HAL
open science

Optimal friction matrix for underdamped Langevin sampling

Martin Chak, Nikolas Kantas, Tony Lelièvre, Grigorios A Pavliotis

To cite this version:

Martin Chak, Nikolas Kantas, Tony Lelièvre, Grigorios A Pavliotis. Optimal friction matrix for underdamped Langevin sampling. 2021. hal-03455809

HAL Id: hal-03455809

https://hal.science/hal-03455809

Preprint submitted on 29 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Optimal friction matrix for underdamped Langevin sampling

Martin Chak Nikolas Kantas Tony Lelièvre Grigorios A. Pavliotis

November 29, 2021

Abstract

A systematic procedure for optimising the friction coefficient in underdamped Langevin dynamics as a sampling tool is given by taking the gradient of the associated asymptotic variance with respect to friction. We give an expression for this gradient in terms of the solution to an appropriate Poisson equation and show that it can be approximated by short simulations of the associated first variation/tangent process under concavity assumptions on the log density. Our algorithm is applied to the estimation of posterior means in Bayesian inference problems and reduced variance is demonstrated when compared to the original underdamped and overdamped Langevin dynamics in both full and stochastic gradient cases.

Contents

1 Introduction **2**
1.1 Outline of approach . . . 3

2 Setting **5**
2.1 Preliminaries and notation . . . 5
2.2 Semigroup bound, Poisson equation and central limit theorem . . . 5

3 Directional derivative of σ^2 **7**
3.1 Preliminary results and the main formula . . . 7
3.2 A formula using a tangent process . . . 8

4 Quadratic cases **10**
4.1 Quadratic observable . . . 11
4.2 Odd polynomial observable . . . 15
4.3 Quartic observable . . . 16

5 Computation of the change in Γ **16**
5.1 Methodology . . . 16
5.1.1 Splitting . . . 17
5.1.2 Gradient procedure in Γ . . . 18
5.1.3 A thinning approach for $\Delta\Gamma$ . . . 18
5.2 Concrete examples . . . 20
5.2.1 One dimensional quadratic case . . . 20
5.2.2 Diffusion bridge sampling . . . 21
5.2.3 Bayesian inference . . . 23

6 Proofs **25**

7 Discussion **34**
7.1 Relation to previous methodologies . . . 34
7.2 The nonconvex case . . . 35
7.3 Position-dependent friction . . . 35
7.4 Metropolisation . . . 35
7.5 Conclusion . . . 35

A Preliminaries **40**

B Solving the Poisson equation with a Galerkin method **41**

C Approximation of $\Delta\Gamma$ using independent realisations **44**

1 Introduction

Let π be a probability measure on $\mathbb{R}^n$ with smooth positive bounded density, also denoted π, with respect to the Lebesgue measure on $\mathbb{R}^n$ and let $f \in L^2(\pi)$ be an observable. In a range of applications including molecular dynamics [12, 52, 54] and machine learning [60, 82, 83], a quantity of interest is the expectation of f with respect to π,

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

which is analytically intractable and is numerically approximated most commonly by Markov Chain Monte Carlo (MCMC) methods, whereby π is sampled by simulating an ergodic Markov chain $(X_k)_{1\leq k\leq N}$ with π as its unique invariant measure and $\pi(f)$ is approximated by the empirical average $\frac{1}{N}\sum_{k=1}^N f(X_k)$. MCMC methods enjoy central limit theorems for many Markov chains employed, the most well-known (class) of such methods being the Metropolis-Hastings algorithm [41, 56]. Recent efforts have been to develop MCMC methods suited to settings where $n \gg 1$ and where point evaluations of π or its gradients are computationally expensive; these methods include slice sampling [25, 61], Hamiltonian Monte Carlo [8, 24, 62], piecewise-deterministic Markov processes [10, 13, 81] and those based on discretisations of continuous-time stochastic dynamics [30, 54, 55] together with divide-and-conquer and subsampling approaches [4].

In this paper we consider the underdamped Langevin dynamics. Denoting $\mathbb{S}^n_{++}$ as the set of real symmetric $n \times n$ positive definite matrices, the underdamped Langevin dynamics[1] 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

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

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

where $\sqrt{\Gamma} \in \mathbb{R}^{n\times n}$ is any matrix satisfying

$$\sqrt{\Gamma}\sqrt{\Gamma}^\top = \Gamma,$$

$U : \mathbb{R}^n \to \mathbb{R}$ is the associated smooth potential or negative log density such that $\pi \propto e^{-U}$ and W_t denotes a standard Wiener process on $\mathbb{R}^n$. The probability distribution from underdamped Langevin dynamics converges under general assumptions to the invariant probability measure given by

$$\tilde{\pi}(dq, dp) = Z^{-1}e^{-U(q)-\frac{p^\top M^{-1}p}{2}}dqdp \tag{1.2}$$

for a normalising constant Z and there have been numerous recent works [19, 20, 28, 32, 42, 50, 58, 73] on its discretisations in terms of the quality of convergence to $\tilde{\pi}$ over time measured by (e.g.) Wasserstein distance; in this paper, the goal is to optimise $\Gamma \in \mathbb{S}^n_{++}$ directly with respect to the asymptotic variance in the convergence of

$$\pi_T(f) := \frac{1}{T}\int_0^T f(q_t)dt$$

[1] also referred to as Langevin, second-order Langevin or kinetic Langevin.

to $\pi(f)$ for any particular f (or a finite set of observables) as $T \to \infty$.
We mention that parameter tuning in MCMC methods is a widely considered topic [2, 84] (and references within). Specifically for underdamped Langevin dynamics, tuning the momentum part of $\tilde{\pi}$ with respect to reducing metastability or computational effort was considered in [70, 78, 80]. The choice of friction (as a scalar) has been a subject of consideration as early as in [44], then in [1, 14, 47, 76] within the context of molecular dynamics and also in [20, 27]. Most of these works make use of different measures for efficiency. The present work constitutes the first systematic gradient procedure for choosing the friction matrix in an optimal manner, with respect to a appropriate cost criterion.

1.1 Outline of approach

We proceed with a formal description of our approach, precise statements can be found in the main Theorems 3.2 and 3.5. It is known using results from [72] and [9] that, under suitable assumptions on U and f, a central limit theorem

$$\frac{1}{\sqrt{T}} \int_0^T (f(q_t) - \pi(f))dt \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2) \qquad \text{as } T \to \infty \tag{1.3}$$

holds and that σ^2, the asymptotic variance, has the form

$$\sigma^2 = 2 \int \phi(f - \pi(f))d\tilde{\pi} \tag{1.4}$$

where ϕ is a solution to the Poisson equation

$$-L\phi = f - \pi(f) \tag{1.5}$$

and L denotes the infinitesimal generator associated to (1.1). Two key observations are then made. Firstly, for any direction $\delta\Gamma \in \mathbb{R}^{n\times n}$ in the friction matrix, the derivative of σ^2 with respect to the entries of Γ in the direction $\delta\Gamma$, denoted $d\sigma^2.\delta\Gamma$, is given by the formula

$$d\sigma^2.\delta\Gamma = -2 \int (\nabla_p \phi)^\top \delta\Gamma \nabla_p \tilde{\phi} d\tilde{\pi}, \tag{1.6}$$

where $\tilde{\phi}$ is given by

$$\tilde{\phi}(q, p) = \phi(q, -p). \tag{1.7}$$

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

$$\Delta\Gamma := \int \nabla_p \phi \otimes \nabla_p \tilde{\phi} d\tilde{\pi} \tag{1.8}$$

where $\otimes$ denotes the outer product. Similarly, taking $\delta\Gamma$ to be the diagonal elements of (1.8) or $\delta\Gamma = I_n \int \nabla_p \phi \cdot \nabla_p \tilde{\phi} d\tilde{\pi}$ give in both cases a negative change in asymptotic variance respectively for diagonal Γ and Γ of the form cI_n. The second observation is that since the solution ϕ to the Poisson equation (1.5) is known to be given by

$$\phi(q, p) = \int_0^\infty \mathbb{E}[f(q_t)]dt, \tag{1.9}$$

where (q_t, p_t) solves (1.1) with initial condition $(q_0, p_0) = (q, p)$, given convexity conditions on the potential U and under suitable assumptions, we have

$$\nabla_p \phi = \int_0^\infty \mathbb{E}[\nabla f(q_t)^\top D_p q_t]dt, \tag{1.10}$$

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 condition p in momentum. Not only does $D_p q_t$ satisfy the dynamics that result from taking partial derivatives in (1.1), which are susceptible to algorithmic simulation, but the process also decays to zero exponentially quickly, so that the infinite time integral (1.10) can be accurately approximated with a truncation using short simulations of $D_p q_t$ for

adaptive estimations of the direction (1.8) in Γ. This leads to an adaptive algorithm involving the selection of Γ in an appropriate constrained set, of which we illustrate the performance with numerical examples.

Examples where improved Γ can be found analytically are presented in Section 4. Numerical illustrations making use of (1.6) and (1.10) are presented in Sections 5.2. In particular, the algorithm is applied on the problem of finding the posterior mean in a Bayesian logistic regression inference problem for two datasets with hundreds of dimensions, where the best friction matrices found in both cases are close to zero (for example $\Gamma = 0.1 I_n$ performs well compared to $\Gamma = I_n$, demonstrating reduced variance of almost an order of magnitude in Tables 5.2 and 5.3). To use the asymptotic variance for a particular observable (or a set of them) and to use measures for the quality of convergence to $\tilde{\pi}$ or to minimise an autocorrelation time as considered in [1, 14, 20, 44, 47, 76] can be conflicting goals. To elaborate, in [44], the autocorrelation time was used as the point of comparison in the Gaussian target measure case for the optimal friction. For $n = 1$, $\omega, \gamma \in \mathbb{R}$, $U(q) = \frac{1}{2}\omega^2 q^2$, $M = 1$, $\Gamma = \gamma > 0$, the autocorrelation time for (1.1) satisfies

$$\partial_t \begin{pmatrix} \mathbb{E}(q_t q_0) \\ \mathbb{E}(p_t q_0) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\omega^2 & -\gamma \end{pmatrix} \begin{pmatrix} \mathbb{E}(q_t q_0) \\ \mathbb{E}(p_t q_0) \end{pmatrix}. \tag{1.11}$$

By considering the eigenvalues, the conclusion in [45] is that the optimal γ for minimising the magnitude of $\mathbb{E}(q_t q_0)$ is given by the critical damping $\gamma = 2\omega$. A similar conclusion can be made when considering the spectral gap[66]. On the other hand, if $f(q) = q$ in our setting, formally, the quantity $\int \int_0^\infty \mathbb{E}(q_t q_0) dt d\pi(q_0)$ is the asymptotic variance

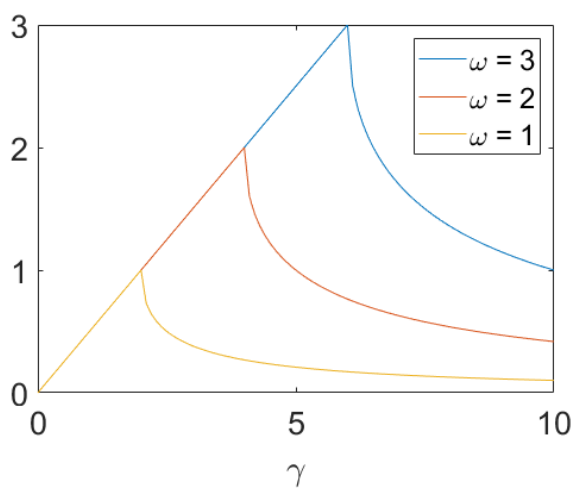

Figure 1.1: The values $\min_i(|\mathrm{Re}(\lambda_i)|)$, where λ_i are the eigenvalues of the matrix appearing in (1.11), also the spectral gap. Critical values of γ are given by 2ω.

due to (1.4) and (1.9). Despite the similarity, Corollary 4.8 asserts that $\gamma = 0$ is optimal. A more detailed discussion about Corollary 4.8 is given in Section 4.2. This difference emphasizes that, at the cost of generic convergence to $\tilde{\pi}$, the tuning of Γ here is directed at variance reduction for a particular observable, in this case $f(q) = q$. However, multiple asymptotic variances can be used for the objective function to minimise, so that Γ can be optimised with respect to several observables of interest simultaneously. Remark 5.1 describes the implementation for a linear combination of asymptotic variances at no extra cost in terms of evaluations of π or its gradients.

