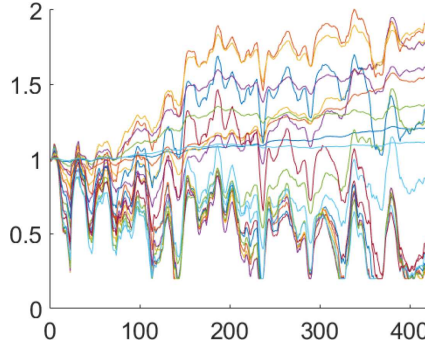


FIGURE 3. Diagonal values of Γ over iterations of (B.8) with $\alpha^i = 0.2$, $G = 5$, $r = 1$ and $\mu = 0.2$.

where U is given by

$$U(q) = \frac{1}{2}q^\top \Sigma^{-1}q = \frac{1}{2}q^\top \begin{pmatrix} \frac{2}{\delta} + \frac{\delta}{4} & -\frac{1}{\delta} & & & \\ -\frac{1}{\delta} & \frac{2}{\delta} + \frac{\delta}{4} & -\frac{1}{\delta} & & \\ & \dots & & & \\ & & -\frac{1}{\delta} & \frac{2}{\delta} + \frac{\delta}{4} & -\frac{1}{\delta} \\ & & & \frac{2}{\delta} + \frac{\delta}{4} & -\frac{1}{\delta} \end{pmatrix} q.$$

From here the Langevin system (1.1) can be used to sample from π and the algorithm given in Section B.1.3 is applied. For this purpose, the observable $f(q) = \frac{1}{2}|q|^2$ is used together with the parameters $\delta = \frac{1}{21}$, $n = 20$, $K = 1$, $L = 5$, $T = 60$, $B = 100$ and $D_{\text{conv}} = 0.01$. Only the diagonal values of Γ are updated and their trajectories are shown in Figure 3.

At the end of approximately 30 000 epochs of simulation (of the Langevin process and of the first variation process), Γ at the end of the update period shown in Figure 3 is given by

$$\Gamma_{\text{final}} = \text{diag}(1.2129, 1.5673, 1.8199, 1.8055, 1.2858, 0.9013, 0.3588, 0.2631, \\ 0.2000, 0.2000, 0.2252, 0.2579, 0.3621, 0.4715, 1.3842, 1.9467, \\ 1.9289, 1.6326, 1.3730, 1.1153).$$

Note that the value $T = 60$ above is the number of time steps in the approximation of (1.1) to be taken between subsequent updates in Γ . The number of iterations in the bottom axis of Figure 3 then indicates the number of updates in Γ , not to be confused with T or the total number of epochs. The value $\Gamma = \Gamma_{\text{final}}$ is then fixed and used for a standard sampling procedure for the same potential and observable. The asymptotic variance is approximated by grouping the epochs after $B = 100$ burn-in iterations into $N_B = 999$ blocks of $T = 300$ epochs, specifically,

$$\sigma_{\text{approx}} = \frac{1}{N_B} \sum_{l=0}^{N_B-1} \left[\frac{1}{\sqrt{T\Delta t}} \sum_{i=1}^T \Delta t \left(f(q^{i+Tl+B}) - \frac{1}{N} \sum_{j=1}^N \Delta t f(q^{j+B}) \right) \right]^2, \quad (5.5)$$

where q^i are iterates in the numerical approximation of q_t , and this is compared to the estimate from the same procedure using different values of fixed Γ in Table 1. Note that $\Gamma = \Sigma^{-\frac{1}{2}}$ is the optimal Γ in the restricted class of matrices commuting with Σ given by Proposition 4.6, where the asymptotic variance is known to be $\text{Tr}(\Sigma^{\frac{3}{2}}) \approx 6.4785$. This is well approximated by the relevant value in Table 1, which validates (5.5) as an approximation of the asymptotic variance. Table 1 also shows that fixing $\Gamma = \Gamma_{\text{final}}$ gives an improved value of

TABLE 1. Empirical asymptotic variances with $N_B = 999$, $T = 300$, $B = 100$, $N = 299\,700$.

	σ_{approx}
$\Gamma = I_n$	6.9834
$\Gamma = \Sigma^{-\frac{1}{2}}$	6.5096
$\Gamma = \Gamma_{\text{final}}$	6.1667

estimated asymptotic variance over $\Gamma = \Sigma^{-\frac{1}{2}}$. There is an extra computational cost to obtain Γ_{final} by using the update procedure in Γ that is illustrated in Figure 3; however, this constitutes a one-time initial cost at the beginning of a simulation, which looks to improve the variance over arbitrarily long simulations.

5.3. Bayesian inference

We adopt the binary regression problem as in [25] on a dataset¹ with datapoints encoding information about images on a webpage and each labelled with “ad” or “non-ad”. The labels $\{Y_i\}_{1 \leq i \leq p}$, taking values in $\{0, 1\}$, of the $p = 2359$ datapoints (counting only those without missing values) given in the dataset are modelled as conditionally independent Bernoulli random variables with probability $\{\rho(\beta^\top X_i)\}_{1 \leq i \leq p}$, where ρ is the logistic function given by $\rho(z) = e^{cz}/(1 + e^{cz})$ for all $z \in \mathbb{R}$, $c \in \mathbb{R}$ is given by (5.8), $\{X_i\}_{1 \leq i \leq p}$, β , both taking values in \mathbb{R}^n , are respectively vectors of known features from each datapoint and regression parameters to be determined. The parameter β has prior distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma^{-1} = \frac{1}{p} \sum_{i=1}^p X_i^\top X_i \in \mathbb{R}^{n \times n}$, and the density of the posterior distribution of β is given up to proportionality by

$$\pi_\beta(\beta | \{(X_i, Y_i)\}_{1 \leq i \leq p}) \propto \exp\left(\sum_{i=1}^p \left\{cY_i\beta^\top X_i - \log\left(1 + e^{c\beta^\top X_i}\right)\right\} - \frac{1}{2}\beta^\top \Sigma^{-1}\beta\right),$$

so that the log-density gradient, in our notation $-\nabla U$, is given by $-\nabla U(\beta) = \sum_{i=1}^p cX_i(Y_i - (1 + e^{-c\beta^\top X_i})^{-1}) - \Sigma^{-1}\beta$. The observable vector $f_i(\beta) = \beta_i$, $1 \leq i \leq n$, corresponding to the posterior mean is used. The coordinate transform $\hat{\beta} = \Sigma^{-\frac{1}{2}}\beta$ is made before applying the symmetric preconditioner $\Sigma^{\frac{1}{2}}$ on the Hamiltonian part of the dynamics so that the dynamics simulated are as in (1.1) with $M = I_n$ and

$$-\nabla U(\hat{\beta}) = \Sigma^{\frac{1}{2}} \sum_{i=1}^p cX_i \left(Y_i - \left(1 + e^{-c(\Sigma^{\frac{1}{2}}\hat{\beta})^\top X_i} \right)^{-1} \right) - \hat{\beta}. \quad (5.6)$$

We use the observable vector $f_i(\hat{\beta}) = \hat{\beta}_i$, $1 \leq i \leq n$ and the sum of their corresponding asymptotic variances as the value to optimise with respect to Γ , but show in Figures 4 and 5 the estimated asymptotic variances for both sets $f_i(\hat{\beta})$, $f_i(\beta)$ of observables. In order to estimate each of the asymptotic variances, point evaluations of each $\nabla_p \phi$ are approximated as in the vector on the left of the outer product in (B.9), which are then used to approximate the asymptotic variances in accordance to

$$\sigma_f = 2 \int \phi(f - \pi(f)) \, d\tilde{\pi} = 2 \int \nabla \phi^\top \Gamma \nabla \phi \, d\tilde{\pi}. \quad (5.7)$$

Equation (5.7) follows from the formula (2.10) after integrating by parts. The approximation (B.5) for the term(s) including the Hessian in (B.4) has been used to test the method despite the explicit availability of the

¹<http://archive.ics.uci.edu/ml/datasets/Internet+Advertisements>. Note that besides missing values at some datapoints, the dataset comes with many quantitatively duplicate features and also some linear dependence between the vectors made up of a single feature across all datapoints; here features have been removed so that the said vectors remaining are linearly independently. In particular, $n = 642$.

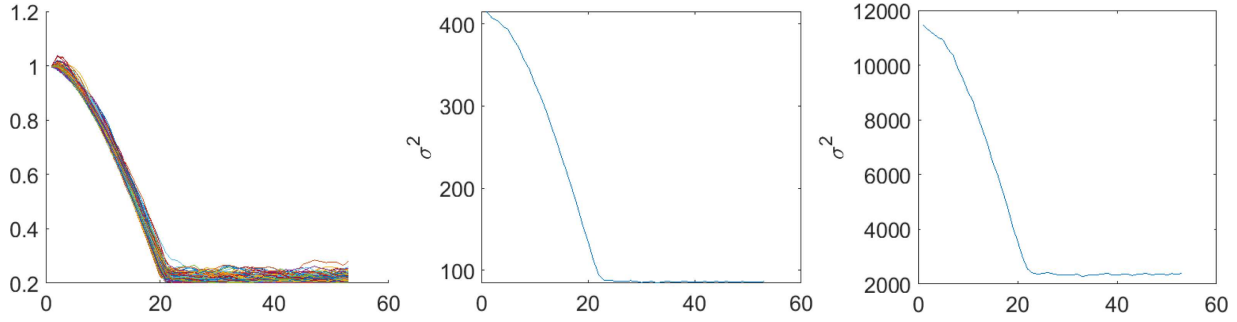


FIGURE 4. *Left*: diagonal values of Γ over iterations of (B.8) with $\alpha^i = 0.1$, $G = 1$, $r = 1$ and $\mu = 0.2$. Note that the mean of the absolute values of all entries of Γ at the end of the iterations is 0.0039. *Middle*: sum over i of estimated asymptotic variances for $f_i(\hat{\beta})$; *Right*: for $f_i(\beta)$. The simulation is based on the dataset with $n = 642$ and $p = 2359$.

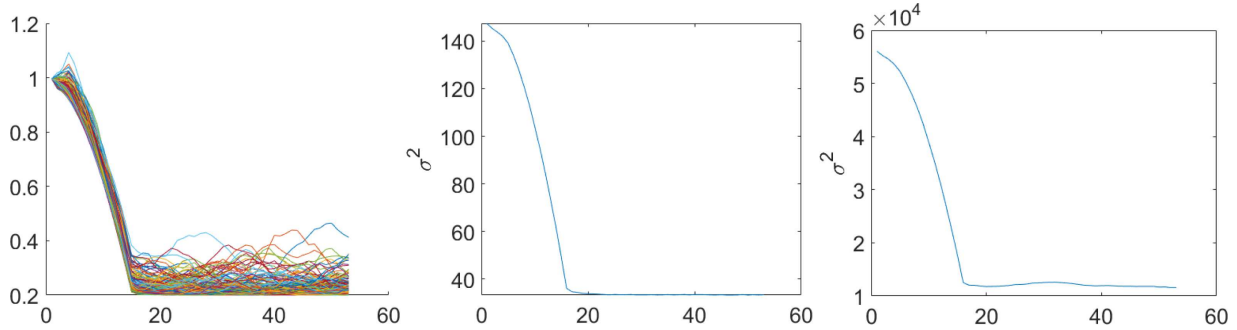


FIGURE 5. The same as in the caption of Figure 4, except $r = 0.5$ and a different dataset ([https://archive.ics.uci.edu/ml/datasets/Musk+\(Version+1\)](https://archive.ics.uci.edu/ml/datasets/Musk+(Version+1))) is used where $n = 167$ and $p = 476$. The mean of the absolute values of all entries of Γ at the end of the iterations is 0.0210.