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 (1.6) and (1.10) are precisely stated. Exact results concerning improvements in Γ including the quadratic U, quadratic f and linear f cases are given in Section 4. Numerical methods in approximating (1.8) together with an algorithm resulting from (1.6) and (1.10) is outlined and detailed in Algorithm 1 and 2 respectively in Section 5, 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 about future work.

2 Setting

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, $(\mathcal{F}_t)_{t\in\mathbb{R}}$ 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 $(\mathcal{F}_t)_{t\in\mathbb{R}}$, $\tilde{\pi}$ be a probability measure given by (1.2).

Assumption 1. $U \in C^\infty(\mathbb{R}^n)$ satisfies $U \geq 0$ and its second derivatives satisfy

$$\|D^2U\|_\infty := \sum_{i,j} \sup_{q\in\mathbb{R}^{2n}} \left|\partial_{q_i}\partial_{q_j}U(q)\right| < \infty. \tag{2.1}$$

Note that (2.1) implies

$$|\nabla U(q)| \leq K_U(1+|q|) \tag{2.2}$$

for some $K_U > 0$. The existence and uniqueness of a strong solution to (1.1) is established in Theorem A.1. Due to the smoothness of U and Γ, the coefficients in (1.1) are locally Lipschitz and well-posedness of equation (1.1) is given by [68], to which we also refer to for the sense of solution. In addition, we make certain to satisfy the joint measurability assumption in [9] of (A.2).

2.1 Preliminaries and notation

The set of smooth compactly supported functions is denoted C_c^∞. The infinitesimal generator L (defined in (A.6)) associated to (1.1) is given formally by its differential operator form, denoted $\mathcal{L}$, when acting on the subset $C_c^\infty(\mathbb{R}^{2n})$,

$$\mathcal{L} = p^\top M^{-1}\nabla_q - \nabla U(q)^\top\nabla_p - p^\top M^{-1}\Gamma\nabla_p + \nabla_p^\top\Gamma\nabla_p. \tag{2.3}$$

Its formal L^2-adjoint $L^\top$ satisfies

$$L^\top\tilde{\pi} = 0, \tag{2.4}$$

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

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

and similar for $\tilde{\pi}$. 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 matrices A, $|A|$ denotes the operator norm associated with the Euclidean norm on $\mathbb{R}^n$. e_i is used to denote the ith Euclidean basis vector. For $A, B \in \mathbb{R}^{n\times n}$, $A : B := \sum_{i,j} A_{ij}B_{ij}$ and $A_S = \frac{1}{2}(A+A^\top)$. $\langle\cdot,\cdot\rangle_{\tilde{\pi}}$ denotes the inner product in $L^2(\tilde{\pi})$ and similar for π.

2.2 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 to optimise Γ with respect to. Specifically, it will be shown that under some weighted L^∞ bound on the observable $f \in L^2(\pi)$, the estimator π_T for the unique solution (q_t, p_t) to (1.1) converges to $\pi(f)$ as $T \to \infty$ such that (1.3) holds with (1.4).

It is well known that the asymptotic variance can be expressed in terms of the solution to the Poisson equation (1.5) using the Kipnis-Varadhan framework, see for example Chapter 2 in [48], Section 3.1.3 in [54], [15] and references therein. In order to show that the expression (1.9) is indeed a solution to the Poisson equation (1.5), exponential decay of the semigroup (A.5) is used. In Theorem 2.1 below, we establish convergence in law to the invariant measure for the Langevin dynamics (1.1).For this, let the Lyapunov function $\mathcal{K}_l : \mathbb{R}^{2n} \to \mathbb{R}$ for all $l \in \mathbb{N}$ be given by

$$\mathcal{K}(z) = \mathcal{K}_l(q,p) = \left(cU(q) + a|q|^2 + b\langle q,p\rangle + \frac{c}{2}|p|^2 + 1\right)^l \tag{2.5}$$

for constants $a, b, c > 0$.

Assumption 2. There exist constants $\beta_1, \beta_2 > 0$ and $\alpha \in \mathbb{R}$ such that

$$\forall q \in \mathbb{R}^n, \langle q, \nabla_q U(q) \rangle \geq \beta_1 U(q) + \beta_2 |q|^2 + \alpha. \tag{2.6}$$

Inequality (2.6) implies

$$U(q) \geq C|q|^2 - C \tag{2.7}$$

for all $q \in \mathbb{R}^n$ and some generic constant $C > 0$.

Theorem 2.1. *Under Assumptions 1 and 2, $\tilde{\pi}$ 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*

$$|\mathbb{E}[\varphi(z_t^z)] - \tilde{\pi}(\varphi)| \leq C_l e^{-t\kappa_l} \mathcal{K}_l(z) \left\| \frac{\varphi - \tilde{\pi}(\varphi)}{\mathcal{K}_l} \right\|_{L^\infty} \tag{2.8}$$

for Lebesgue almost all initial $z \in \mathbb{R}^{2n}$, $\mathcal{K}_l \geq 1$ given by (2.5) *and all Lebesgue measurable φ satisfying*

$$\frac{\varphi}{\mathcal{K}_l} \in L^\infty \tag{2.9}$$

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

$$\int \mathcal{K}_l d\tilde{\pi} < \infty \tag{2.10}$$

and

$$\mathcal{L}K_l \leq -a_l \mathcal{K}_l + b_l \tag{2.11}$$

for some constants $a_l, b_l > 0$.

The proof is from [72], in which 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 equation (6.2) in [66]) does not hold, but of course, it applies in particular to our setting.

Remark 2.1. Inequality (2.8) 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 (2.8) together with the strong Feller property given by Theorem 4.2 in [23].

Proof. The measure $\tilde{\pi}$ is invariant due to (2.4). For the rest of the statements, see Theorem 3 in [72]. □

The following corollary holds by taking φ as indicator functions and Remark 2.1.

Corollary 2.2. *Under Assumptions 1 and 2, for all initial $z \in \mathbb{R}^{2n}$, the transition probability p_t^z of* (1.1)*, given by $p_t^z(A) = \mathbb{P}(z_t^z \in A)$, satisfies*

$$\|p_t^z - \tilde{\pi}\|_{TV} \to 0 \qquad \text{as } t \to \infty$$

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

The solution to the Poisson equation is given next following the direction of [15].

Theorem 2.3. *Under Assumptions 1 and 2, 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* (1.5)*. Moreover, the solution is given by*

$$\phi = \int_0^\infty P_t(f) dt. \tag{2.12}$$

Proof. 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 2.1

$$g_T \to \int_0^\infty P_t(f)dt \tag{2.13}$$

in $L^2(\tilde{\pi})$ as $T \to \infty$, specifically (2.8) with $\varphi = f$ and using (2.10) for $2l$ in place of l. Applying L, it holds that

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

where the exchange in the order of integration is justified by Fubini, (2.8) and the last equality follows by the strong continuity of $(P_t)_{t\geq 0}$ given by Proposition A.2 in Section 6. Inequalities (2.8) and (2.10) (with $2l$ in place of l) also give

$$P_T(f) \to 0 \qquad \text{in } L^2(\tilde{\pi}) \tag{2.14}$$

as $T \to \infty$, so that since L is a closed operator, equations (1.5) and (2.12) hold. In addition, $\int \phi d\tilde{\pi} = 0$ follows from the invariance of $\tilde{\pi}$, Theorem 2.1 and Fubini's theorem. □

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

Theorem 2.4. *Under Assumptions 1 and 2, if $f \in L^2(\tilde{\pi})$ satisfies $\frac{f}{\mathcal{K}_l} \in L^\infty$ for some $l \in \mathbb{N}$, 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 $t \to \infty$ for any initial distribution, where*

$$\sigma_f^2 = 2\int \phi(f - \pi(f))d\tilde{\pi} \tag{2.15}$$

and $\phi \in L_0^2(\tilde{\pi})$ is the solution to (1.5).

Proof. By Corollary 2.2 and Theorem 2.3, the result follows by Theorem 2.6 in [9]. See also Theorem 3.1 in [15]. □

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 3.4, in which a formula for the derivative (1.6) of σ^2 with respect to Γ is provided. The proofs of Proposition 3.1, Lemma 3.2 and Theorem 3.4 are deferred to Section 6.

3.1 Preliminary results and the main formula

In order to establish the formula for the directional derivative, heavy use of the differential operator form (2.3) for the generator is made. Proposition 3.1 establishes that ϕ solves the Poisson equation also as a partial differential equation, which makes use of the Feynman-Kac representation formula for the solution to the Kolmogorov (backward) equation.

Proposition 3.1. *Under Assumptions 1 and 2, if $f \in L_0^2(\tilde{\pi})$ satisfies $\frac{f}{\mathcal{K}_l} \in L^\infty$ for some $l \in \mathbb{N}$, the solution ϕ given by* (2.12) *solves $-\mathcal{L}\phi = f$ in the distributional sense for $\mathcal{L}$ given by* (2.3)*, hence classically if in addition $f \in C^\infty$.*

In order for the integral in a formula like (1.6) to be finite, control on the derivatives in p is required. This will also be used in the proof of Theorem 3.4 and it is given by the following lemma.

Lemma 3.2. *Under Assumptions 1 and 2, if $f \in L_0^2(\tilde{\pi})$ satisfies $\frac{f}{\mathcal{K}_l} \in L^\infty$ for some $l \in \mathbb{N}$, the weak derivative in p of the solution ϕ to $-\mathcal{L}\phi = f$ satisfies*

$$\int |\nabla_p \phi|^2 d\tilde{\pi} < \infty.$$

The following preliminary result is about the solution ϕ under a momentum reversal. It turns out that this is the solution to the Poisson equation associated to the formal $L^2(\tilde{\pi})$-adjoint $\mathcal{L}^*$ of $\mathcal{L}$ which appears in the proof of Theorem 3.4.

Lemma 3.3. *Let Assumptions 1 and 2 hold and let $f \in C^\infty$ and $\tilde{\phi} \in L^2_0(\tilde{\pi}) \cap C^\infty$ be given by* (1.7)*, where ϕ is a classical solution to $-\mathcal{L}\phi = f$. Then $\tilde{\phi}$ is a classical solution to the equation*

$$-\mathcal{L}^*\tilde{\phi} = f, \tag{3.1}$$

where $\mathcal{L}^*$ is the formal $L^2(\tilde{\pi})$-adjoint of $\mathcal{L}$ given by

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

Proof. The equation $\mathcal{L}^*\tilde{\phi} = \widetilde{\mathcal{L}\phi}$ follows by a straightforward calculation. □

If f is not smooth, the equation (3.1) still holds in the distributional sense, since for $g \in C_c^\infty$ and keeping the notation (1.7) for the momentum reversal on arbitrary functions,

$$\int f g d\tilde{\pi} = \int f\tilde{g} d\tilde{\pi} = -\int \phi \mathcal{L}^*\tilde{g} d\tilde{\pi} = -\int \phi \widetilde{\mathcal{L}g} d\tilde{\pi} = -\int \tilde{\phi}\mathcal{L}g d\tilde{\pi}.$$

The main formula of this section for the directional derivative of the asymptotic variance is given next. The directional derivative of $E : \mathbb{S}^n_{++} \to \mathbb{R}$ at $\Gamma \in \mathbb{S}^n_{++}$ in the direction $\delta\Gamma \in \mathbb{S}^n_{++}$ is denoted by $dE(\Gamma).\delta\Gamma = \lim_{\epsilon\to 0} \frac{1}{\epsilon}(E(\Gamma + \epsilon\delta\Gamma) - E(\Gamma))$ whenever the limit exists.

Theorem 3.4. *Under Assumptions 1 and 2, if $f = f(q) \in L^2_0(\pi)$ is continuous, satisfies $\frac{f}{\mathcal{K}_l} \in L^\infty$ for some $l \in \mathbb{N}$ and there exists $\epsilon' > 0$ such that $\Gamma, \Gamma + \epsilon\delta\Gamma \in \mathbb{S}^n_{++}$ for $0 < \epsilon \le \epsilon'$, then the directional derivative of the asymptotic variance σ^2 at Γ in the direction $\delta\Gamma$ has the form*

$$d\sigma^2.\delta\Gamma = -2\int (\nabla_p\phi)^\top \delta\Gamma \nabla_p \tilde{\phi} d\tilde{\pi}, \tag{3.2}$$

where ϕ is the solution (2.12) *to the Poisson equation for the dynamics* (1.1) *at Γ and $\tilde{\phi}$ is given by* (1.7)*.*

As mentioned in the introduction, from (3.2), the direction (1.8) guarantees a decrease in asymptotic variance; similarly the scalar change in Γ given by (1.8) where the outer product is replaced by a dot product guarantees a decrease in σ^2.

3.2 A formula using a tangent process

Equation (3.2) has a more useful form. The first variation process of (1.1) is used here to calculate (1.10); this will be the main methodology used in the numerical sections. This alternative formula given in Theorem 3.5 provides a way to avoid using a finite difference Monte Carlo estimate of the derivative of an expectation. For simplicity, we set $M = I_n$ here. The first variation process associated to (1.1), denoted by $(D_p q_t, D_p p_t) \in \mathbb{R}^{n\times 2n}$ for $t \ge 0$, is defined as the matrix-valued solution to

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

with the initial condition $D_p q_0 = 0$, $D_p p_0 = I_n$. By Theorem 39 of Chapter V in [69], the partial derivatives of (q_t, p_t) with respect to the initial values in p indeed satisfy (3.3) and $(D_p q_t, D_p p_t)$ is continuous with respect to those initial values. Note that there exists a unique solution to (3.3) by Theorem 38 in the same chapter of [69]. 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 3.5. *Let Assumptions 1 and 2 hold. If in addition,*

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

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

where $D : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ *is small enough everywhere, in particular,*

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

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*, is assumed;*

- $f = f(q) \in L^2_0(\pi)$ *is everywhere differentiable and satisfies* $\frac{|f|+|\nabla f|}{\mathcal{K}_l} \in L^\infty$ *for some* $l \in \mathbb{N}$,

then the weak derivative $\nabla_p \phi$ *has the form*

$$\nabla_p \phi(q,p) = \int_0^\infty \mathbb{E}[\nabla f(q_t)^\top D_p q_t] dt, \tag{3.5}$$

where q_t *solves* (1.1) *with initial condition* $(q_0, p_0) = (q, p)$ *and* $D_p q_t$ *solves* (3.3)*, the latter satisfying*

$$|D_p q_t|^2 + |D_p p_t|^2 \leq C' e^{-Ct} \tag{3.6}$$

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

The 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 in order for the integral in (3.5) to be finite. Specifically, U is assumed to be close to some particular quadratic function $q^\top Q q$, cf. [11].

Remark 3.1. Exponential decay of the first variation process is not required, only some uniform (in initial (q_0, p_0)) integrability in time of $D_p q_t$ together with a boundedness assumption on ∇f. On the other hand, Proposition 1 in [20] and Proposition 4 in [57] explores more detailed conditions under which contractivity holds and does not hold.

Proof. Let $b > 0$ be the constant

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

so that $\hat{\lambda}$ reduces to

$$\hat{\lambda} = \min\left(\frac{\lambda_m}{2}, b\frac{U_0}{4}\right)$$

and we have the following bound

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

for some generic constant $C > 0$ independent of the initial values (q_0, p_0) and $\omega \in \Omega$. Consequently, using the (weighted) boundedness assumption on $|\nabla f|$ and for each index i,

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

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

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

Using Theorem 2.1, (3.8) again and dominated convergence to take $T \to \infty$ on both sides concludes the proof. □

At any t, equation (3.5) can be used in a practical way in order to estimate the gradient direction (1.8). Specifically, the following estimator can be used. Given $(q, p) \in \mathbb{R}^{2n}$, which we think of as solutions (q_t, p_t) to (1.1) hence approximately distributed according to $\tilde{\pi}$, let

$$\delta\Gamma = \int_0^T \nabla f(q_t^{(q,p)})^\top D_p q_t^{(q,p)} dt \otimes \int_0^T \nabla f(q_t^{(q,-p)})^\top D_p q_t^{(q,-p)} dt \tag{3.9}$$

for any large enough $T > 0$, where $(q_t^{(q,p)}, p_t^{(q,p)})$, $(q_t^{(q,-p)}, p_t^{(q,-p)})$ solve (1.1) with independent realisations of W_t and D_p has been used to denote the derivative with respect to the initial p as above. The next result is that this estimator has finite variance. Note that for (q, p) distributed away from stationarity, the estimator cannot be expected to be an unbiased estimator of (1.8).

Theorem 3.6. *Let the assumptions of Theorem 3.5 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 (3.7),

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

so that using the (weighted) boundedness assumption on $|\nabla f|$ together with Theorem 2.1 and Fubini's theorem, the proof concludes. □

4 Quadratic cases

Throughout this section, the target measure π is assumed to be Gaussian, when mean zero this is given by $\pi \propto \exp(-\frac{1}{2} q^\top \Sigma^{-1} q)$ 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 observables, we look for solutions to the Poisson equation 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 3.4 as well as intuition for how Γ can be improved in concrete cases. We will consider the following cases.

1. Quadratic $f = \frac{1}{2}q^\top U_0 q$, given commutativity between U_0 and Σ (Proposition 4.5), also $f = \frac{1}{2}U_0 q^2 + lq$ in one dimension (Proposition 4.6);
2. Odd polynomial f, where the asymptotic variance will be shown to decrease to zero as $\Gamma \to 0$ (Proposition 4.7, Corollary 4.8 and Proposition 4.9);
3. Quartic f in one dimension, in which the situation is similar to quadratic f (Proposition 4.10);

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 $f : \mathbb{R}^{2n} \to \mathbb{R}$ are given by

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

and the value $\pi(f)$ becomes

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

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

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

Consider the natural candidate solution ϕ to the Poisson equation (1.5) given by

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

for some constant matrices $G, E, H \in \mathbb{R}^{n \times n}$ and vectors $g, h \in \mathbb{R}^n$.

Lemma 4.1. *Given f in* (4.1)*, $\pi(f)$ in* (4.2) *and $\mathcal{L}$ of the form* (4.3)*, ϕ given by* (4.4) *is a solution to the Poisson equation* (1.5) *if and only if*

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

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

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

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

Proof. Substituting (4.3), (4.4) and (4.1) into the Poisson equation (1.5), one obtains

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

Comparing the constant, first order and second order coefficients in p give respectively the sufficient conditions (4.5), (4.6) and (4.7) as stated. □

4.1 Quadratic observable

Similar calculations in this situation have appeared previously in Proposition 1 in [26], where explicit expressions analogous to G, E, H and for σ^2 are given. For our purposes of finding an optimal Γ, instead of taking these explicit expressions, we keep unknown antisymmetric matrices (such as A_1) as they appear and eventually use commutativity between Σ and U_0 to show that the antisymmetric matrices are zero. We continue from (4.5), (4.6) and (4.7) with finding explicit expressions for the coefficients G, E, H of ϕ.

Lemma 4.2. *Given f in* (4.1)*, $\pi(f)$ in* (4.2)*, $\mathcal{L}$ of the form* (4.3)*, ϕ given by* (4.4) *is a solution to the Poisson equation* (1.5) *with* (4.3) *if and only if*

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

$$E = \frac{1}{2}U_0\Sigma + \frac{1}{2}A_2\Sigma, \tag{4.9}$$

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

$$h = \Sigma l \qquad \text{and} \qquad g = \Gamma\Sigma l. \tag{4.11}$$

for some antisymmetric matrices A_1, A_2.

Proof. Comparing coefficients in q in equation (4.5) gives

$$2\Gamma : H_S = U_0 : \Sigma \tag{4.12}$$

$$h^\top \Sigma^{-1} = l^\top \tag{4.13}$$

$$2E\Sigma^{-1} = U_0 + A_2 \tag{4.14}$$

and the same for condition (4.6) gives

$$M^{-1}G_S = H_S\Sigma^{-1} + M^{-1}\Gamma E^\top, \tag{4.15}$$

$$M^{-1}g = M^{-1}\Gamma h. \tag{4.16}$$

Condition (4.14) yields (4.9). Together with (4.7), this gives (4.10). From the expression (4.10) and by symmetry of U_0, condition (4.12) 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 (4.15) then gives (4.8). Equations (4.13) and (4.16) give the equations (4.11) for g and h. □

The asymptotic variance from Theorem 2.4 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 [26].

Lemma 4.3. *If the solution ϕ to the Poisson equation* (1.5) *for f given by* (4.1)*, $\pi(f)$ given by* (4.2) *and $\mathcal{L}$ given by* (4.3) *is of the form* (4.4)*, the asymptotic variance σ^2 given by* (2.15) *has the expression*

$$2\langle \phi, f - \pi(f)\rangle_{\tilde{\pi}} = Tr(G_S\Sigma U_0\Sigma) + 2g \cdot \Sigma l. \tag{4.17}$$

Proof. 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 and $f - \pi(f)$ are given by

$$\phi(z) = \frac{1}{2}z \cdot \bar{G}z - \bar{g} \cdot z - \frac{1}{2}\bar{G} : \bar{\Sigma}$$

$$f(z) - \pi(f) = \frac{1}{2}z \cdot \bar{U}_0 z - \bar{l} \cdot z - \frac{1}{2}\bar{U}_0 : \bar{\Sigma}$$

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

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

where

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

As a result,

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

□

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. *Let $\Sigma U_0 = U_0\Sigma$, 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* (1.5)*. The following holds.*

$$\lim_{M \to 0} \int \phi(f - \pi(f))d\tilde{\pi} = \inf_{M \in \mathbb{S}^n_{++}} \int \phi(f - \pi(f))d\tilde{\pi},$$

where the limit on the left hand side is in the sense of $M = mI_n$, $m \to 0^+$.