Hessian. During the execution of Algorithm 2, the constant c in (5.6) has been set to

$$\bar{c} := \frac{5}{\max_i \left(\Sigma^{\frac{1}{2}} \sum_j X_j Y_j \right)_i}. \quad (5.8)$$

The particular choice of $c = \bar{c}$ is not important, except that it gives reasonable numerics at $\hat{\beta} = 0$, which can be seen by substitution into (5.6). In detail, 30 000 epochs are simulated; after 100 burn-in iterations of the Langevin discretisation (B.2), 2 parallel simulations of (B.2) and 2 of the first variation discretisation (B.4) are run according to Section B.1.3 with time-step $\Delta t = 0.1$, block-size $T = 100$, $L = 1$ block per update in Γ , $K = 1$ and tolerance $D_{\text{conv}} = 0.01$.

In Figures 4 and 5, Γ starts initially from the identity I_n and descends towards $0.2I_n$ (restricted as in (B.7)), as expected for a linear observable and potential close to a quadratic (see Cor. 4.9). We note that in the gradient descent procedure for Γ , using the minibatch gradient (the corresponding results of which are not shown here) does not change the behaviour shown in Figures 4 and 5. In addition, although the trajectory of Γ seems to go directly to zero, we expect the optimal Γ to be close but away from zero since the potential is close but not exactly quadratic.

TABLE 2. $(\frac{1}{n} \sum_{k=1}^n \sigma_{k,\text{approx}}^2, \frac{1}{n} \sum_{k=1}^n (\sigma_{k,\text{approx}}^2 - \frac{1}{n} \sum_{l=1}^n \sigma_{l,\text{approx}}^2)^2)$ – Empirical asymptotic variances, mean and variance over observable entries, where full gradients have been used for the dataset with $n = 642$, $p = 2359$.

	Block-size $T = 300$	Block-size $T = 9900$
$\Gamma = I_n$	(1.2669, 0.0320)	(0.8667, 0.7190)
$\Gamma = 0.2I_n$	(0.2939, 0.0018)	(0.1571, 0.0243)
$\Gamma = 0.1I_n$	(0.1739, 0.0007)	(0.0890, 0.0092)
Overdamped	(1.2298, 0.0319)	(0.8687, 0.8662)
Irreversible overdamped	(0.5642, 0.0077)	(0.3835, 0.1614)

Next, the value for Γ is fixed at various values and used for hyperparameter training on the same problem for the first dataset, using both the full gradient (5.6) and a minibatch² version where the sum in (5.6) is replaced by $\frac{p}{10}$ times a sum over a subset S of $\{1, \dots, p\}$ with 10 elements randomly drawn without replacement such that S changes once for each i in (B.2). In the minibatch gradient case, c is set to a fraction of (5.8), specifically $\bar{c}(\frac{p}{10})^{-1}$. In Tables 2 and 3, variances for the posterior mean estimates are shown (similar variance reduction results persist when using the probability of success for features taken from a single datapoint in the dataset).

In detail, for each row of Tables 2 and 3, $N = 29\,700$ epochs of (B.2) are simulated with the same parameters as above. The asymptotic variance for each observable entry is approximated using block averaging (Sect. 2.3.1.3 in [52]) by grouping the epochs after $B = 100$ burn-in iterations into $N_B = 99$ blocks of $T = 300$ epochs, that is,

$$\sigma_{k,\text{approx}}^2 = \frac{1}{N_B} \sum_{l=0}^{N_B-1} \left[\frac{1}{\sqrt{T}\Delta t} \sum_{i=1}^T \Delta t \left(f_k(q^{i+Tl+B}) - \frac{1}{N} \sum_{j=1}^N f_k(q^{j+B}) \right) \right]^2$$

and $N_B = 3$ blocks of $T = 9900$ epochs (respectively for each column of Tabs. 2 and 3); the values 0.8667 and 0.1571 approach and correspond to values in the middle plot of Figure 4 after multiplying by $n = 642$. The variances are compared to those using a gradient oracle: unadjusted (overdamped) Langevin dynamics [25] and with an irreversible perturbation [23], where the antisymmetric matrix J is given by

$$J_{i,j} = \begin{cases} 1 & \text{if } j - i = 1 \text{ or } 1 - n, \\ -1 & \text{if } i - j = 1 \text{ or } 1 - n, \\ 0 & \text{otherwise} \end{cases}$$

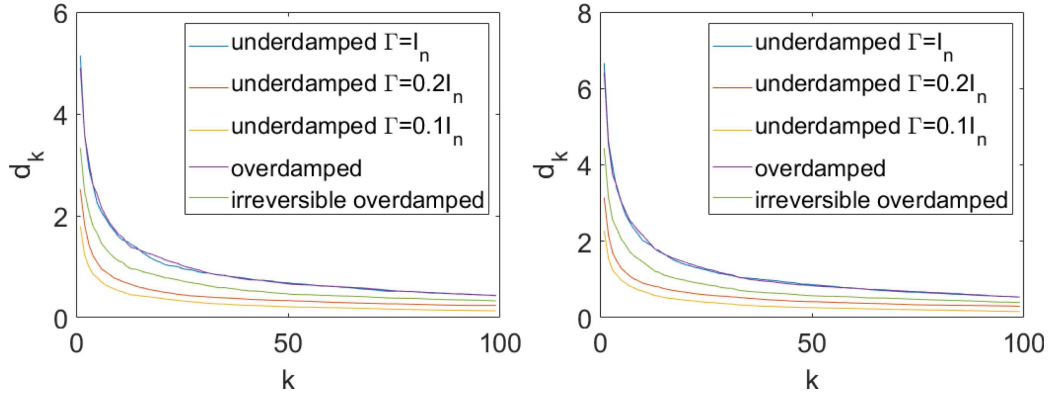
for $1 \leq i, j \leq n$ and the stepsizes are the same as for underdamped implementations. In addition, the Euclidean distance from intermediate estimates of the posterior mean to a total, combined estimate is shown for each method. Specifically, $d_k := \left| \frac{1}{300k} \sum_{i=1}^{300k} f(q^{i+B}) - \hat{\pi}(f) \right|$ is plotted against k in Figure 6, where $\hat{\pi}(f)$ is the mean (over the methods listed in Tabs. 2 and 3) of the final posterior mean estimates. A weighted mean with unit weights except one half on the $\Gamma = 0.2I_n$ and $\Gamma = 0.1I_n$ methods also gave similar results, though this is not shown explicitly.

These figures demonstrate improvement of an order of magnitude in observed variances for Γ close to that resulting from the gradient procedure over $\Gamma = I_n$. The improvement is also seen when compared to overdamped Langevin dynamics with and without irreversible perturbation.

²The control variate stochastic gradient on underdamped dynamics [14, 57] is not directly considered here but the benefits of an improved Γ is expected to carry over to such variations of the stochastic gradient.

TABLE 3. The same as in Table 2, except minibatch gradients have been used for the dataset with $n = 642$, $p = 2359$.

	Block-size $T = 300$	Block-size $T = 9900$
$\Gamma = I_n$	(1.9575, 0.0744)	(1.3338, 1.6650)
$\Gamma = 0.2I_n$	(0.4600, 0.0042)	(0.2781, 0.0784)
$\Gamma = 0.1I_n$	(0.2646, 0.0016)	(0.1335, 0.0208)
Overdamped	(1.9137, 0.0791)	(1.3065, 1.9714)
Irreversible overdamped	(0.8764, 0.0150)	(0.5778, 0.3266)

FIGURE 6. Euclidean distances to a combined posterior mean estimate over time. *Left*: full gradient. *Right*: minibatch gradient.

6. PROOFS

Some additional preliminaries are presented here for the proof of Theorem 3.2. For small $\epsilon \in \mathbb{R}$ and some direction $\delta\Gamma \in \mathbb{R}^{n \times n}$ such that $\Gamma + \epsilon\delta\Gamma \in \mathbb{S}_{++}^n$, let L_ϵ be the infinitesimal generator of (1.1) with the perturbed friction matrix $\Gamma + \epsilon\delta\Gamma$ in place of Γ , given formally by the differential operator

$$-\mathcal{L}_\epsilon = -p^\top M^{-1} \nabla_q + \nabla U(q)^\top \nabla_p + p^\top M^{-1} (\Gamma + \epsilon\delta\Gamma) \nabla_p - \nabla_p^\top (\Gamma + \epsilon\delta\Gamma) \nabla_p.$$

The formal $L^2(\tilde{\pi})$ -adjoint of \mathcal{L}_ϵ is denoted

$$-\mathcal{L}_\epsilon^* = p^\top M^{-1} \nabla_q - \nabla U(q)^\top \nabla_p + p^\top M^{-1} (\Gamma + \epsilon\delta\Gamma) \nabla_p - \nabla_p^\top (\Gamma + \epsilon\delta\Gamma) \nabla_p$$

just as for \mathcal{L}^* .

Proof of Theorem 3.2. For $\epsilon \leq \epsilon'$, by Theorem 2.5 there exists a solution $\phi + \delta\phi_\epsilon \in L_0^2(\tilde{\pi})$ to the Poisson equation with the perturbed generator $-L_\epsilon(\phi + \delta\phi_\epsilon) = f - \pi(f)$. By Theorem 2.6, the directional derivative of $\sigma^2(\Gamma)$ in the direction $\delta\Gamma : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ is

$$\frac{1}{2} d\sigma^2 \cdot \delta\Gamma = \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \int \delta\phi_\epsilon f d\tilde{\pi}. \quad (6.1)$$

By Proposition A.3, there are $\phi_k, \phi_{k,\epsilon} \in C_c^\infty$ such that $(\phi_k, -\mathcal{L}\phi_k)_{k \in \mathbb{N}}$, $(\phi_{k,\epsilon}, -\mathcal{L}_\epsilon\phi_{k,\epsilon})$ are approximating sequences to $(\phi, f - \pi(f))$, $(\phi + \delta\phi_\epsilon, f - \pi(f))$ respectively in $L^2(\tilde{\pi})^2$. Furthermore, in the same way as in the proof of Lemma 3.1 to obtain (A.11), it holds that

$$\|\nabla_p \phi_k - \nabla_p \phi\|_{L^2(\tilde{\pi})} + \|\nabla_p \phi_{k,\epsilon} - \nabla_p(\phi + \delta\phi_\epsilon)\|_{L^2(\tilde{\pi})} \rightarrow 0 \quad \text{as } k \rightarrow \infty. \quad (6.2)$$

Using the obvious extension on the notation from (1.8),

$$\int (\phi_{k,\epsilon} - \phi_k)(f - \pi(f)) d\tilde{\pi} = \int (\phi_{k,\epsilon} - \phi_k) \left(f - \pi(f) + \mathcal{L}^* \tilde{\phi}_k \right) d\tilde{\pi} - \int (\phi_{k,\epsilon} - \phi_k) \mathcal{L}^* \tilde{\phi}_k d\tilde{\pi}, \quad (6.3)$$

where the first term on the right hand side is negligible as $k \rightarrow \infty$ for any fixed ϵ due to $\mathcal{L}^* \tilde{\phi}_k = \widetilde{\mathcal{L} \phi_k}$ and the second term is

$$- \int (\phi_{k,\epsilon} - \phi_k) \mathcal{L}^* \tilde{\phi}_k d\tilde{\pi} = \int (-\mathcal{L}_\epsilon \phi_{k,\epsilon} + \mathcal{L} \phi_k) \tilde{\phi}_k d\tilde{\pi} - \int \epsilon (M^{-1}p - \nabla_p)^\top \delta \Gamma \nabla_p \phi_{k,\epsilon} \tilde{\phi}_k d\tilde{\pi}.$$

Again, the first term on the right hand side is negligible for any fixed ϵ as $k \rightarrow \infty$ since both terms in the bracket converge to $\pm(f - \pi(f))$. Integration by parts on the last term gives

$$- \int \epsilon (M^{-1}p - \nabla_p)^\top \delta \Gamma \nabla_p \phi_{k,\epsilon} \tilde{\phi}_k d\tilde{\pi} = - \int \epsilon \nabla_p \phi_{k,\epsilon}^\top \delta \Gamma \nabla_p \tilde{\phi}_k d\tilde{\pi}.$$

Collecting the above, for any fixed ϵ , taking $k \rightarrow \infty$ and using (6.2),

$$\int \delta \phi_\epsilon f d\tilde{\pi} = - \int \epsilon \nabla_p \phi^\top \delta \Gamma \nabla_p (\tilde{\phi} + \delta \tilde{\phi}_\epsilon) d\tilde{\pi}$$

holds. Plugging into (6.1), the directional derivative becomes

$$\frac{1}{2} d\sigma^2 \cdot \delta \Gamma = - \lim_{\epsilon \rightarrow 0} \int \nabla_p \phi^\top \delta \Gamma \nabla_p (\tilde{\phi} + \delta \tilde{\phi}_\epsilon) d\tilde{\pi}. \quad (6.4)$$

From here, for any ϵ , the unwanted term under the limit can be controlled by approximating again with $\tilde{\phi}_{k,\epsilon}$,

$$\begin{aligned} \lambda_m \int \left| \nabla_p (\tilde{\phi}_{k,\epsilon} - \tilde{\phi}_k) \right|^2 d\tilde{\pi} &\leq \int \nabla_p (\tilde{\phi}_{k,\epsilon} - \tilde{\phi}_k)^\top (\Gamma + \epsilon \delta \Gamma) \nabla_p (\tilde{\phi}_{k,\epsilon} - \tilde{\phi}_k) d\tilde{\pi} \\ &= \int (\tilde{\phi}_{k,\epsilon} - \tilde{\phi}_k) (M^{-1}p - \nabla_p)^\top (\Gamma + \epsilon \delta \Gamma) \nabla_p (\tilde{\phi}_{k,\epsilon} - \tilde{\phi}_k) d\tilde{\pi}, \\ &= - \int (\tilde{\phi}_{k,\epsilon} - \tilde{\phi}_k) \mathcal{L}_\epsilon^* (\tilde{\phi}_{k,\epsilon} - \tilde{\phi}_k) d\tilde{\pi} \\ &= - \epsilon \int (\tilde{\phi}_{k,\epsilon} - \tilde{\phi}_k) (M^{-1}p - \nabla_p)^\top \delta \Gamma \nabla_p \tilde{\phi}_k d\tilde{\pi} \\ &= - \epsilon \int \nabla_p (\tilde{\phi}_{k,\epsilon} - \tilde{\phi}_k)^\top \delta \Gamma \nabla_p \tilde{\phi}_k d\tilde{\pi} \\ &\leq \epsilon C \int \left(\left| \nabla_p (\tilde{\phi}_{k,\epsilon} - \tilde{\phi}_k) \right|^2 + \left| \nabla_p \tilde{\phi}_k \right|^2 \right) d\tilde{\pi}, \end{aligned}$$

where $\lambda_m = \inf_{0 < \epsilon \leq \epsilon'} \lambda_m^\epsilon$, λ_m^ϵ is the smallest eigenvalue of $\Gamma + \epsilon \delta \Gamma$ and $C > 0$ is a constant depending on $\delta \Gamma$ and independent of k . Therefore taking $k \rightarrow \infty$ and using (6.2) gives

$$\int \left| \nabla_p \delta \tilde{\phi}_\epsilon \right|^2 d\tilde{\pi} \leq \frac{\epsilon C}{\lambda_m - \epsilon C} \int \left| \nabla_p \tilde{\phi} \right|^2 d\tilde{\pi}$$

holds for small enough ϵ and putting into (6.4) concludes the proof. \square

Proof of Lemma 4.1. Substituting (4.3), (4.4) and (4.1) into the Poisson equation (1.6), one obtains

$$- \begin{pmatrix} 0 & M^{-1} \\ -\Sigma^{-1} & -\Gamma M^{-1} \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix} \cdot \begin{pmatrix} G_S q + E p + g \\ E^\top q + H_S p + h \end{pmatrix} - \Gamma : H_S = \frac{1}{2} q \cdot U_0 q + l \cdot q - \frac{1}{2} U_0 : \Sigma.$$

Comparing the constant, first order and second order coefficients in p give respectively the sufficient conditions (4.5)–(4.7) as stated. \square

Proof of Lemma 4.2. Comparing coefficients in q in equation (4.5) gives

$$2\Gamma : H_S = U_0 : \Sigma \quad (6.5)$$

$$h^\top \Sigma^{-1} = l^\top \quad (6.6)$$

$$2E\Sigma^{-1} = U_0 + A_2 \quad (6.7)$$

and the same for condition (4.6) gives

$$M^{-1}G_S = H_S\Sigma^{-1} + M^{-1}\Gamma E^\top, \quad (6.8)$$

$$M^{-1}g = M^{-1}\Gamma h. \quad (6.9)$$

Condition (6.7) yields (4.9). Together with (4.7), this gives (4.10). From the expression (4.10) and by symmetry of U_0 , condition (6.5) is in turn satisfied:

$$\begin{aligned} 2\Gamma : H_S &= \Gamma : ((\Sigma U_0 - \Sigma A_2 - 2A_1 M)\Gamma^{-1}) \\ &= \sum_{i,j,k,l} \Gamma_{ji} (\Sigma_{ik}(U_0)_{kl} - \Sigma_{ik}(A_2)_{kl} - (A_1)_{ik} M_{kl}) (\Gamma^{-1})_{lj} \\ &= \sum_{i,k} (U_0)_{ki} \Sigma_{ki} = U_0 : \Sigma, \end{aligned}$$

where symmetry of Σ and M have been used. Substituting (4.9) and (4.10) into equation (6.8) then gives (4.8). Equations (6.6) and (6.9) give the equations (4.11) for g and h . \square

Proof of Lemma 4.3. Denote

$$\bar{G} = \begin{pmatrix} G_S & E \\ E^\top & H_S \end{pmatrix}, \quad \bar{U}_0 = \begin{pmatrix} U_0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \bar{\Sigma} = \begin{pmatrix} \Sigma & 0 \\ 0 & M \end{pmatrix}, \quad \bar{g} = \begin{pmatrix} g \\ h \end{pmatrix}, \quad \bar{l} = \begin{pmatrix} l \\ 0 \end{pmatrix}.$$

Each of ϕ in (4.4) and $f - \pi(f)$ from (4.1), (4.2) are given by

$$\begin{aligned} \phi(z) &= \frac{1}{2} z \cdot \bar{G} z - \bar{g} \cdot z - \frac{1}{2} \bar{G} : \bar{\Sigma} \\ f(z) - \pi(f) &= \frac{1}{2} z \cdot \bar{U}_0 z - \bar{l} \cdot z - \frac{1}{2} \bar{U}_0 : \bar{\Sigma} \end{aligned}$$

for $z = (q, p) \in \mathbb{R}^{2n}$. Substituting into $\sigma^2 = 2\langle \phi, f - \pi(f) \rangle_{\bar{\pi}}$ gives

$$\begin{aligned} 2 \int \phi(f - \pi(f)) \, d\bar{\pi} &= \frac{1}{2} \int (z \cdot \bar{G} z) (z \cdot \bar{U}_0 z) \, d\bar{\pi} - \frac{1}{2} \int (z \cdot \bar{G} z) \bar{U}_0 : \bar{\Sigma} \, d\bar{\pi} + 2 \int (\bar{g} \cdot z) (\bar{l} \cdot z) \, d\bar{\pi} \\ &\quad - \frac{1}{2} \int \bar{G} : \bar{\Sigma} (z \cdot \bar{U}_0 z) \, d\bar{\pi} + \frac{1}{2} (\bar{G} : \bar{\Sigma}) (\bar{U}_0 : \bar{\Sigma}), \end{aligned}$$

where

$$\begin{aligned} \int (z \cdot \bar{G} z) (z \cdot \bar{U}_0 z) \, d\bar{\pi} &= \sum_{i,j,u,v} \bar{G}_{ij} (\bar{U}_0)_{uv} \int z_i z_j z_u z_v \, d\bar{\pi} \\ &= \sum_{i,j,u,v} \bar{G}_{ij} (\bar{U}_0)_{uv} (\bar{\Sigma}_{ij} \bar{\Sigma}_{uv} + \bar{\Sigma}_{iu} \bar{\Sigma}_{jv} + \bar{\Sigma}_{iv} \bar{\Sigma}_{ju}) \\ &= (\bar{G} : \bar{\Sigma}) (\bar{U}_0 : \bar{\Sigma}) + 2\text{Tr}(\bar{G} \bar{\Sigma} \bar{U}_0 \bar{\Sigma}). \end{aligned}$$

As a result,

$$\begin{aligned} 2 \int \phi(f - \pi(f)) \, d\tilde{\pi} &= \frac{1}{2}(\bar{G} : \bar{\Sigma})(\bar{U}_0 : \bar{\Sigma}) + \text{Tr}(\bar{G}\bar{\Sigma}\bar{U}_0\bar{\Sigma}) - \frac{1}{2}(\bar{G} : \bar{\Sigma})(\bar{U}_0 : \bar{\Sigma}) \\ &\quad + 2 \int (\bar{g} \cdot z)(\bar{l} \cdot z) \, d\tilde{\pi} \\ &= \text{Tr}(\bar{G}\bar{\Sigma}\bar{U}_0\bar{\Sigma}) + 2\bar{g} \cdot \bar{\Sigma}\bar{l}. \end{aligned}$$