Proof. Let

$$G = \frac{1}{2}M\Sigma U_0\Gamma^{-1}\Sigma^{-1} + \frac{1}{2}\Gamma U_0\Sigma, \quad E = \frac{1}{2}U_0\Sigma, \quad H = \frac{1}{2}\Sigma U_0\Gamma^{-1}, \tag{4.18a}$$

$$g = \Gamma\Sigma l, \quad h = \Sigma l \tag{4.18b}$$

so that by Lemma 4.2, ϕ given by (4.4) is the solution to the Poisson equation (1.5) and inserting G, g into (4.17) gives

$$2\langle \phi, f - \pi(f)\rangle_{\tilde{\pi}} = \frac{1}{2}\mathrm{Tr}(M\Sigma U_0\Gamma^{-1}U_0\Sigma + \Gamma U_0\Sigma^2 U_0\Sigma) + 2l^\top\Sigma\Gamma\Sigma l. \tag{4.19}$$

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

Proposition 4.4 solves the optimisation problem in M in the stated setting, but highlights the corresponding discrete time problem since one cannot take $M^{-1} \to \infty$ in practice. We focus on the optimisation of Γ and fix $M = I_n$ in the following.

Proposition 4.5. *Let Σ, U_0, l, M be such that*

$$\Sigma U_0 = U_0 \Sigma, \qquad l = 0, \qquad M = I_n, \tag{4.20}$$

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 (1.5). The following holds.

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

where $\mathbb{S}_\Sigma$ is the set of symmetric positive definite matrices commuting with Σ and 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 $(\lambda_i)_{1 \le i \le n}$, $\lambda_i \ge 0$. Setting again (4.18), ϕ given by (4.4) is the solution to the Poisson equation (1.5) and the asymptotic variance σ^2 given by (2.15) becomes

$$2\langle \phi, f - \pi(f) \rangle_{\tilde{\pi}} = \frac{1}{2} \mathrm{Tr}(\Sigma U_0 \Gamma^{-1} U_0 \Sigma + \Gamma U_0 \Sigma^2 U_0 \Sigma), \tag{4.21}$$

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. □

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

Proposition 4.6. *If $n = 1$, $U_0 \neq 0$, $l \neq 0$, $f : \mathbb{R} \to \mathbb{R}$ 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 (1.5), then*

$$\min_{\Gamma > 0} 2 \int \phi(f - \pi(f)) d\tilde{\pi} = M^{\frac{1}{2}} \Sigma^2 U_0^2 (\Sigma + 4 l^2 U_0^{-2})^{\frac{1}{2}}$$

and the minimum is attained by

$$\Gamma = \frac{M^{\frac{1}{2}}}{(\Sigma + 4 l^2 U_0^{-2})^{\frac{1}{2}}}. \tag{4.22}$$

Proof. By Lemma 4.2, the solution (4.4) to the Poisson equation (1.5) is

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

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 (4.22) as claimed. □

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 4.7. *Under Assumption 1 and 2, 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 (2.3) *and ϕ is the solution to the Poisson equation* (1.5)*, then the asymptotic variance satisfies*

$$\inf_{\Gamma \in \{\gamma I_n : \gamma > 0\}} 2 \int \phi (f - \pi(f)) d\tilde{\pi} = 0. \tag{4.23}$$

Proof. Let $\Gamma = \gamma I_n$, $\gamma \in \mathbb{R}$.Note there is a unique solution $\phi \in L^2_0(\tilde{\pi})$ to (1.5) by Theorem 2.3. The solution ϕ to (1.5) 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 $dq_1 \dots dq_{j-1} dq_{j+1} \dots dq_n$. □

Corollary 4.8. *Given a Gaussian target measure with density $\pi \propto e^{-U}$ on $\mathbb{R}^n$, observable $f : \mathbb{R}^n \to \mathbb{R}$ as in* (4.1) *with $U_0 = 0$, that is,*

$$f(q) = l \cdot q,$$

where $l \in \mathbb{R}^n$, $\pi(f) = 0$, $\mathcal{L}$ of the form (2.3) *and ϕ the solution to the Poisson equation* (1.5)*, equation* (4.23) *holds.*

There is some intuition in the situation in Corollary 4.8. 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 4.7. 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 starting point 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 4.9. *If $n = 1$, $\hat{k} \in \mathbb{N}$ and $f : \mathbb{R} \to \mathbb{R}$ is an odd finite order polynomial observable given by*

$$f(q) = \sum_{i=0}^{\hat{k}} a_i q^{2i+1}, \tag{4.24}$$

$\pi(f) = 0$, $\mathcal{L}$ is of the form (4.3) *and ϕ is the solution to the Poisson equation* (1.5)*, then the asymptotic variance satisfies* (4.23)*.*

The proof of Proposition 4.9 is deferred to Section 6.

4.3 Quartic observable

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

Proposition 4.10. *If $n = 1$ and $f : \mathbb{R} \to \mathbb{R}$ is a quartic observable given by*

$$f(q) = q^4, \tag{4.25}$$

$\pi(f) = 3\Sigma^2$ for some $\Sigma > 0$, $\mathcal{L}$ is of the form (4.3), $M = 1$ and ϕ is the solution to the Poisson equation (1.5), then there exists $\sigma_{quar} > 0$ such that

$$\min_{\Gamma=\gamma>0} 2 \int \phi(f - \pi(f)) d\tilde{\pi} = \sigma_{quar}.$$

The proof of Proposition 4.10 can be found in Section 6.

5 Computation of the change in Γ

Throughout this section, the $M = I_n$ case is considered. As mentioned, the formula (3.2) gives a natural gradient descent direction (1.8) to take Γ in order to optimise σ^2 from Theorem 2.4. In Theorem 3.4 and in the form (1.8), the expression for the gradient is already susceptible to a Green-Kubo approach in the sense that the form (2.12) for ϕ can be substituted in to obtain a trajectory based formula, where finite difference is used to approximate ∇_p and independent realisations of (q_t, p_t) is used for the expectations. However, this is too inaccurate in the implementation to be useful. The more directly calculable form as stated in the introduction in (1.10) is used involving the derivative of (q_t, p_t) with respect to the initial condition in Section 3.2.
We focus the discussion on a Monte Carlo method to approximate $\nabla\phi$ and gradient directions in Γ (e.g. (1.8)) based on Theorem 3.4, but a spectral method to solve (1.5) and compute the change in Γ is given in Appendix B, which is computationally feasible in low dimensions. Algorithm 1 summarises the resulting continuous-time procedure, where all expectations within (1.8) are approximated by single realisations; further justifications, alternative methods, refinements and a concrete implementation (Algorithm 2) along with examples follow.

Algorithm 1: Continuous-time outline of Γ update using (1.6) and (1.10)

Result: $\Gamma \in \mathbb{S}^n_{++}$
Start from arbitrary $(q_0, p_0) \in \mathbb{R}^{2n}$ and set $(\tilde{q}_0, \tilde{p}_0) = (q^0, -p^0)$, $Dq_0 = D\tilde{q}_0 = 0$, $Dp_0 = D\tilde{p}_0 = I_n$, $\zeta = \tilde{\zeta} = 0$, $k = 0$, $\Gamma = I_n \quad \forall 1 \le j \le B$, $t = t_0 = 0$;
for N *epochs* **do**
  simulate one step in q_t, $\tilde{q}_t$ then in $D_p q_t$ and $D_p \tilde{q}_t$ from t to $t + \Delta t$;
  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

5.1 Methodology

Here we describe an on-the-fly procedure to repeatedly calculate the change (1.8) 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 5.1.1. Two gradient procedures, namely gradient descent and the Heavy ball method, for evolving Γ given a gradient are detailed in Section 5.1.2. Then iterates from Section 5.1.1 are used to approximate each change in Γ in Section 5.1.3 (see also Appendix C. The key idea linking the above is that if equation (3.5) holds, then

$$\begin{aligned}\Delta\Gamma &= \int \nabla_p\phi \otimes \nabla_p\tilde{\phi} d\tilde{\pi} \\ &= -\int\int_0^\infty \mathbb{E}[\nabla f(q_s)^\top D_p q_s]^\top ds \int_0^\infty \mathbb{E}[\nabla f(\tilde{q}_t)^\top D_p\tilde{q}_t] dt d\tilde{\pi},\end{aligned} \tag{5.1}$$

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 (5.3) with $\tilde{q}_t$ replacing q_t for the latter and the integral in (5.1) is with respect to (q, p).

5.1.1 Splitting

A BAOAB splitting scheme [50, 51] will be used to integrate the Langevin dynamics (1.1), given explicitly by

$$\begin{cases} p^{i+\frac{1}{3}} &= p^i - \nabla U(q^i)\frac{\Delta t}{2} \\ q^{i+\frac{1}{2}} &= q^i + p^{i+\frac{1}{3}}\frac{\Delta t}{2} \\ p^{i+\frac{2}{3}} &= \exp(-\Delta t\Gamma^i)p^{i+\frac{1}{3}} + \sqrt{1-\exp(-2\Delta t\Gamma^i)}\xi^i \\ q^{i+1} &= q^{i+\frac{1}{2}} + p^{i+\frac{2}{3}}\frac{\Delta t}{2} \\ p^{i+1} &= p^{i+\frac{2}{3}} - \nabla U(q^{i+1})\frac{\Delta t}{2} \end{cases} \tag{5.2}$$

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. [19, 20, 32, 58, 73, 75], 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 (1.8) for $\Delta\Gamma$, so that no further deliberation is made about the choice of discretisation for the purposes here. The first variation process (5.3) together with its initial condition is

$$D_p q_t = \int_0^t D_p p_s ds, \tag{5.3a}$$

$$D_p p_t = I_n - \int_0^t (D^2U(q_s)D_p q_s + \Gamma D_p p_s) ds. \tag{5.3b}$$

In order to simulate (5.3), an analogous splitting scheme is used:

$$\begin{cases} Dp^{i+\frac{1}{3}} &= Dp^i - D^2U(q^i)Dq^i\frac{\Delta t}{2} \\ Dq^{i+\frac{1}{2}} &= Dq^i + Dp^{i+\frac{1}{3}}\frac{\Delta t}{2} \\ Dp^{i+\frac{2}{3}} &= \exp(-\Delta t\Gamma^i)Dp^{i+\frac{1}{3}} \\ Dq^{i+1} &= Dq^{i+\frac{1}{2}} + Dp^{i+\frac{2}{3}}\frac{\Delta t}{2} \\ Dp^{i+1} &= Dp^{i+\frac{2}{3}} - D^2U(q^{i+1})Dq^i\frac{\Delta t}{2}. \end{cases} \tag{5.4}$$

The kth column of the first term including the Hessian of U (and similarly for the last) can be approximated by

$$-\nabla U(q^i + \frac{\Delta t}{2}(Dq^i)_k) + \nabla U(q^i) \tag{5.5}$$

where $(Dq^i)_k$ denotes the kth column of Dq^i, so that (5.3) can still be approximated in the absence of Hessian evaluations. The approximation (5.5) will be used only when explicitly stated in the sequel.

5.1.2 Gradient procedure in Γ

Suppose we have available a series of proposal updates $(b_0, \ldots, b_{L-1}) \in \mathbb{R}^{n\times n\times L}$ for 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)

$$\Gamma^{i+1} = \Pi^{\mu}_{\mathrm{pd}}\left(\Gamma^i + \frac{\alpha^i}{2L}\sum_{j=0}^{L-1}(b_j + b_j^\top)\right) \tag{5.6}$$

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

$$\Pi^{\mu}_{\mathrm{pd}}(M) = \sum_{i=1}^{n} \max(\lambda_i, \mu) v_i v_i^\top \tag{5.7}$$

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

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

Alternatively, a Heavy-ball method [67, 35] (with projection) can be used. The method is considered in the stochastic gradient context in [18], given here as

$$\begin{cases} \Gamma^{i+1} & = \Pi^{\mu}_{\mathrm{pd}}(\Gamma^i + \alpha^i \Theta^{i+1}), \\ \Theta^{i+1} & = (1 - \alpha^i r)\Theta^i + \frac{\alpha^i}{2L}\sum_{j=0}^{L-1}(b_j + b_j^\top). \end{cases} \tag{5.8}$$

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 $\nabla\sigma^2(\Gamma^i_k)$ and variance $\sigma_b^2 > 0$, then the system (5.8) 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.

5.1.3 A thinning approach for $\Delta\Gamma$

The most straightforward way of approximating the integral in (5.1) is to use independent realisations of (5.2), as described in Appendix C, but we draw alternatively a thinned sample [64] from a single trajectory here in order to run only a single parallel set of realisations of (5.2) and (5.4) at a time. More specifically, we consider a single realisation of (5.2) 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 (5.4) is calculated in parallel. A precise description follows.
Let $K = 1$ for simplicity. The Γ direction (5.1) is approximated by

$$\begin{aligned} -\frac{1}{(L+L^*)}\sum_{l=0}^{L+L^*-1}&\left(\sum_{i=1}^{T}\frac{\Delta t}{K}\sum_{k=1}^{K}\nabla f(q_{(k)}^{i+Tl+B})^\top Dq_{(k)}^{i+Tl+B}\right)\otimes \\ &\left(\sum_{i=1}^{T}\frac{\Delta t}{K}\sum_{k=1}^{K}\nabla f(\tilde{q}_{(k)}^{i+Tl+B})^\top D\tilde{q}_{(k)}^{i+Tl+B}\right), \end{aligned} \tag{5.9}$$

where $L \in \mathbb{N}$, $((q^i_{(k)}, p^i_{(k)}))_{i\in\mathbb{N}}$, $((\tilde{q}^i_{(k)}, \tilde{p}^i_{(k)}))_{i\in\mathbb{N}}$ denote solutions to (5.2)

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

with initial condition $(0, 0)$, noise $\xi^i = \xi^i_{(k)}, \tilde{\xi}^i_{(k)}$ for all $i \in \mathbb{N}$ satisfying $\xi^i_{(k)} = \xi^i_{(k')} = \tilde{\xi}^i_{(k)} = \tilde{\xi}^i_{(k')}$ for all $i < B, 1 \leq k \leq K$, $1 \leq k' \leq K$, independent otherwise as i and k vary, along with corresponding $(Dq^i_{(k)}, Dp^i_{(k)})$, $(D\tilde{q}^i_{(k)}, D\tilde{p}^i_{(k)})$ satisfying (5.4) for $i \neq B + Tl - 1$, $l \in \mathbb{N}$ (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,

$$q^{Tl+B}_{(k)} = q^{Kl+B}_{(1)}, \qquad p^{Tl+B}_{(k)} = p^{Tl+B}_{(1)}, \qquad Dq^{Tl+B}_{(k)} = 0, \qquad Dp^{Tl+B}_{(k)} = I_n \tag{5.10a}$$

$$\tilde{q}^{Tl+B}_{(k)} = q^{Tl+B}_{(1)}, \qquad \tilde{p}^{Tl+B}_{(k)} = -p^{Tl+B}_{(1)}, \qquad D\tilde{q}^{Tl+B}_{(k)} = 0, \qquad D\tilde{p}^{Tl+B}_{(k)} = I_n \tag{5.10b}$$

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

$$\max_{i,j,k} \left| (Dq^{Tl+B}_{(k)})_{ij} \right| < D_{\text{conv}}, \quad \max_{i,j,k} \left| (D\tilde{q}^{Tl+B}_{(k)})_{ij} \right| < D_{\text{conv}}, \tag{5.11a}$$

$$\max_{i,j,k} \left| (Dp^{Tl+B}_{(k)})_{ij} \right| < D_{\text{conv}}, \quad \max_{i,j,k} \left| (D\tilde{p}^{Tl+B}_{(k)})_{ij} \right| < D_{\text{conv}} \tag{5.11b}$$

and $L^* \in \mathbb{N}$ is such that the number of elements in $\{l \in \mathbb{N} : 1 \leq l \leq L + L^*\}$ satisfying (5.11) 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 (5.9) for the continuous-time expression (5.1) may well worsen, since the integrand (of the outermost integral) in (5.1) is not f. Increasing L is expected to solve any resulting issues; extremely small L have been successful in the experiments here.

Remark 5.1. 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 ith observable. In the implementation in Algorithm 2, instead of only the vectors ζ, $\tilde{\zeta}$, this amounts to calculating at each iteration the vectors $\zeta^{(i)}$, $\tilde{\zeta}^{(i)}$ corresponding to the ith 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 5.2. *(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 ((5.3) and) (5.4) by v, one obtains

$$\begin{cases} Dpv^{i+\frac{1}{3}} &= Dpv^i - D^2U(q^i)Dqv^i \frac{\Delta t}{2} \\ Dqv^{i+\frac{1}{2}} &= Dqv^i + Dpv^{i+\frac{1}{3}} \frac{\Delta t}{2} \\ Dpv^{i+\frac{2}{3}} &= \exp(-\Delta t \Gamma^i) Dpv^{i+\frac{1}{3}} \\ Dqv^{i+1} &= Dqv^{i+\frac{1}{2}} + Dpv^{i+\frac{2}{3}} \frac{\Delta t}{2} \\ Dpv^{i+1} &= Dpv^{i+\frac{2}{3}} - D^2U(q^{i+1})Dqv^i \frac{\Delta t}{2}, \end{cases} \tag{5.12}$$

where the first term involving the Hessian of U in (5.12) can be approximated by

$$-\nabla U(q^i + \frac{\Delta t}{2} Dqv^i) + \nabla U(q^i)$$

and similarly for the last such term. In continuous time, the resulting direction in Γ is $\int \nabla\phi^\top v \nabla\tilde{\phi}^\top v d\tilde{\pi} v \otimes v$ and from (3.2) the rate of change in asymptotic variance in this direction is $-2(\int \nabla\phi^\top v \nabla\tilde{\phi}^\top v d\tilde{\pi})^2$. However, the resulting gradient procedure in Γ turns out to be painstakingly slow to converge in high dimensions in comparison to simulating a full first variation process; as illustration, one can think of the situation where the randomly chosen vector v is taken from the restricted set of standard Euclidean basis vectors, where only one diagonal value in Γ is changed at a time. For a directional derivative, we also mention [79, 36].

Algorithm 2: Gradient procedure in Γ

Result: $\Gamma^i, 1 \le i \le N+1$
Start from arbitrary $(q^0, p^0) \in \mathbb{R}^{2n}$ and set $Dq^0 = D\tilde{q}^0 = 0$, $Dp^0 = D\tilde{p}^0 = I_n$, $\zeta = \tilde{\zeta} = 0$, $k = 0$,
$\Gamma^j = I_n \quad \forall 1 \le j \le B$;
for $i = 1 : B-1$ **do**
  compute q^{i+1} according to (5.2);
end
if $i = B$ **then**
  set $(\tilde{q}^i, \tilde{p}^i) \leftarrow (q^i, -p^i)$;
end
for $i = B : N$ **do**
  compute q^{i+1} and $\tilde{q}^{i+1}$ according to (5.2);
  compute Dq^{i+1} and $D\tilde{q}^{i+1}$ from (5.4) corresponding to q^{i+1} and $\tilde{q}^{i+1}$ respectively;
  compute the row vectors $\zeta \leftarrow \zeta + \nabla f(q^{i+1})^\top Dq^{i+1}\Delta t$;
  $\tilde{\zeta} \leftarrow \tilde{\zeta} + \nabla f(\tilde{q}^{i+1})^\top D\tilde{q}^{i+1}\Delta t$
  if $l := i - B \in T\mathbb{N}$ *and* (5.11) *hold (ignoring appearances of* $(k) = (1)$*)* **then**
    save the matrix $b_{(\frac{k}{G} - \lfloor \frac{k}{G} \rfloor)G} = -\zeta \otimes \tilde{\zeta}$;
    reset as follows: $\zeta, \tilde{\zeta} \leftarrow 0$, $(\tilde{q}^{i+1}, \tilde{p}^{i+1}) \leftarrow (q^{i+1}, -p^{i+1})$;
    $Dq^{i+1}, D\tilde{q}^{i+1} \leftarrow 0$, $Dp^{i+1}, D\tilde{p}^{i+1} \leftarrow I_n$
    and update the counter $k \leftarrow k+1$;
  end
  if $k \in G\mathbb{N}$ **then**
    compute Γ^{i+1} according to (5.8);
  else
    set $\Gamma^{i+1} = \Gamma^i$;
  end
end

5.2 Concrete examples

In Sections 5.2.1, 5.2.2 and 5.2.3, the Monte Carlo approach is applied on concrete problems. Section 5.2.1 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 5.2.2, where the algorithm is shown to approximate a Γ matrix that exhibits an even better empirical asymptotic variance than the one given by Proposition 4.5. Finally, the algorithm is applied to finding the optimal Γ in estimating the posterior mean in a Bayesian inference problem in Section 5.2.3, where the situation is shown to be similar to Proposition 4.8, in the sense that the optimal Γ is close to 0; after and separately from such a finding, the empircal 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.

5.2.1 One dimensional quadratic case

Here the algorithm given in Section 5.1.3 is used in the simplest one dimensional

$$U(q) = \frac{V_0}{2}q^2, \qquad f(q) = \frac{1}{2}q^2, \tag{5.13}$$

$V_0 > 0$, 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 4.5 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 (4.17), which reduces in this case to

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

The case $V_0 = 5$ is illustrated in Figure 5.3. In the middle and right plot of Figure 5.3, the procedure in Section 5.1.3 is used for $5 \cdot 10^4$ epochs, with $\Delta t = 0.08$, block-size $T = 125$, $L = 1$ and $D_{\text{conv}} = 2 \cdot 10^{-4}$. Changing the observable to the linear

$$f(q) = q \tag{5.14}$$

gives that the 'optimal' (but unreachable in the algorithm due to the constraints) friction is 0 by Corollary 4.8. The right plot in Figure 5.3 shows that the procedure arrives at a similar conclusion in the sense that the Γ hits and stays at $\mu = 0.2$.

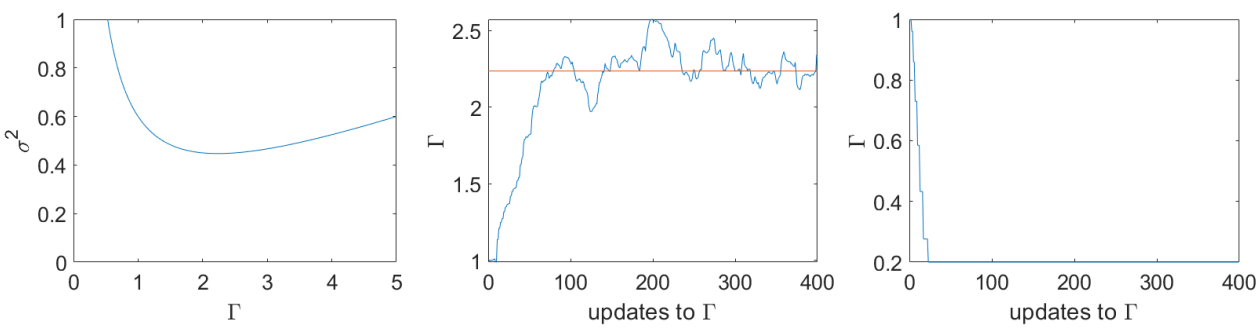

Figure 5.1: Left: Relationship between asymptotic variance and Γ for (5.13). Middle and right: Trajectory of Γ for (5.13) and (5.14) respectively by (5.8) with $\alpha^i = 1$, $G = 1$, $r = 0.5$ and $\mu = 0.2$. Middle: the red line is the optimal value $\Gamma = \sqrt{5}$ given by Proposition 4.5. All plots are for $V_0 = 5$.