□

Proof of Proposition 4.8. Let $\Gamma = \gamma I_n$, $\gamma \in \mathbb{R}$. Note there is a unique solution $\phi \in L_0^2(\tilde{\pi})$ to (1.6) by Theorem 2.5. The solution ϕ to (1.6) has the expression $\phi = \sum_i \alpha_i(\gamma q_i + p_i)$. The asymptotic variance is equal to

$$\begin{aligned} 2\langle \phi, f - \pi(f) \rangle_{\tilde{\pi}} &= 2\gamma \sum_{i,j} \alpha_i \alpha_j \int_{\mathbb{R}^n} q_i \partial_{q_j} U(q) \pi(dq) \\ &= -2\gamma \sum_i \alpha_i^2 \int_{\mathbb{R}^n} q_i \partial_{q_i} \pi(q) \, dq - 2\gamma \sum_{i \neq j} \alpha_i \alpha_j \int_{\mathbb{R}^n} q_i \partial_{q_j} \pi(q) \, dq \\ &= 2\gamma \sum_i \alpha_i^2 \int_{\mathbb{R}^n} \pi(q) \, dq - 2\gamma \sum_{i \neq j} \alpha_i \alpha_j \int_{\mathbb{R}^{n-1}} q_i \int_{\mathbb{R}} \partial_{q_j} \pi(q) \, dq_j \, dq_{-j} \\ &= 2\gamma \sum_i \alpha_i^2 \end{aligned}$$

where dq_{-j} denotes $dq_1 \dots dq_{j-1} dq_{j+1} \dots dq_n$. Taking $\gamma \rightarrow 0$ gives (4.18). □

For the proof of Proposition 4.10, some notation is introduced. For $\tilde{k} \in \mathbb{N}_0$, let the tridiagonal matrix $M_{\tilde{k}} \in \mathbb{R}^{(\tilde{k}+1) \times (\tilde{k}+1)}$ be given by its elements

$$(M_{\tilde{k}})_{i,j} = \begin{cases} i & \text{if } i+1 = j, \\ (i-1)\gamma & \text{if } i = j, \\ i - \tilde{k} - 2 & \text{if } i-1 = j, \\ 0 & \text{otherwise} \end{cases} \quad (6.10)$$

for indices $1 \leq i, j \leq \tilde{k} + 1$.

Lemma 6.1. *Let $m \in \mathbb{N}$ and let $\tilde{M} \in \mathbb{R}^{m \times m}$ be a tridiagonal matrix of the form*

$$(\tilde{M})_{i,j} = \begin{cases} b_i & \text{if } i+1 = j, \\ b'_i \gamma & \text{if } i = j, \\ b''_i & \text{if } i-1 = j, \\ 0 & \text{otherwise} \end{cases}$$

for constants $b'_i \in \mathbb{R}$, $b_i, b''_i \in \mathbb{R} \setminus \{0\}$. The following statements hold.

- (i) *If m is odd, then $\det(\tilde{M}) = \mathcal{O}(\gamma)$ as $\gamma \rightarrow 0$.*
- (ii) *If m is even, then $\lim_{\gamma \rightarrow 0} \det(\tilde{M}) \neq 0$.*

Lemma 6.1 is straightforwardly proved by repeatedly taking Laplace expansions. An explicit proof is not given here.

Proof of Proposition 4.10. Only a standard Gaussian and $M = 1$ is considered, the arguments for the general centered Gaussian case are the same. First consider the observable

$$f(q) = q^k \quad (6.11)$$

for some odd $k \in \mathbb{N}_0$. Take the polynomial ansatz

$$\phi(q, p) = \sum_{i,j=0}^k a_{i,j} q^i p^j \quad (6.12)$$

for $a_{i,j} \in \mathbb{R}$ and $\Gamma = \gamma > 0$. It will be shown that arbitrarily small asymptotic variance is achieved in the $\gamma \rightarrow 0$ limit. Note that only pairs (i, j) with odd i and even j make nonzero contributions to the asymptotic variance. Applying $-\mathcal{L}$ to the ansatz,

$$\begin{aligned} -\mathcal{L}\phi &= \sum_{i,j=0}^k -ia_{i,j} q^{i-1} p^{j+1} + ja_{i,j} q^{i+1} p^{j-1} + \gamma ja_{i,j} q^i p^j - \gamma j(j-1) a_{i,j} q^i p^{j-2} \\ &= \sum_{i,j=0}^{k+1} (-(i+1)a_{i+1,j-1} + (j+1)a_{i-1,j+1} + \gamma ja_{i,j} - \gamma(j+2)(j+1)a_{i,j+2}) q^i p^j \end{aligned}$$

where

$$a_{i,j} = 0 \quad \forall i, j < 0 \text{ and } \forall i, j > k. \quad (6.13)$$

Comparing with coefficients in (6.11) via (1.6),

$$-(i+1)a_{i+1,j-1} + (j+1)a_{i-1,j+1} + \gamma ja_{i,j} - \gamma(j+2)(j+1)a_{i,j+2} = 0 \quad (6.14)$$

for all $(i, j) \neq (k, 0)$. It holds by strong induction (in j') that

$$a_{i'+j',k+1-j'} = 0 \quad \forall i', j' \geq 0 \quad (6.15)$$

because of the following. The base case $j' = 0$ follows by (6.13), the induction step follows by taking $(i, j) = (i' + j' - 1, k + 2 - j')$ for $i' \geq 0$ in (6.14) and again using (6.13) where necessary; an illustration is given in Figure 7.

Comparing coefficients in the Poisson equation (1.6) for $(i, j) = (k, 0)$ and using (6.13), (6.15) yields

$$a_{k-1,1} = 1. \quad (6.16)$$

Combining (6.16) with setting $(i, j) = (j' - 1, k + 1 - j')$ for $j' = 1, \dots, k$ in (6.14), the entries $a_{j',k-j'}$ satisfy the linear system

$$M_k(a_{k,0}, a_{k-1,1}, \dots, a_{0,k})^\top = (1, 0, \dots, 0)^\top, \quad (6.17)$$

where $M_k \in \mathbb{R}^{k+1 \times k+1}$ is the tridiagonal matrix given in (6.10). In order to find the order in γ as $\gamma \rightarrow 0$ of the elements of $(a_{k,0}, \dots, a_{0,k})^\top$ appearing in (6.17), it suffices to find the order of the entries in the leftmost column of M_k^{-1} . For this, let $C_i \in \mathbb{R}$ be the i th minor appearing in the top row of the cofactor matrix of M_k . On the corresponding submatrix, repeatedly taking the Laplace expansion on the leftmost column until only the determinant of a $(k+1-i)$ -by- $(k+1-i)$ square matrix from the bottom right corner of M_k remains to be calculated, then using Lemma 6.1 for this $(k+1-i)$ -by- $(k+1-i)$ matrix gives that C_i is of order γ as $\gamma \rightarrow 0$ for odd i . Furthermore, the determinant of M_k is bounded away from zero as $\gamma \rightarrow 0$ by Lemma 6.1. Therefore the elements of $(a_{k,0}, \dots, a_{0,k})$ in the left hand side of (6.17) with an odd index, that is $a_{k-j,j}$ for even j , have order γ and at most order 1 otherwise as $\gamma \rightarrow 0$. These elements with odd indices are exactly those from the

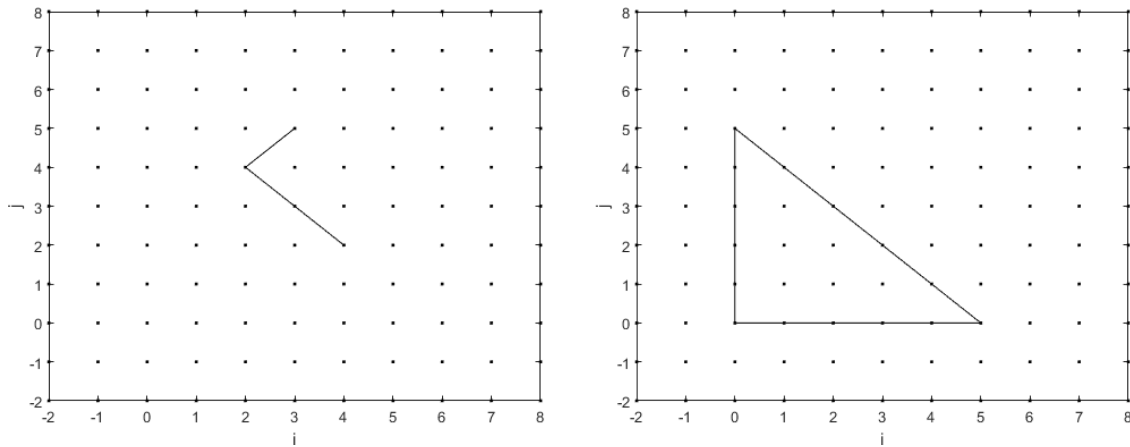


FIGURE 7. *Left*: an illustration of the relation (6.14) with $(i, j) = (3, 3)$, where each dot represents a coefficient $a_{i,j}$. *Right*: points on and inside the triangle show in the case $k = 5$ all of the coefficients that can be nonzero. All other points outside of the triangle are asserted to be zero by (6.13) and (6.15). The induction step for proving (6.15) amounts to translating the chain on the left figure horizontally, in order to show that the point at the bottom of each translated chain is zero by the fact that the other points in the same chain are zero.

vector $(a_{k,0}, \dots, a_{0,k})^\top$ that make a contribution to the asymptotic variance. The “next” set of contributions come from the vector $(a_{k-2,0}, a_{k-3,1}, \dots, a_{0,k-2})$. Using again (6.13) and (6.14), the vector satisfies

$$M_{k-2}(a_{k-2,0}, a_{k-3,1}, \dots, a_{0,k-2})^\top = v_{k-2},$$

for some vector v_{k-2} (from the last term on the left hand side of (6.14)) of order γ as $\gamma \rightarrow 0$ and since the determinant of M_{k-2} is of order 1 (by Lem. 6.1), the contributions here to the asymptotic variance are again of order γ . Continuing for $(a_{k-2j,0}, a_{k-2j-1,1}, \dots, a_{0,k-2j})^\top$, $j \in \mathbb{N}$, it follows that all contributions are of order γ as $\gamma \rightarrow 0$. The resulting coefficients indeed make up a solution ϕ to the Poisson equation because the matrices M_k are invertible and because the coefficients $a_{i,j}$ for even $i+j$ are equal to zero from repeating the above procedure for the coefficients associated to M_{k-1} , M_{k-3} and so on.