5.2.2 Diffusion bridge sampling

The algorithm in Section 5.1.3 is applied in the context of diffusion bridge sampling [38, 40] (see also for example [7, 21, 39]), where the SDE

$$dx_t = -\nabla V(x_t)dt + \sqrt{2\beta^{-1}}dW_t' \tag{5.15}$$

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

$$x_0 = x_- \quad \text{and} \quad x_T = x_+ \tag{5.16}$$

for some fixed $T > 0, x_0, x_+ \in \mathbb{R}^d$ and the problem setting is to sample from the path space of solutions to (5.15) conditioned on (5.16). For the derivation of the following formulations, we refer to Section 5 in [38] and Section 6.1 in [6]; here we extract a simplified potential U to apply our algorithm on after a brief description.
Let

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

Using the measure given by Brownian motion conditioned on (5.16) as the reference measure μ_0 on the path space of continuous functions $C([0,1], \mathbb{R})$, the measure μ associated to (5.15) conditioned on (5.16) satisfies

$$\frac{d\mu}{d\mu_0}(x) \propto \exp\left(-\frac{1}{4}\int_0^T |x|^2 dt\right),$$

where the left hand side denotes the Radon-Nikodym derivative, so that discretising μ on a grid in $[0,1]$ with grid-size $\delta > 0$ gives the approximating measure

$$\pi(q_1, \ldots, q_n) \propto e^{-U(q_1,\ldots,q_n)}$$

where U is given by

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

From here the Langevin system (1.1) can be used to sample from π and the algorithm given in Section 5.1.3 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 5.2.

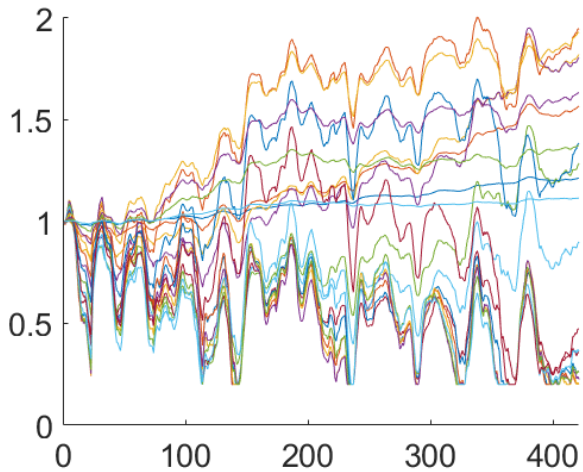

Figure 5.2: Diagonal values of Γ over iterations of (5.8) with $\alpha^i = 0.2$, $G = 5$, $r = 1$ and $\mu = 0.2$.

At the end of 300000 epochs, Γ is given by

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

This Γ is 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,

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

and this is compared to the estimate from the same procedure using different values of fixed Γ in Table 5.1. Note that $\Gamma = \Sigma^{-\frac{1}{2}}$ is the optimal Γ in the restricted class of matrices commuting with Σ given by Proposition 4.5, where the asymptotic variance is known to be $\text{Tr}(\Sigma^{\frac{5}{2}}) \approx 6.4785$.

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

Table 5.1: Empirical asymptotic variances with $N_B = 999$, $T = 300$, $B = 100$, $N = 299700$.

5.2.3 Bayesian inference

We adopt the binary regression problem as in [29] on a dataset[2] with datapoints encoding information about images on a webpage and each labelled with 'ad' or 'non-ad'. The labels $\{Y_i\}_{1\le i\le p}$, 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 $\{\rho(\beta^\top X_i)\}_{1\le i\le p}$, 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 (5.18), $\{X_i\}_{1\le i\le p}$, β, both taking values in $\mathbb{R}^n$, are respectively vectors of known features from each datapoint and regression parameters to be determined. The parameters β are given the 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

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

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 $f_i(q) = q_i$, $1 \le i \le n$, 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

$$-\nabla U(\hat\beta) = \Sigma^{\frac{1}{2}}\sum_{i=1}^{p} cX_i(Y_i - (1+e^{-c(\Sigma^{\frac{1}{2}}\hat\beta)^\top X_i})^{-1}) - \hat\beta. \tag{5.17}$$

We use the observable vector

$$f_i(\hat\beta) = \hat\beta_i, \quad 1 \le i \le n$$

and the sum of their corresponding asymptotic variances as the value to optimise with respect to Γ, but show in Figures 5.3 and 5.4 the estimated asymptotic variances for both sets $f_i(\hat\beta)$, $f_i(\beta)$ of observables, where the estimation is calculated using the vector on the left of the outer product in (5.9) in accordance with $2\int \nabla\phi^\top\Gamma\nabla\phi d\tilde\pi$ which follows from the formula (2.15) after integrating by parts with truncation. The approximation (5.5) for the term(s) including the Hessian in (5.4) has been used to test the method despite the explicit availability of the Hessian. During the execution of Algorithm 2, the constant c has been set to

$$c = \bar c := \frac{5}{\max_i(\Sigma^{\frac{1}{2}}\sum_j X_jY_j)_i}. \tag{5.18}$$

[2] 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, $n = 642$.

In detail, 30000 epochs are simulated; after 100 burn-in iterations of the Langevin discretisation (5.2), 2 parallel simulations of (5.2) and 2 of the first variation discretisation (5.4) are run according to Section 5.1.3 with time-step $\Delta t = 0.1$, block-size $T = 100$, $L = 1$ block per update in Γ, $K = 1$ and tolerance $D_{\text{conv}} = 0.01$.

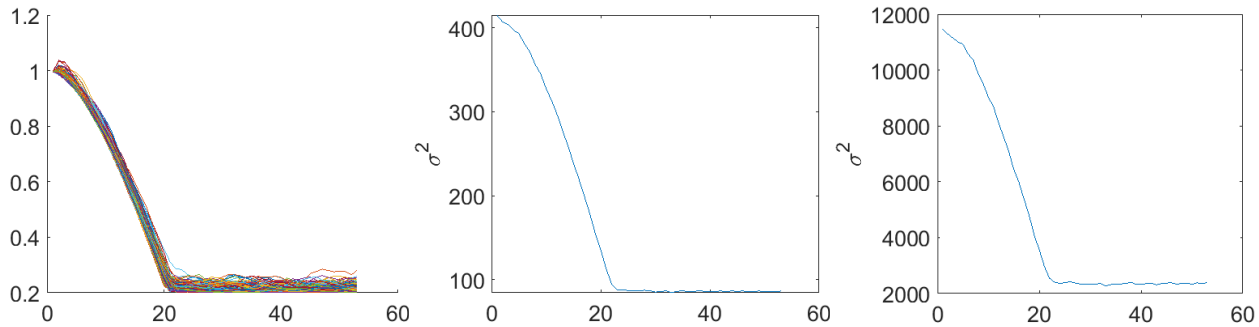

Figure 5.3: Left: Diagonal values of Γ over iterations of (5.8) with $\alpha^i = 0.1$, $G = 1$, $r = 1$ and $\mu = 0.2$. Note that the mean of the absolute values of all entries of Γ at the end of the iterations is 0.0039. Middle: Sum over i of estimated asymptotic variances for $f_i(\hat{\beta})$; right: for $f_i(\beta)$.

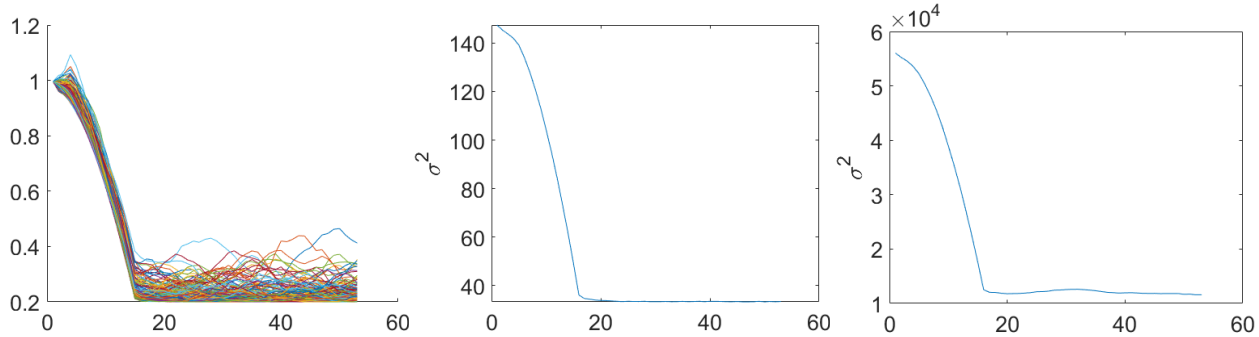

Figure 5.4: The same as in the caption of Figure 5.4, except $r = 0.5$ and a different dataset (https://archive.ics.uci.edu/ml/datasets/Musk+(Version+1)) is used where $n = 167$ and $p = 476$. The mean of the absolute values of all entries of Γ at the end of the iterations is 0.0210.

In Figures 5.3 and 5.4, Γ starts initially from the identity I_n and descends towards $0.2I_n$ (restricted as in (5.7)), as expected for a linear observable and potential close to a quadratic (see Proposition 4.9). We note that in the gradient descent procedure for Γ, using the minibatch gradient does not change the behaviour shown in Figures 5.3 and 5.4. 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.

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 (5.17) and a minibatch[3] version where the sum in (5.17) is replaced by $\frac{p}{10}$ times a sum over a subset S of $\{1, \ldots, p\}$ with 10 elements randomly drawn without replacement such that S changes once for each i in (5.2). In the minibatch gradient case, c is set to a fraction of (5.18), specifically $\bar{c}(\frac{p}{10})^{-1}$. In Tables 5.2 and 5.3, 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 5.2 and 5.3, $N = 29700$ epochs of (5.2) are simulated with the same parameters as above. The asymptotic variance for each observable entry is approximated using block averaging (Section 2.3.1.3

[3]The control variate stochastic gradient on underdamped dynamics [17, 63] is not directly considered here but the benefits of an improved Γ is expected to carry over to such variations of the stochastic gradient.

in [53]) by grouping the epochs after $B = 100$ burn-in iterations into $N_B = 99$ blocks of $T = 300$ epochs, that is,

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

and $N_B = 3$ blocks of $T = 9900$ epochs (respectively for each column of Tables 5.2 and 5.3); the values 0.8667 and 0.1571 approach and correspond to values in the middle plot of Figure 5.3 after multiplying by $n = 642$. The variances are compared to those using a gradient oracle: unadjusted (overdamped) Langevin dynamics[29] and with an irreversible perturbation[26], 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 $1 \le i, j \le n$ 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 5.5, where $\hat{\pi}(f)$ is the mean (over the methods listed in Tables 5.2 and 5.3) 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.

	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)

Table 5.2: $(\frac{1}{n}\sum_{k=1}^{n}\sigma^2_{k,\text{approx}}, \frac{1}{n}\sum_{k=1}^{n}(\sigma^2_{k,\text{approx}} - \frac{1}{n}\sum_{l=1}^{n}\sigma^2_{l,\text{approx}})^2)$ - Empirical asymptotic variances, mean and variance over observable entries, where full gradients have been used.