For the general case of (4.19), since \mathcal{L} is a linear differential operator and the contributions to the value of $\int \phi(f - \pi(f)) d\tilde{\pi}$ come from exactly the same (odd i , even j) $a_{i,j}$ coefficients from the corresponding solution ϕ to each summand in (4.19), the proof concludes. \square

Proof of Proposition 4.11. Take the polynomial ansatz

$$\phi(q, p) = \sum_{i,j=0}^4 a_{i,j} q^i p^j \quad (6.18)$$

for $a_{i,j} \in \mathbb{R}$, where $a_{i,j}$ not appearing in the sum are taken to be zero in the following. Again, only the standard Gaussian is considered, it turns out the arguments follow similarly otherwise. Comparing coefficients in (1.6) and using the same strong induction argument as in the proof of Proposition 4.10 leads to (6.14) for all $(i, j) \neq (4, 0), (0, 0)$ and equation (6.15). Taking $(i, j) = (j' - 1, 5 - j')$ for $1 \leq j' \leq 4$ in (6.14) and comparing the q^4 coefficients in the Poisson equation, it holds that

$$M_4(a_{4,0}, a_{3,1}, a_{2,2}, a_{1,3}, a_{0,4})^\top = (1, 0, \dots, 0)^\top \quad (6.19)$$

and taking $(i, j) = (j' - 1, 3 - j')$ for $j' \geq 1$ in (6.14) yields

$$M_2(a_{2,0}, a_{1,1}, a_{0,2})^\top = \gamma(2a_{2,2}, 6a_{1,3}, 12a_{0,4})^\top. \quad (6.20)$$

Equations (6.19), (6.20) can be solved explicitly and the asymptotic variance is a weighted sum of the resulting coefficients. Those in (6.18) that make contributions are $a_{4,0}, a_{2,2}, a_{2,0}$, which gives the asymptotic variance $\frac{12(21\gamma^4 + 55\gamma^2 + 27)}{\gamma(3\gamma^2 + 4)}$ that goes to infinity as $\gamma \rightarrow 0$ or $\gamma \rightarrow \infty$; note that the coefficients associated to M_3 and M_1 are zero by a similar procedure as above. Finally, to check that the solution coefficients given by (6.19) and (6.20) indeed solve the Poisson equation, it remains to show the constant coefficient of $\mathcal{L}\phi$ with these solution coefficients matches that of the right-hand side of the Poisson equation. The constant term on the right-hand side of the Poisson equation is given by

$$\int q^4 \frac{e^{\frac{q^2}{2}}}{\sqrt{2\pi}} dq = 3,$$

which is readily verified to be equal to $2\Gamma a_{0,2}$ from (6.19) and (6.20). \square

7. DISCUSSION

7.1. The nonconvex case

In the case where U is nonconvex, the Monte Carlo procedure in Section B.1.3 may continue to be used as presented, however the first variation process (that is, the solution to (3.3)) could easily stray from the case of exponential decay as in Theorem 3.3. Transitions from one metastable state to another cause the tangent process to increase in magnitude. In a one dimension double well potential $U(q) = \frac{q^4}{4} - q^2 + \frac{q}{2}$, linear observable $f(q) = q$ case, these transitions occur frequently enough during the gradient procedure in Γ that Dq blows up in simulation. Even in cases for which the metastabilities are strong, so that transitions occur less frequently, simulations show that Γ dives to zero in periods where no transitions are occurring (as if the case of Cor. 4.9), but increase dramatically in value once a transition does occur, causing the trajectory in Γ to decay over time but occasionally jumping in value, so that there is no convergence for Γ . On the other hand, a Galerkin approach tends to give good convergence for Γ in such cases.

7.2. Position-dependent friction

It is possible to adapt the formula (3.1) to the case of position-dependent gradient direction in Γ . The gradient direction is the same as (1.9) with the change that the integral is replaced by the corresponding marginal integral in p . Ideas using such a formula need to take into account that the first variation process retains a non-vanishing stochastic integral with respect to Brownian motion, so that the truncation in calculating the corresponding infinite time integral in Section B.1.3 is not as well justified, or rather, does not happen in the execution of Algorithm 2 due to (B.11) not being satisfied.

7.3. Metropolisation

Throughout Section 5, the implementation has not involved accept-reject steps. Metropolisation of discretisations of the underdamped Langevin dynamics was given in [42], see also Section 2.2.3.2 in [52] and [54, 65]. The systematic discretisation error is removed with the inclusion of this step but the momentum is reversed upon rejection (to avoid high rejection rates [65]), which raises the question of whether friction matrices arising from Algorithm 1 improve the Metropolised situation where dynamics no longer imitate those in the continuous-time. For example the intuition in the Gaussian target measure, linear observable case discussed in Section 4.2 no longer applies.

7.4. Possible connections for future work

Taking $\Gamma \rightarrow \infty$ together with a time rescaling, the dynamics (1.1) become the overdamped Langevin equation [60]. An analogous result holds [43] when $\Gamma = \Gamma(q)$ is position dependent, where a preconditioner for the corresponding overdamped dynamics appears in terms of Γ^{-1} ; see Section 7.2 for a consideration of our method in the position dependent friction case. On the other hand, the Hessian of U makes a good preconditioner in the overdamped dynamics because of the Brascamp-Lieb inequality, see Remark 1 in [2]. To our knowledge, it is an open question whether optimisers for our problem are related in the limit to optimisers for analogous problems for the overdamped dynamics.

The infinite time integral (1.10) has been used for the calculation of transport coefficients in molecular dynamics [50, 59] and the derivative of the expectation appearing in (1.10) with respect to initial conditions is a problem considered when calculating the “greeks” in mathematical finance [30]. On the topic of the latter and in contrast to [30], there is previous work dealing with cases of degenerate noise in the system, some of these references are given in Remark B.2.

Variance reduction by modifying the observable instead of changing the dynamics has been considered for example in [5, 6, 68]. As of writing, the methods there are not directly compatible with the framework in the present work due to the improved observable being unknown before the simulation of the Markov chain.

Finally, we mention that underdamped Langevin dynamics has been considered with (variance reduced) stochastic gradients. In [70], the authors present a comparison between such an application with Hamiltonian Monte Carlo. In [14], convergence guarantees are provided for control variate stochastic gradients in underdamped Langevin dynamics, along with numerical comparisons in low dimensional, tall dataset regimes. Furthermore, the underdamped dynamics with single, randomly selected component gradient update in place of the full gradient is considered in [21].

7.5. Conclusion

We have presented the central limit theorem for the underdamped Langevin dynamics and provided a formula for the directional derivative of the corresponding asymptotic variance with respect to a friction matrix Γ . A number of methods for approximating the gradient direction in Γ have been discussed together with numerical results giving improved observed variances. Some cases where an improved friction matrix can be explicitly found have been given to guide the expectation of an optimal Γ . In particular, in cases where the observable is linear and the potential is close to quadratic, which is the case when finding the posterior mean in Bayesian inference with Gaussian priors, the optimal friction is expected to be close to zero (due to Cor. 4.9). This is consistent with the numerical results from Algorithm 2. Moreover, it is shown that the improvement in variance is retained when using minibatch stochastic gradients in a case of Bayesian inference.

We mention that the gradient procedure using (1.7) and (1.11) can be used to guide Γ in arbitrarily high dimension; given a high dimensional problem of interest, the gradient procedure can be used on similar, intermediate dimensional problems in order to obtain a friction matrix. In particular, for the Bayesian inference problem as formulated in Section 5.3, the algorithm recommends the choice of a small friction scalar, which can be expected to apply for datasets in an arbitrary number of dimensions.

Future directions not mentioned above include the study of well-posedness of the optimisation in Γ , extension to higher-order Langevin samplers methods as in [13, 56] and gradient formulae in the discrete time case analogous to Theorem 3.2.

APPENDIX A. PRELIMINARIES

Theorem A.1. *Let Assumption 2.1 hold. For any \mathcal{F}_0 -measurable $z_0 : \Omega \rightarrow \mathbb{R}^{2n}$, there exists an almost surely continuous in t solution $(q_t, p_t) = z_t : \Omega \rightarrow \mathbb{R}^{2n}$ to (1.1) that is \mathcal{F}_t -adapted and unique up to equivalence. Furthermore, for any $z \in \mathbb{R}^{2n}$, $t \geq 0$, let ρ_t^z be the probability measure given by*

$$\rho_t^z(A) = \mathbb{P}(z_t^z \in A) \tag{A.1}$$

for any Borel measurable A , where z_t^z denotes the solution to (1.1) starting at $z_0 = z$, then ρ_t^z

- (1) is a transition probability in the sense that
 - (a) $(t, z) \mapsto \rho_t^z(A)$ is Borel measurable on $(0, \infty) \times \mathbb{R}^{2n}$,
 - (b) the Chapman–Kolmogorov relation [31] holds and;
- (2) admits a density denoted $\rho(z, \cdot, t) : \mathbb{R}^{2n} \rightarrow \mathbb{R}$ with respect to the Lebesgue measure on \mathbb{R}^{2n} at every $(t, z) \in (0, \infty) \times \mathbb{R}^{2n}$ such that ρ is a measurable function satisfying for every $z \in \mathbb{R}^{2n}$,

$$\rho(z, \cdot, \cdot) \in C^\infty(\mathbb{R}^{2n} \times (0, \infty)). \quad (\text{A.2})$$

Proof. Theorem 3.5 in [45] together with (2.8) yields existence and uniqueness of the solution to (1.1). Theorem 3.1 and 3.6 in Section 5 of [31] give that $\rho_t^z(A)$ given by (A.1) is a probability kernel, that is, $\rho_t^z(A)$ is Borel measurable in z for fixed A, t , is a probability measure in A for fixed z, t and satisfies the Chapman–Kolmogorov relation. For Borel measurability of $(t, z) \mapsto \rho_t^z(A)$ for fixed A , consider \hat{z}_t^z given by

$$\hat{z}_t^z(\omega) = \begin{cases} z_t^z(\omega) & \text{if } \omega : z_\bullet^z(\omega) \in C([0, \infty)), \\ 0 & \text{otherwise.} \end{cases} \quad (\text{A.3})$$

The process \hat{z}_t^z is continuous in t and \mathcal{F} -measurable in ω , therefore $\mathbb{P}(\hat{z}_t^z \in A) = \mathbb{P}(z_t^z \in A)$ is continuous in t hence Borel measurable in (t, z) . Finally, ρ_t^z admits a density at every $(t, z) \in (0, \infty) \times \mathbb{R}^{2n}$ satisfying (A.2) due to Itô's rule and Hörmander's theorem [40]; measurability with respect to the starting point z and therefore jointly in all of the arguments ([1], Lem. 4.51) follows by the strong Feller property given by Theorem 4.2 in [22], because $\rho(\cdot, \zeta, t)$ is the pointwise limit of the continuous functions $(\int \eta_k(\zeta - \zeta')\rho(\cdot, \zeta', t)d\zeta')_{k>0}$, where η_k denotes the standard scaled mollifiers. \square

For all $t \geq 0$, all $z \in \mathbb{R}^{2n}$ and all $f : \mathbb{R}^{2n} \rightarrow \mathbb{R}$ integrable under the law $\mathcal{L}((z_t)_{t \geq 0} | z_0 = z)$ of z_t starting at z , let

$$P_t(f) : z \mapsto \mathbb{E}(f(z_t^z)) = \mathbb{E}(f(z_t) | z_0 = z). \quad (\text{A.4})$$

The family $(P_t)_{t \geq 0}$ forms a strongly continuous (Prop. A.2) Markov semigroup on $L^2(\tilde{\pi})$ with unit operator norm. Denote by L the infinitesimal generator associated to this semigroup, given by

$$Lu = \lim_{t \rightarrow 0} \frac{P_t(u) - u}{t} \quad (\text{A.5})$$

for all functions $u \in \mathcal{D}(L) \subset L^2(\tilde{\pi})$, where the domain $\mathcal{D}(L)$ consists of the functions for which the above limit in $L^2(\tilde{\pi})$ exists.