	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)

Table 5.3: The same as in Table 5.2, except for minibatch gradients

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 irrreversible perturbation.

6 Proofs

Proof. (of Proposition 3.1) Take an approximating sequence $(f_k)_{k\in\mathbb{N}}$ such that $f_k \in C_c^\infty(\mathbb{R}^{2n})$, $f_k \to f$ in $L^2(\tilde{\pi})$. Fix $h \in C_c^\infty(\mathbb{R}^{2n})$ and for the differential operator $\mathcal{L}^* = -p^\top M^{-1}\nabla_q + \nabla U(q)^\top\nabla_p - p^\top M^{-1}\Gamma\nabla_p + \nabla_p^\top\Gamma\nabla_p$ consider

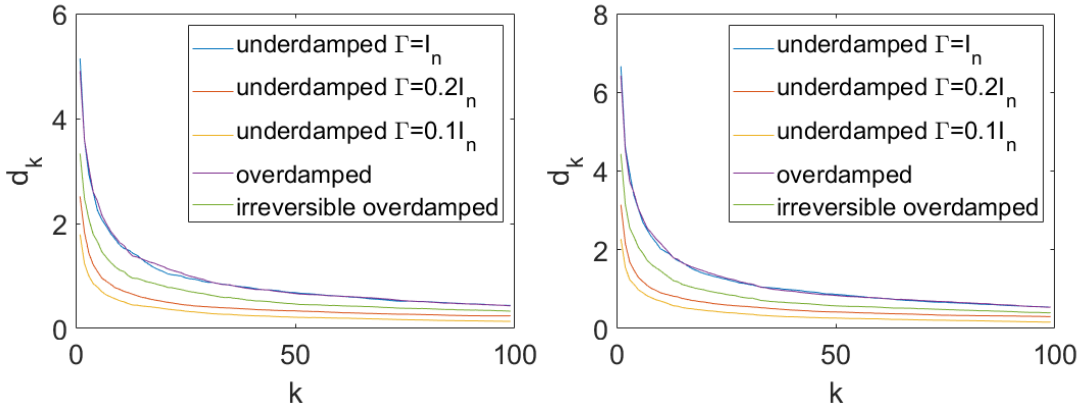

Figure 5.5: Euclidean distances to a combined posterior mean estimate over time. Left: full gradient. Right: minibatch gradient.

the expression

$$\begin{aligned}\left|\int \mathcal{L}^* h\phi d\tilde{\pi} - \int hf d\tilde{\pi}\right| \leq & \left|\int \mathcal{L}^* h\phi d\tilde{\pi} - \int \mathcal{L}^* h \int_{\frac{1}{T}}^{T} P_t(f) dt d\tilde{\pi}\right| \\ & + \left|\int h(f - P_{\frac{1}{T}}(f) + P_T(f)) d\tilde{\pi}\right| \\ & + \left|\int \mathcal{L}^* h \int_{\frac{1}{T}}^{T} P_t(f - f_k) dt d\tilde{\pi}\right| \\ & + \left|\int h(P_{\frac{1}{T}}(f - f_k) - P_T(f - f_k)) d\tilde{\pi}\right| \\ & + \left|\int \mathcal{L}^* h \int_{\frac{1}{T}}^{T} P_t(f_k) dt d\tilde{\pi} + \int h(P_{\frac{1}{T}}(f_k) - P_T(f_k)) d\tilde{\pi}\right|. \end{aligned} \tag{6.1}$$

For any $\epsilon > 0$, T can be chosen so that the first two terms on the right are each bounded by $\frac{\epsilon}{4}$ due to (2.13), the strongly continuous property of P_t and (2.14); subsequently k can be chosen so that the third and fourth terms are each bounded by $\frac{\epsilon}{4}$. For the remaining term, using Assumption 1 for Theorem 5.13 in the first chapter of [49] and Hörmander's theorem [43],

$$\partial_t P_t(f_k) = \mathcal{L} P_t(f_k) \tag{6.2}$$

holds classically on $(0, T) \times \mathbb{R}^{2n}$ and

$$P_t(f_k) \in C^\infty((0, T) \times \mathbb{R}^{2n}). \tag{6.3}$$

By Fubini and equation (6.2),

$$\begin{aligned}\int \mathcal{L}^* h \int_{\frac{1}{T}}^{T} P_t(f_k) dt d\tilde{\pi} &= \int_{\frac{1}{T}}^{T} \int \mathcal{L}^* h P_t(f_k) d\tilde{\pi} dt \\ &= \int_{\frac{1}{T}}^{T} \int h \mathcal{L} P_t(f_k) d\tilde{\pi} dt \\ &= \int_{\frac{1}{T}}^{T} \int h \partial_t P_t(f_k) d\tilde{\pi} dt,\end{aligned}$$

so that since (6.3) holds and therefore $P_t(f_k)$ is bounded on $(\frac{1}{T}, T) \times B_R$ for any $k \in \mathbb{N}$, $T, R > 0$, B_R denoting the Euclidean ball in $\mathbb{R}^{2n}$ of radius R, Fubini's theorem can be applied again to obtain

$$\int \mathcal{L}^* h \int_{\frac{1}{T}}^{T} P_t(f_k) dt d\tilde{\pi} = \int h(P_T(f_k) - P_{\frac{1}{T}}(f_k)) d\tilde{\pi},$$

which is that the last term in (6.1) is equal to zero.

□

Before giving the proof for the main formula of the directional derivative of the asymptotic variance, a truncation function is introduced and the membership of $\nabla_p \phi$ in $L^2(\tilde{\pi})$ is shown. The truncation function is constructed to satisfy a property (6.5) related to the generator (2.3); it will be used to robustly integrate by parts when establishing both $\nabla_p \phi \in L^2(\tilde{\pi})$ and the main formula.

Firstly, let $\varphi : \mathbb{R} \to \mathbb{R}$, $\varphi_k : \mathbb{R} \to \mathbb{R}$ be the standard mollifiers together with $\nu_k : \mathbb{R} \to \mathbb{R}$ be given by

$$\varphi(x) := \begin{cases} e^{\frac{1}{x^2-1}} \left(\int_{-1}^{1} e^{\frac{1}{y^2-1}} dy \right)^{-1} & \text{if } -1 < x \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$\varphi_k(x) := \frac{1}{k} \varphi\left(\frac{x}{k}\right), \tag{6.4}$$

$$\nu_k := \varphi_k * \mathbb{1}_{(-\infty, k^2]} \leq 1$$

for $k > 0$.

Lemma 6.1. *Under Assumption 1 and for $k > 0$, let $\eta_k : \mathbb{R}^{2n} \to \mathbb{R}$ be the smooth functions given by*

$$\eta_k(\zeta) = \nu_k(\ln(1 + |\zeta|^2))$$

for all $\zeta \in \mathbb{R}^{2n}$, then the following properties hold:

1. *η_k is compactly supported;*
2. *η_k converges to 1 pointwise as $k \to \infty$;*
3. *for some constant $C > 0$ independent of k,*

$$|\mathcal{L}\eta_k| + (1 + |\zeta|)|\nabla \eta_k| \leq \frac{C}{k}. \tag{6.5}$$

Proof. The first two properties easily hold by definition of η_k. For the third property, denoting

$$\begin{aligned} \mathcal{L}\eta_k = {} & \nu_k'(\ln(1 + |\zeta|^2)) \mathcal{L} \ln(1 + |\zeta|^2) \\ & + \nu_k''(\ln(1 + |\zeta|^2))(\nabla_p \ln(1 + |\zeta|^2))^\top \Gamma \nabla_p \ln(1 + |\zeta|^2). \end{aligned}$$

It can be seen that ν_k' and ν_k'' are estimated by terms at most of order k^{-1}; to see this, for all $x \in \mathbb{R}$,

$$\nu_k(x) = \int_{-\infty}^{k^2} \varphi_k(x - y) dy = \int_{x-k^2}^{\infty} \varphi_k(z) dz,$$

so that

$$0 \geq \nu_k'(x) = -\varphi_k(x - k^2) \geq -k^{-1} \max \varphi$$

and

$$|\nu_k''(x)| = \left|\varphi_k'(x - k^2)\right| \leq k^{-2} \max \varphi'.$$

Therefore there exists a constant $\bar{C} > 0$ such that

$$|\mathcal{L}\eta_k| \leq \bar{C}\Big(k^{-1}\Big|\mathcal{L}\ln(1+|\zeta|^2)\Big| + k^{-2}\Big|(\nabla\ln(1+|\zeta|^2))^\top\Gamma\nabla\ln(1+|\zeta|^2)\Big|\Big). \tag{6.6}$$

Moreover, the expression

$$\begin{aligned}\mathcal{L}\ln(1+|\zeta|^2) &= \frac{\mathcal{L}|\zeta|^2}{1+|\zeta|^2} - \frac{4p^\top\Gamma p}{(1+|\zeta|^2)^2} \\ &= \frac{p^\top M^{-1}q - \nabla U(q)^\top p - p^\top M^{-1}\Gamma p + \mathrm{Tr}\Gamma}{1+|\zeta|^2} - \frac{4p^\top\Gamma p}{(1+|\zeta|^2)^2}\end{aligned}$$

is bounded from above and below by (2.2) and also the second term in (6.6) is bounded above by a direct calculation. Similarly, $(1+|\zeta|)|\nabla\eta_k|$ can be bounded as required by a direct calculation. □

Proof. (of Lemma 3.2) Consider the functions $(f_{k,R})_{k,R\in\mathbb{N}}$ given by

$$f_{k,R} := \varphi_{\frac{1}{k}} * (f\mathbb{1}_{B_R}) \tag{6.7}$$

for $\varphi_{\frac{1}{k}}$ given in (6.4), B_R the radius R ball in $\mathbb{R}^{2n}$ centered at 0 and $\mathbb{1}_{B_R}$ its indicator function. For any $k, R \in \mathbb{N}$, $f_{k,R} \in C_c^\infty$; moreover

$$f_{k_i,R_i} \to f \tag{6.8}$$

in $L^2(\tilde{\pi})$ as $i \to \infty$ for any non-decreasing sequences $(k_i)_{i\in\mathbb{N}}, (R_i)_{i\in\mathbb{N}}$ such that $k_i, R_i \to \infty$. By Theorem 2.3 and Proposition 3.1 together with Hörmander's theorem, there exists $\phi_{k,R} \in C^\infty \cap L_0^2(\tilde{\pi})$ such that

$$-\mathcal{L}\phi_{k,R} = f_{k,R} - \tilde{\pi}(f_{k,R})$$

for each $k, R \in \mathbb{N}$. For $r \in \mathbb{N}$, η_r from Lemma 6.1 and the smallest eigenvalue λ_m of Γ,

$$\begin{aligned}\lambda_m \int \eta_r|\nabla_p\phi_{k,R}|^2 d\tilde{\pi} &\leq \int \eta_r\nabla_p\phi_{k,R}^\top\Gamma\nabla_p\phi_{k,R}d\tilde{\pi} \\ &= -\int \phi_{k,R}\nabla_p\eta_r^\top\Gamma\nabla_p\phi_{k,R}d\tilde{\pi} \\ &\quad - \int \eta_r\phi_{k,R}(-M^{-1}p+\nabla_p)^\top(\Gamma\nabla_p\phi_{k,R})d\tilde{\pi}.\end{aligned} \tag{6.9}$$