Proposition A.2. *The family $(P_t)_{t \geq 0}$ is strongly continuous in $L^2(\tilde{\pi})$.*

Proof. Fix $\epsilon > 0$. For any $f \in L^2(\tilde{\pi})$, there exists $g \in C_c^\infty$ such that $\|f - g\|_{L^2(\tilde{\pi})} \leq \frac{\epsilon}{3}$. By triangle inequality, it holds that

$$\|P_t f - f\|_{L^2(\tilde{\pi})} \leq \|P_t f - P_t g\|_{L^2(\tilde{\pi})} + \|f - g\|_{L^2(\tilde{\pi})} + \|P_t g - g\|_{L^2(\tilde{\pi})}. \quad (\text{A.6})$$

The last term on the right hand side converges to 0 as $t \rightarrow 0$ by Itô's rule. Since the measures $\int \mathbb{E}[\mathbb{1}_{(z_t^z)}] \tilde{\pi}(dz)$ solve the associated Fokker–Planck equation in the distributional sense, it is equal to the unique solution $\tilde{\pi}$, therefore the first term on the right-hand side of (A.6) can be bounded by $\frac{\epsilon}{3}$ after Jensen's inequality. \square

Here, we give the definition of a maximally accretive operator in $L^2(\tilde{\pi})$ from Definition 5.2 of [39]. For a linear operator $A : \mathcal{D}(A) \rightarrow L^2(\tilde{\pi})$ defined on $\mathcal{D}(A) \subset L^2(\tilde{\pi})$, the operator A is called maximally accretive if

(i) it holds that

$$\int f A f \, d\tilde{\pi} \geq 0, \quad \forall f \in \mathcal{D}(A), \quad (\text{A.7})$$

(ii) and there does not exist an extension $\tilde{A} : \mathcal{D}(\tilde{A}) \rightarrow L^2(\tilde{\pi})$ of A with $\mathcal{D}(\tilde{A})$ strictly larger than $\mathcal{D}(A)$, satisfying (A.7) with A and $\mathcal{D}(A)$ replaced by \tilde{A} and $\mathcal{D}(\tilde{A})$.

In this work, we take for granted the results ([39], Thm. 5.4 [28], Thm. 2.12, Prop. 3.1) on operators with maximally accretive closures.

Proposition A.3. *The differential operator $-\mathcal{L}$ defined on C_c^∞ has a maximally accretive closure in $L^2(\tilde{\pi})$.*

Proof. Let K denote the differential operator

$$\begin{aligned} K &= e^{-\frac{1}{2}\left(U(q)+\frac{v^2}{2}\right)} \mathcal{L} \left(e^{\frac{1}{2}\left(U(q)+\frac{v^2}{2}\right)} \right) \\ &= p^\top M^{-1} \nabla_q - \nabla U(q)^\top \nabla_p + \frac{1}{2} \text{Tr} \Gamma - \frac{1}{4} p^\top \Gamma p + \nabla_p^\top \Gamma \nabla_p \end{aligned}$$

acting on C_c^∞ . By a straightforward adaptation of the proof of Proposition 5.5 in [39], the closure of $-K$ in $L^2(\mathbb{R}^{2n})$ and therefore the closure of $-\mathcal{L}$ in $L^2(\tilde{\pi})$ are maximally accretive. \square

Proof of Theorem 2.5. For $T > 0$, let $g_T := \int_0^T P_t(f) \, dt$. Note that $g_T \in L^2(\tilde{\pi})$ for $T \in \mathbb{R}_+ \cup \{\infty\}$ and by Theorem 2.2

$$g_T \rightarrow \int_0^\infty P_t(f) \, dt \quad (\text{A.8})$$

in $L^2(\tilde{\pi})$ as $T \rightarrow \infty$, specifically (2.5) with $\varphi = f$ and using (2.7) for $2l$ in place of l . Applying L , it holds that

$$Lg_T = \lim_{s \rightarrow 0} \frac{P_s(g_T) - g_T}{s} = \lim_{s \rightarrow 0} \frac{1}{s} \left(\int_s^{T+s} - \int_0^T \right) P_u(f) \, du = P_T(f) - f,$$

where the exchange in the order of integration is justified by Fubini, equation (2.5) and the last equality follows by the strong continuity of $(P_t)_{t \geq 0}$ (given by Prop. A.2). Inequalities (2.5) and (2.7) (with $2l$ in place of l) also give

$$P_T(f) \rightarrow 0 \quad \text{in } L^2(\tilde{\pi}) \quad (\text{A.9})$$

as $T \rightarrow \infty$, so that since L is a closed operator, equations (1.6) and (2.9) hold. In addition, $\int \phi \, d\tilde{\pi} = 0$ follows from the invariance of $\tilde{\pi}$, Theorem 2.2 and Fubini's theorem. \square

Proof of Lemma 3.1. By Proposition A.3, there are $\phi_k \in C_c^\infty$ such that $(\phi_k, -\mathcal{L}\phi_k)_{k \in \mathbb{N}}$ is an approximating sequence to $(\phi, -L\phi)$ in $L^2(\tilde{\pi})^2$. By integrating by parts and using the antisymmetric property of the relevant part of \mathcal{L} , we have

$$\begin{aligned} \lambda_m \int |\nabla_p \phi_k - \nabla_p \phi_{k'}|^2 \, d\tilde{\pi} &\leq \int \nabla_p(\phi_k - \phi_{k'})^\top \Gamma \nabla_p(\phi_k - \phi_{k'}) \, d\tilde{\pi} \\ &= - \int (\phi_k - \phi_{k'}) (\mathcal{L}\phi_k - \mathcal{L}\phi_{k'}) \, d\tilde{\pi}, \end{aligned} \quad (\text{A.10})$$

where λ_m is the smallest eigenvalue of Γ , so that $\nabla_p \phi_k$ is Cauchy in $L^2(\tilde{\pi})$, with limit denoted as $g \in L^2(\tilde{\pi})$. For any $h \in C_c^\infty$,

$$\left| \int gh + \int \phi \nabla_p h \right| \leq \left| \int gh - \int \nabla_p \phi_k h \right| + \left| \int \phi \nabla_p h - \int \phi_k \nabla_p h \right|,$$

hence

$$\nabla_p \phi_k \rightarrow g = \nabla_p \phi \in L^2(\tilde{\pi}). \quad (\text{A.11})$$

\square

APPENDIX B. DETAILS OF THE IMPLEMENTATION

B.1. Methodology

Here we describe an on-the-fly procedure to repeatedly calculate the change (1.9) in Γ by simulating the first variation process parallel to underdamped Langevin processes. The discretisation schemes used to simulate (1.1) and (3.3) are given in Section B.1.1. Two gradient procedures, namely gradient descent and the Heavy ball method, for evolving Γ given a gradient are detailed in Section B.1.2. Then iterates from Section B.1.1 are used to approximate each change in Γ in Section B.1.3. The key idea linking the above is that if equation (1.11) holds, then

$$\begin{aligned}\Delta\Gamma &= \int \nabla_p \phi \otimes \nabla_p \tilde{\phi} \, d\tilde{\pi} \\ &= - \int \left(\int_0^\infty \mathbb{E}[\nabla f(q_s)^\top D_p q_s]^\top ds \right) \left(\int_0^\infty \mathbb{E}[\nabla f(\tilde{q}_t)^\top D_p \tilde{q}_t] dt \right) d\tilde{\pi},\end{aligned}\quad (\text{B.1})$$

where (q_t, p_t) and $(\tilde{q}_t, \tilde{p}_t)$ denote the solutions to (1.1) with initial values (q, p) , $(q, -p)$ respectively, $(D_p q_t, D_p p_t)$ and $(D_p \tilde{q}_t, D_p \tilde{p}_t)$ denote the solutions to (B.3) with \tilde{q}_t replacing q_t for the latter and the integral in (B.1) is with respect to (q, p) .

B.1.1. Splitting

A BAOAB splitting scheme [48, 49] will be used to integrate the Langevin dynamics (1.1), given explicitly by

$$\begin{cases} p^{i+\frac{1}{3}} = p^i - \nabla U(q^i) \frac{\Delta t}{2} \\ q^{i+\frac{1}{2}} = q^i + p^{i+\frac{1}{3}} \frac{\Delta t}{2} \\ p^{i+\frac{2}{3}} = \exp(-\Delta t \Gamma^i) p^{i+\frac{1}{3}} + \sqrt{1 - \exp(-2\Delta t \Gamma^i)} \xi^i \\ q^{i+1} = q^{i+\frac{1}{2}} + p^{i+\frac{2}{3}} \frac{\Delta t}{2} \\ p^{i+1} = p^{i+\frac{2}{3}} - \nabla U(q^{i+1}) \frac{\Delta t}{2} \end{cases} \quad (\text{B.2})$$

for $i \in \mathbb{N}$, $\Delta t > 0$, where ξ^i are independent n -dimensional standard normal random variables and $\Gamma^i \in \mathbb{S}_{++}^n$ are a sequence of friction matrices to be updated throughout the duration of the algorithm, but we mention again recent developments, *e.g.* [16, 19, 29, 54, 64, 66], on discretisations of the underdamped Langevin dynamics; the majority of the numerical error involved in updating Γ is expected to come from the small number of particles in approximating the integrals in the expression (1.9) for $\Delta\Gamma$, so that no further deliberation is made about the choice of discretisation for the purposes here. The first variation process (B.3) together with its initial condition is

$$D_p q_t = \int_0^t D_p p_s \, ds, \quad (\text{B.3a})$$

$$D_p p_t = I_n - \int_0^t (D^2 U(q_s) D_p q_s + \Gamma D_p p_s) \, ds. \quad (\text{B.3b})$$

In order to simulate (B.3), an analogous splitting scheme is used:

$$\begin{cases} Dp^{i+\frac{1}{3}} = Dp^i - D^2 U(q^i) Dq^i \frac{\Delta t}{2} \\ Dq^{i+\frac{1}{2}} = Dq^i + Dp^{i+\frac{1}{3}} \frac{\Delta t}{2} \\ Dp^{i+\frac{2}{3}} = \exp(-\Delta t \Gamma^i) Dp^{i+\frac{1}{3}} \\ Dq^{i+1} = Dq^{i+\frac{1}{2}} + Dp^{i+\frac{2}{3}} \frac{\Delta t}{2} \\ Dp^{i+1} = Dp^{i+\frac{2}{3}} - D^2 U(q^{i+1}) Dq^i \frac{\Delta t}{2}. \end{cases} \quad (\text{B.4})$$

In the case where the second derivatives of U are not directly available, the k th column of (for example) $D^2U(q^i)Dq^i \frac{\Delta t}{2}$ can be approximated by

$$-\nabla U\left(q^i + \frac{\Delta t}{2}(Dq^i)_k\right) + \nabla U(q^i) \quad (\text{B.5})$$

where $(Dq^i)_k$ denotes the k th column of Dq^i , so that (B.3) can still be approximated in the absence of Hessian evaluations. The approximation (B.5) will be used only when explicitly stated in the sequel.

B.1.2. Gradient procedure in Γ

Suppose we have available a series of proposal updates $(b_0, \dots, b_{L-1}) \in \mathbb{R}^{n \times n \times L}$ for Γ , each element of which being noisy estimates of the same gradient direction in Γ . Given stepsizes $\alpha^i = \alpha \in \mathbb{R}$ and an annealing factor $r \in \mathbb{R}$, the following constrained stochastic gradient descent (for i where proposal updates are produced)

$$\Gamma^{i+1} = \Pi_{\text{pd}}^\mu \left(\Gamma^i + \frac{\alpha^i}{2L} \sum_{j=0}^{L-1} (b_j + b_j^\top) \right) \quad (\text{B.6})$$

can be considered, where $L \in \mathbb{N}$ and Π_{pd}^μ is the projection to a positive definite matrix, for some minimum value $\mu > 0$, given by

$$\Pi_{\text{pd}}^\mu(M) = \sum_{i=1}^n \max(\lambda_i, \mu) v_i v_i^\top \quad (\text{B.7})$$

for symmetric $M \in \mathbb{R}^{n \times n}$ and its the eigenvalue decomposition

$$M = \sum_{i=1}^n \lambda_i v_i v_i^\top.$$

Alternatively, a Heavy-ball method [32, 61] (with projection) can be used. The method is considered in the stochastic gradient context in [15], given here as

$$\begin{cases} \Gamma^{i+1} = \Pi_{\text{pd}}^\mu(\Gamma^i + \alpha^i \Theta^{i+1}), \\ \Theta^{i+1} = (1 - \alpha^i r) \Theta^i + \frac{\alpha^i}{2L} \sum_{j=0}^{L-1} (b_j + b_j^\top). \end{cases} \quad (\text{B.8})$$

The heavy-ball method offers a smoother trajectory of Γ over the course of the algorithm. Under appropriate assumptions on b_j , in particular if

$$\frac{1}{2L} \sum_{j=0}^{L-1} (b_j + b_j^\top) \sim \mathcal{N}(\nabla \sigma^2(\Gamma^i), \sigma_b^2 I_{n^2}),$$

for some gradient $\nabla \sigma^2(\Gamma_k^i)$ and variance $\sigma_b^2 > 0$, then the system (B.8) has the interpretation of an Euler discretisation of a constrained Langevin dynamics, in which case $\frac{r}{\sqrt{\alpha^i \sigma_b^2}}$ is the inverse temperature. By increasing r , the analogous invariant distribution “sharpens” around the maximum in its density and in this way reduces the effect of noise at equilibrium; on the other hand, decreasing r reduces the decay in the momentum.

B.1.3. A thinning approach for $\Delta\Gamma$

The most straightforward way of approximating the integral in (B.1) is to use independent realisations of (B.2), but we draw alternatively a thinned sample [58] from a single trajectory here in order to run only a single parallel set of realisations of (B.2) and (B.4) at a time. More specifically, we consider a single realisation of (B.2) and regularly-spaced points from its trajectory (possibly after a burn-in) as sample points from $\tilde{\pi}$.

Starting at each of these sample points and ending at each subsequent one, the process is replicated albeit starting with a momentum reversal and simulated in parallel. In addition, for each of the two processes, a corresponding first variation process (B.4) is calculated in parallel. A precise description follows.

Let $K = 1$ for simplicity. The Γ direction (B.1) is approximated by

$$-\frac{1}{(L+L^*)} \sum_{l=0}^{L+L^*-1} \left(\sum_{i=1}^T \frac{\Delta t}{K} \sum_{k=1}^K \nabla f(q_{(k)}^{i+Tl+B})^\top Dq_{(k)}^{i+Tl+B} \right) \otimes \left(\sum_{i=1}^T \frac{\Delta t}{K} \sum_{k=1}^K \nabla f(\tilde{q}_{(k)}^{i+Tl+B})^\top D\tilde{q}_{(k)}^{i+Tl+B} \right), \quad (\text{B.9})$$

where $L \in \mathbb{N}$, $((q_{(k)}^i, p_{(k)}^i))_{i \in \mathbb{N}}$, $((\tilde{q}_{(k)}^i, \tilde{p}_{(k)}^i))_{i \in \mathbb{N}}$ denote solutions to (B.2)

- for $i \neq B + Tl - 1$, $l \in \mathbb{N}_0$ if $k \neq 1$ and;
- for all i if $k = 1$

with initial condition $(0, 0)$, noise $\xi^i = \xi_{(k)}^i, \tilde{\xi}_{(k)}^i$ for all $i \in \mathbb{N}$ satisfying $\xi_{(k)}^i = \xi_{(k')}^i = \tilde{\xi}_{(k)}^i = \tilde{\xi}_{(k')}^i$ for all $i < B, 1 \leq k \leq K, 1 \leq k' \leq K$, independent otherwise as i and k vary, along with corresponding $(Dq_{(k)}^i, Dp_{(k)}^i), (D\tilde{q}_{(k)}^i, D\tilde{p}_{(k)}^i)$ satisfying (B.4) for $i \neq B + Tl - 1, l \in \mathbb{N}_0$ (regardless of k), and where the $k \neq 1$ processes are “reset” at $i = B + Tl$ corresponding to the values of the $k = 1$ chain if the first variation processes have converged to zero, that is,

$$q_{(k)}^{Tl+B} = q_{(1)}^{Kl+B}, \quad p_{(k)}^{Tl+B} = p_{(1)}^{Tl+B}, \quad Dq_{(k)}^{Tl+B} = 0, \quad Dp_{(k)}^{Tl+B} = I_n \quad (\text{B.10a})$$

$$\tilde{q}_{(k)}^{Tl+B} = q_{(1)}^{Tl+B}, \quad \tilde{p}_{(k)}^{Tl+B} = -p_{(1)}^{Tl+B}, \quad D\tilde{q}_{(k)}^{Tl+B} = 0, \quad D\tilde{p}_{(k)}^{Tl+B} = I_n \quad (\text{B.10b})$$

for all $1 \leq k \leq K$ if for some $D_{\text{conv}} > 0$,

$$\max_{i,j,k} \left| \left(Dq_{(k)}^{Tl+B} \right)_{ij} \right| < D_{\text{conv}}, \quad \max_{i,j,k} \left| \left(D\tilde{q}_{(k)}^{Tl+B} \right)_{ij} \right| < D_{\text{conv}}, \quad (\text{B.11a})$$

$$\max_{i,j,k} \left| \left(Dp_{(k)}^{Tl+B} \right)_{ij} \right| < D_{\text{conv}}, \quad \max_{i,j,k} \left| \left(D\tilde{p}_{(k)}^{Tl+B} \right)_{ij} \right| < D_{\text{conv}} \quad (\text{B.11b})$$

and $L^* \in \mathbb{N}$ is such that the number of elements in $\{l \in \mathbb{N} : 1 \leq l \leq L + L^*\}$ satisfying (B.11) is L . The approach is summarised in Algorithm 2. Of course, the above for generic $K \in \mathbb{N}$ constitutes improving approximations to $\Delta\Gamma$. Note that as Γ changes through the prescribed procedure, the asymptotic variance associated to the given observable f is expected to improve, but on the contrary, the estimator (B.9) for the continuous-time expression (B.1) may well worsen, since the integrand (of the outermost integral) in (B.1) is not f . Increasing L is expected to solve any resulting issues; on the other hand extremely small L have been successful in the experiments here.

Remark B.1. If it is of interest to approximate expectations of $P \in \mathbb{N}$ observables with respect to π , the quantity $\sum_i^P \sigma_i^2$ for example can be used as an objective function, where σ_i^2 is the asymptotic variance from the i th observable. In the implementation in Algorithm 2, instead of only the vectors $\zeta, \tilde{\zeta}$, this amounts to calculating at each iteration the vectors $\zeta^{(i)}, \tilde{\zeta}^{(i)}$ corresponding to the i th observable and taking the sum of the resulting update matrices in Γ to update Γ . This calls for no extra evaluations of ∇U over the single observable case.

Remark B.2 (Tangent processes along random directions). We mention the situation where simulating the full first variation processes $(D_p q_t, D_p p_t)$ in $\mathbb{R}^{n \times 2n}$ is prohibitively expensive. A directional tangent process can be used instead of $(D_p q_t, D_p p_t)$. Consider for a unit vector $v \in \mathbb{R}^n$, that is $|v| = 1$, randomly chosen at the beginning of each estimation of $\Delta\Gamma$, the pair of vectors $(D_p q_t v, D_p p_t v) \in \mathbb{R}^{n \times 2}$. Multiplying on the right of

Acknowledgements. The authors would like to thank Gabriel Stoltz for insightful comments on an earlier draft of the paper. M.C. was funded under a EPSRC studentship. G.A.P. was partially supported by the EPSRC through grant EP/P031587/1. N.K. and G.A.P. were funded in part by JPMorgan Chase & Co under a J.P. Morgan A.I. Research Awards 2021. Part of this project was carried out as T.L. was a visiting professor at Imperial College of London, with a visiting professorship grant from the Leverhulme Trust. The Department of Mathematics at ICL and the Leverhulme Trust are warmly thanked for their support.

REFERENCES

- [1] C.D. Aliprantis and K.C. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, 3rd edition. Springer, Berlin (2006).
- [2] H. AlRachid, L. Mones and C. Ortner, Some remarks on preconditioning molecular dynamics. *SMAI J. Comput. Math.* **4** (2018) 57–80.
- [3] C. Andrieu and J. Thoms, A tutorial on adaptive MCMC. *Stat. Comput.* **18** (2008) 343–373.
- [4] L. Angeli, D. Crisan and M. Ottobre, Uniform in time convergence of numerical schemes for stochastic differential equations via Strong Exponential stability: Euler methods. Split-Step and Tamed Schemes. Preprint [arXiv:2303.15463](https://arxiv.org/abs/2303.15463) (2023).
- [5] J. Baker, P. Fearnhead, E.B. Fox and C. Nemeth, Control variates for stochastic gradient MCMC. *Stat. Comput.* **29** (2019) 599–615.
- [6] D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov and S. Samsonov, Variance reduction for Markov chains with application to MCMC. *Stat. Comput.* **30** (2020) 973–997.
- [7] A. Beskos and A. Stuart, MCMC methods for sampling function space, in *ICIAM 07 – 6th International Congress on Industrial and Applied Mathematics*. Eur. Math. Soc., Zürich (2009) 337–364.
- [8] A. Beskos, G. Roberts, A. Stuart and J. Voss, MCMC methods for diffusion bridges. *Stoch. Dyn.* **8** (2008) 319–350.
- [9] R.N. Bhattacharya, On the functional central limit theorem and the law of the iterated logarithm for Markov processes. *Z. Wahrsch. Verw. Gebiete* **60** (1982) 185–201.
- [10] F. Bolley, A. Guillin and F. Malrieu, Trend to equilibrium and particle approximation for a weakly selfconsistent Vlasov–Fokker–Planck equation. *M2AN Math. Model. Numer. Anal.* **44** (2010) 867–884.
- [11] G. Bussi and M. Parrinello, Accurate sampling using Langevin dynamics. *Phys. Rev. E* **75** (2007) 056707.
- [12] P. Cattiaux, D. Chafaï and A. Guillin, Central limit theorems for additive functionals of ergodic Markov diffusions processes. *ALEA Lat. Am. J. Probab. Math. Stat.* **9** (2012) 337–382.
- [13] M. Chak, N. Kantas and G.A. Pavliotis, On the generalised Langevin equation for simulated annealing. *SIAM/ASA J. Uncertainty Quantif.* **11** (2023) 139–167.
- [14] N.S. Chatterji, N. Flammarion, Y.-A. Ma, P.L. Bartlett and M.I. Jordan, On the theory of variance reduction for stochastic gradient Monte Carlo. *PMLR* **80** (2018) 764–773.
- [15] X. Chen, S. Liu, R. Sun and M. Hong, On the convergence of a class of adam-type algorithms for non-convex optimization, in 2019. 7th International Conference on Learning Representations, ICLR 2019. Conference date: 06-05-2019 Through 09-05-2019 (2019).
- [16] X. Cheng, N.S. Chatterji, P.L. Bartlett and M.I. Jordan, Underdamped Langevin MCMC: a non-asymptotic analysis, in *Proceedings of the 31st Conference On Learning Theory*. Vol. 75 of *Proceedings of Machine Learning Research*, edited by S. Bubeck, V. Perchet and P. Rigollet. 06–09 Jul 2018. *PMLR* (2018) 300–323.
- [17] D. Crisan, P. Dobson and M. Ottobre, Uniform in time estimates for the weak error of the Euler method for SDEs and a pathwise approach to derivative estimates for diffusion semigroups. *Trans. Am. Math. Soc.* **374** (2021) 3289–3330.
- [18] D. Crisan, P. Dobson, B. Goddard, M. Ottobre and I. Souttar, Poisson equations with locally-Lipschitz coefficients and uniform in time averaging for stochastic differential equations via strong exponential stability. Preprint [arXiv:2204.02679](https://arxiv.org/abs/2204.02679) (2022).
- [19] A.S. Dalalyan and L. Riou-Durand, On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli* **26** (2020) 1956–1988.
- [20] B. Delyon and Y. Hu, Simulation of conditioned diffusion and application to parameter estimation. *Stochastic Process. Appl.* **116** (2006) 1660–1675.
- [21] Z. Ding, Q. Li, J. Lu and S.J. Wright, Random coordinate underdamped Langevin Monte Carlo. Preprint [arXiv:2010.11366](https://arxiv.org/abs/2010.11366) (2020).
- [22] Z. Dong and X. Peng, Malliavin matrix of degenerate SDE and gradient estimate. *Electron. J. Probab.* **19** (2014) 26.
- [23] A.B. Duncan, T. Lelièvre and G.A. Pavliotis, Variance reduction using nonreversible Langevin samplers. *J. Stat. Phys.* **163** (2016) 457–491.
- [24] A.B. Duncan, N. Nüsken and G.A. Pavliotis, Using perturbed underdamped Langevin dynamics to efficiently sample from probability distributions. *J. Stat. Phys.* **169** (2017) 1098–1131.
- [25] A. Durmus and E. Moulines, High-dimensional Bayesian inference via the unadjusted Langevin algorithm. Preprint [arXiv:1605.01559](https://arxiv.org/abs/1605.01559) (2018).
- [26] 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](https://arxiv.org/abs/2107.14542) (2021).
- [27] J.-P. Eckmann and M. Hairer, Non-equilibrium statistical mechanics of strongly anharmonic chains of oscillators. *Comm. Math. Phys.* **212** (2000) 105–164.

- [28] S.N. Ethier and T.G. Kurtz, Markov Processes: Characterization and Convergence. *Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. John Wiley & Sons, Inc., New York (1986).
- [29] J. Foster, T. Lyons and H. Oberhauser, The shifted ODE method for underdamped Langevin MCMC. Preprint [arXiv:2101.03446](https://arxiv.org/abs/2101.03446) (2021).
- [30] E. Fournié, J.-M. Lasry, J. Lebuchoux, P.-L. Lions and N. Touzi, Applications of Malliavin calculus to Monte Carlo methods in finance. *Finan. Stoch.* **3** (1999) 391–412.
- [31] A. Friedman, Stochastic differential equations and applications. Vol. 1, in Probability and Mathematical Statistics, Vol. 28. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London (1975).
- [32] E. Ghadimi, H.R. Feyzmahdavian and M. Johansson, Global convergence of the heavy-ball method for convex optimization, in 2015 European Control Conference (ECC). (2015) 310–315.
- [33] A. Guillin and P. Monmarché, Optimal linear drift for the speed of convergence of an hypoelliptic diffusion. *Electron. Commun. Probab.* **21** (2016) 14.
- [34] A. Guillin and F.-Y. Wang, Degenerate Fokker–Planck equations: Bismut formula, gradient estimate and Harnack inequality. *J. Differ. Equ.* **253** (2012) 20–40.
- [35] M. Hairer, A.M. Stuart, J. Voss and P. Wiberg, Analysis of SPDEs arising in path sampling. I. The Gaussian case. *Commun. Math. Sci.* **3** (2005) 587–603.
- [36] M. Hairer, A.M. Stuart and J. Voss, Analysis of SPDEs arising in path sampling. II. The nonlinear case. *Ann. Appl. Probab.* **17** (2007) 1657–1706.
- [37] M. Hairer, A.M. Stuart and J. Voss, Sampling conditioned hypoelliptic diffusions. *Ann. Appl. Probab.* **21** (2011) 669–698.
- [38] Y. He, K. Balasubramanian and M.A. Erdogdu, On the ergodicity, bias and asymptotic normality of randomized midpoint sampling method. *Adv. Neural Inf. Process. Syst.* **33** (2020) 7366–7376.
- [39] B. Helffer and F. Nier, Hypoelliptic Estimates and Spectral Theory for Fokker–Planck Operators and Witten Laplacians. Vol. 1862 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin (2005).
- [40] L. Hörmander, Hypoelliptic second order differential equations. *Acta Math.* **119** (1967) 147–171.
- [41] A.M. Horowitz, The second order Langevin equation and numerical simulations. *Nucl. Phys. B* **280** (1987) 510–522.
- [42] A.M. Horowitz, A generalized guided Monte Carlo algorithm. *Phys. Lett. B* **268** (1991) 247–252.
- [43] S. Hottovy, A. McDaniel, G. Volpe and J. Wehr, The Smoluchowski–Kramers limit of stochastic differential equations with arbitrary state-dependent friction. *Comm. Math. Phys.* **336** (2015) 1259–1283.
- [44] A. Kavalur, V. Guduguntla and W.K. Kim, Effects of Langevin friction and time steps in the molecular dynamics simulation of nanoindentation. *Mol. Simul.* **46** (2020) 911–922.
- [45] R. Khasminskii, Stochastic Stability of Differential Equations. Vol. 66 of *Stochastic Modelling and Applied Probability*, 2nd edition. Springer, Heidelberg (2012). With contributions by G.N. Milstein and M.B. Nevelson.
- [46] T. Komorowski, C. Landim and S. Olla, Fluctuations in Markov Processes: Time Symmetry and Martingale Approximation. Vol. 345 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Heidelberg (2012).
- [47] N.V. Krylov, On Kolmogorov’s equations for finite-dimensional diffusions, in Stochastic PDE’s and Kolmogorov Equations in Infinite Dimensions (Cetraro, 1998). Vol. 1715 of *Lecture Notes in Math.* Springer, Berlin (1999) 1–63.
- [48] B. Leimkuhler and C. Matthews, Rational construction of stochastic numerical methods for molecular sampling. *Appl. Math. Res. Express. AMRX* **2013** (2013) 34–56.
- [49] B. Leimkuhler and C. Matthews, Robust and efficient configurational molecular sampling via Langevin dynamics. *J. Chem. Phys.* **138** (2013) 174102.
- [50] 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.
- [51] T. Lelièvre and G. Stoltz, Partial differential equations and stochastic methods in molecular dynamics. *Acta Numer.* **25** (2016) 681–880.
- [52] T. Lelièvre, M. Rousset and G. Stoltz, Free Energy Computations: A Mathematical Perspective. Imperial College Press, London (2010).
- [53] T. Lelièvre, F. Nier and G.A. Pavliotis, Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *J. Stat. Phys.* **152** (2013) 237–274.
- [54] P. Monmarché, High-dimensional MCMC with a standard splitting scheme for the underdamped Langevin diffusion. *Electron. J. Stat.* **15** (2021) 4117–4166.
- [55] P. Monmarché, Almost sure contraction for diffusions on \mathbb{R}^d . Application to generalized Langevin diffusions. *Stochastic Process. Appl.* **161** (2023) 316–349.
- [56] W. Mou, Y.-A. Ma, M.J. Wainwright, P.L. Bartlett and M.I. Jordan, High-order Langevin diffusion yields an accelerated MCMC algorithm. *J. Mach. Learn. Res.* **22** (2021) 41.
- [57] C. Nemeth and P. Fearnhead, Stochastic gradient Markov Chain Monte Carlo. *J. Am. Stat. Assoc.* **116** (2021) 433–450.
- [58] A.B. Owen, Statistically efficient thinning of a Markov chain sampler. *J. Comput. Graph. Stat.* **26** (2017) 738–744.
- [59] G.A. Pavliotis, Asymptotic analysis of the Green–Kubo formula. *IMA J. Appl. Math.* **75** (6) 951–967.
- [60] G.A. Pavliotis, Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations. Vol. 60 of *Texts in Applied Mathematics*. Springer, New York (2014).
- [61] B. Polyak, Some methods of speeding up the convergence of iteration methods. *USSR Comput. Math. Math. Phys.* **4** (1964) 1–17.

- [62] P.E. Protter, Stochastic Integration and Differential Equations: Stochastic Modelling and Applied Probability. Vol. 21 of *Applications of Mathematics (New York)*, 2nd edition. Springer-Verlag, Berlin (2004).
- [63] M. Sachs, B. Leimkuhler and V. Danos, Langevin dynamics with variable coefficients and nonconservative forces: from stationary states to numerical methods. *Entropy* **19** (2017) 647.
- [64] 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) 37.
- [65] A. Scemama, T. Lelièvre, G. Stoltz, E. Cancès and M. Caffarel, An efficient sampling algorithm for variational Monte Carlo. *J. Chem. Phys.* **125** (2006) 114105.
- [66] R. Shen and Y.T. Lee, The randomized midpoint method for log-concave sampling, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett. Vol. 32. Curran Associates, Inc. (2019).
- [67] R.D. Skeel and C. Hartmann, Choice of damping coefficient in Langevin dynamics. *Eur. Phys. J. B* **94** (2021) 1–13.
- [68] L.F. South, C.J. Oates, A. Mira and C. Drovandi, Regularised zero-variance control variates for high-dimensional variance reduction. *Bayesian Anal.* **18** (2023) 865–888.
- [69] J. Teichmann, Calculating the Greeks by cubature formulae. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **462** (2006) 647–670.
- [70] D. Zou and Q. Gu, On the convergence of Hamiltonian Monte Carlo with stochastic gradients, in *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139 of *Proceedings of Machine Learning Research*, edited by M. Meila and T. Zhang. PMLR (2021) 13012–13022.



Please help to maintain this journal in open access!

This journal is currently published in open access under the Subscribe to Open model (S2O). We are thankful to our subscribers and supporters for making it possible to publish this journal in open access in the current year, free of charge for authors and readers.

Check with your library that it subscribes to the journal, or consider making a personal donation to the S2O programme by contacting subscribers@edpsciences.org.

More information, including a list of supporters and financial transparency reports, is available at <https://edpsciences.org/en/subscribe-to-open-s2o>.