The first term on the right can be written as

$$\begin{aligned}-\int \phi_{k,R}\nabla_p\eta_r^\top\Gamma\nabla_p\phi_{k,R}d\tilde{\pi} &= \int \phi_{k,R}(-M^{-1}p+\nabla_p)^\top\Gamma(\nabla_p\eta_r\phi_{k,R})d\tilde{\pi} \\ &= \int \phi_{k,R}^2(-M^{-1}p+\nabla_p)^\top\Gamma\nabla_p\eta_r d\tilde{\pi} \\ &\quad + \int \phi_{k,R}\nabla_p\eta_r^\top\Gamma\nabla_p\phi_{k,R}d\tilde{\pi},\end{aligned}$$

so that

$$\begin{aligned}-\int \phi_{k,R}\nabla_p\eta_r^\top\Gamma\nabla_p\phi_{k,R}d\tilde{\pi} &= \frac{1}{2}\int \phi_{k,R}^2(-M^{-1}p+\nabla_p)^\top\Gamma\nabla_p\eta_r d\tilde{\pi} \\ &= \frac{1}{2}\int \phi_{k,R}^2\mathcal{L}\eta_r d\tilde{\pi} \\ &\quad + \frac{1}{2}\int \phi_{k,R}^2(-p^\top M^{-1}\nabla_q + \nabla U(q)^\top\nabla_p)\eta_r d\tilde{\pi} \\ &\leq \frac{C}{r}\|\phi_{k,R}\|_{L^2(\tilde{\pi})}^2\end{aligned} \tag{6.10}$$

for some generic constant $C > 0$, where the last line follows by (6.5). The second term on the right of (6.9) is

$$
\begin{aligned}
&-\int \eta_r \phi_{k,R}(-M^{-1}p + \nabla_p)^\top (\Gamma \nabla_p \phi_{k,R}) d\tilde{\pi} \\
&\quad = -\int \eta_r \phi_{k,R} \mathcal{L} \phi_{k,R} d\tilde{\pi} + \int \eta_r \phi_{k,R}(p^\top M^{-1} \nabla_q - \nabla U(q)^\top \nabla_p) \phi_{k,R} d\tilde{\pi},
\end{aligned} \tag{6.11}
$$

where the last term integrates by parts to obtain

$$
\begin{aligned}
&\int \eta_r \phi_{k,R}(p^\top M^{-1} \nabla_q - \nabla U(q)^\top \nabla_p) \phi_{k,R} d\tilde{\pi} \\
&\quad = -\frac{1}{2} \int (p^\top M^{-1} \nabla_q - \nabla U(q)^\top \nabla_p) \eta_r \phi_{k,R}^2 d\tilde{\pi},
\end{aligned}
$$

so that plugging back into (6.11) and using again (6.5),

$$
\begin{aligned}
-\int \eta_r \phi_{k,R}(-M^{-1}p + \nabla_p)^\top (\Gamma \nabla_p \phi_{k,R}) d\tilde{\pi} &\leq \|\phi_{k,R}\|_{L^2(\tilde{\pi})} \|f_{k,R} - \tilde{\pi}(f_{k,R})\|_{L^2(\tilde{\pi})} \\
&\quad + \frac{C}{r} \|\phi_{k,R}\|_{L^2(\tilde{\pi})}^2.
\end{aligned}
$$

Plugging into (6.9), together with (6.10),

$$
\lambda_m \int \eta_r |\nabla_p \phi_{k,R}|^2 d\tilde{\pi} \leq \|\phi_{k,R}\|_{L^2(\tilde{\pi})} \|f_{k,R} - \tilde{\pi}(f_{k,R})\|_{L^2(\tilde{\pi})} + \frac{C}{r} \|\phi_{k,R}\|_{L^2(\tilde{\pi})}^2,
$$

that is, $\nabla_p \phi_{k,R} \in L^2(\tilde{\pi})$. Since $\mathcal{L}$ and its parts are linear, the same arguments as above give

$$
\begin{aligned}
&\lambda_m \int |\nabla_p \phi_{k,R} - \nabla_p \phi_{k',R'}|^2 d\tilde{\pi} \\
&\quad \leq \|\phi_{k,R} - \phi_{k',R'}\|_{L^2(\tilde{\pi})} \|f_{k,R} - f_{k',R'} - \tilde{\pi}(f_{k,R}) + \tilde{\pi}(f_{k',R'})\|_{L^2(\tilde{\pi})}
\end{aligned} \tag{6.12}
$$

for any $k', R' \in \mathbb{N}$. Moreover,

$$
|f_{k,R}| = \left| \varphi_{\frac{1}{k}} * (f \mathbb{1}_{B_R}) \right| \leq C \mathbb{1}_{B_{R+1}} \varphi_{\frac{1}{k}} * \mathcal{K}_l \tag{6.13}
$$

for a generic constant $C > 0$ independent of k, R and since $\varphi_{\frac{1}{k}} * \mathcal{K}_l \to \mathcal{K}_l$ uniformly on compact subsets, for any fixed $R \in \mathbb{N}$, there exists $K_R \in \mathbb{N}$ such that $k \geq K_R$ implies

$$
|f_{k,R}| \leq 2C\mathcal{K}_l. \tag{6.14}
$$

Choosing now the sequences $R_i = i$, $k_i = i + \max_{j \leq i} K_j$ for all $i \in \mathbb{N}$. By Theorems 2.1 and 2.3, $i, j \in \mathbb{N}$,

$$
\|\phi_{k_i,R_i} - \phi_{k_j,R_j}\|_{L^2(\tilde{\pi})} \leq C \left\| \frac{f_{k_i,R_i} - f_{k_j,R_j} - \tilde{\pi}(f_{k_i,R_i}) + \tilde{\pi}(f_{k_j,R_j})}{\mathcal{K}_l} \right\|_{L^\infty},
$$

where $C > 0$ is again independent of k, R. Using the definition (6.7), the terms $\left| \frac{\tilde{\pi}(f_{k,R})}{\mathcal{K}_l} \right|$ are bounded uniformly in k and R, which, together with (6.14), implies $\|\phi_{k_i,R_i} - \phi_{k_j,R_j}\|_{L^2(\tilde{\pi})}$ is bounded uniformly in i, j, so that inserting into (6.12) gives

$$
\lambda_m \int \left| \nabla_p \phi_{k_i,R_i} - \nabla_p \phi_{k_j,R_j} \right|^2 d\tilde{\pi} \leq C \|f_{k_i,R_i} - f_{k_j,R_j} - \tilde{\pi}(f_{k_i,R_i}) + \tilde{\pi}(f_{k_j,R_j})\|_{L^2(\tilde{\pi})}.
$$

Together with (6.8), $\nabla_p \phi_{k_i,R_i}$ is a Cauchy sequence, with limit denoted as $g \in L^2(\tilde{\pi})$, so that for any $h \in C_c^\infty$,

$$
\left| \int gh + \int \phi \nabla_p h \right| \leq \left| \int gh - \int \nabla_p \phi_{k_i,R_i} h \right| + \left| \int \phi \nabla_p h - \int \phi_{k_i,R_i} \nabla_p h \right|,
$$

hence

$$
\nabla_p \phi_{k_i,R_i} \to g = \nabla_p \phi \in L^2(\tilde{\pi}). \tag{6.15}
$$

□

Some additional preliminaries are presented here for the proof of Theorem 3.4. For small $\epsilon \in \mathbb{R}$ and some direction $\delta\Gamma \in \mathbb{R}^{n\times n}$ in the space of smooth friction matrices 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

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

where the notation $-\epsilon\mathcal{S}$ will be used for the perturbation on $\mathcal{L}$, that is,

$$-\mathcal{S} = p^\top M^{-1}\delta\Gamma\nabla_p - \nabla_p^\top \delta\Gamma\nabla_p.$$

The formal $L^2(\tilde{\pi})$-adjoint of $\mathcal{L}_\epsilon$ 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 3.4) For $\epsilon \le \epsilon'$, by Theorem 2.3 there exists a solution $\phi + \delta\phi_\epsilon \in L^2_0(\tilde{\pi})$ to the Poisson equation with the perturbed generator

$$-L_\epsilon(\phi + \delta\phi_\epsilon) = f - \pi(f).$$

By Theorem 2.4, the directional derivative of $\sigma^2(\Gamma)$ in the direction $\delta\Gamma : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is

$$\frac{1}{2}d\sigma^2.\delta\Gamma = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\int \delta\phi_\epsilon f d\tilde{\pi}. \tag{6.16}$$

Let $(f_{k,R})_{k,R\in\mathbb{N}}$ be given by (6.7). Since inequality (6.13) holds by definition and $\varphi_{\frac{1}{k}} * g \to g$ uniformly on compact subsets for any continuous g, there exists for each $R \in \mathbb{N}$ a constant $\hat{K}_R \in \mathbb{N}$ such that $k \ge \hat{K}_R$ implies (6.14) for C independent of k, R and also

$$|f_{k,R} - f| \le \frac{1}{R} \quad \text{on } B_{R-1}.$$

The sequences $R_i = i$, $k_i = i + \max_{j\le i}\hat{K}_j$ for $i \in \mathbb{N}$ then give the sequence $(f_i)_{i\in\mathbb{N}}$, $f_i := f_{k_i,R_i} \in C_c^\infty$ for all $i \in \mathbb{N}$, that satisfies

$$\left\|\frac{f_i - f}{\mathcal{K}_{2l}}\right\|_{L^\infty} \to 0 \qquad \text{as } i \to \infty,$$

which implies

$$\|f_i - f\|_{L^2(\tilde{\pi})} \to 0 \tag{6.17}$$

by (2.10) or by definition. Moreover, $\pi(f_i) \to 0$ by (6.14) and dominated convergence. Therefore the solutions $\phi_i, \phi_{i,\epsilon} \in L^2_0(\tilde{\pi})$ to the Poisson equations

$$-L\phi_i = f_i - \pi(f_i), \qquad -L_\epsilon\phi_{i,\epsilon} = f_i - \pi(f_i)$$

given by Theorem 2.3 satisfy

$$\|\phi_i - \phi\|_{L^2(\tilde{\pi})} + \|\phi_{i,\epsilon} - (\phi + \delta\phi_\epsilon)\|_{L^2(\tilde{\pi})} \to 0 \quad \text{as } i \to \infty \tag{6.18}$$

by (2.12) and Theorem 2.1. Since $f_i \in C^\infty$, Hörmander's theorem together with Proposition 3.1 say that $\phi_i, \phi_{i,\epsilon} \in C^\infty$ and so $\phi_i, \phi_{i,\epsilon}$ solves $-\mathcal{L}\phi_i = -\mathcal{L}_\epsilon\phi_{i,\epsilon} = f_i - \pi(f_i)$ classically. Furthermore, in the same way as in the proof of Lemma 3.2 to obtain (6.15), it holds that

$$\|\nabla_p\phi_i - \nabla_p\phi\|_{L^2(\tilde{\pi})} + \|\nabla_p\phi_{i,\epsilon} - \nabla_p(\phi + \delta\phi_\epsilon)\|_{L^2(\tilde{\pi})} \to 0 \quad \text{as } i \to \infty, \tag{6.19}$$

where $\nabla_p\phi_i, \nabla_p\phi_{i,\epsilon} \in L^2(\tilde{\pi})$ by Lemma 3.2 itself. The term under the limit in (6.16) is now approximated with a term involving f_i and the truncation functions η_k from Lemma 6.1. Working now with the approximating integral

and using Lemma 3.3 together with the obvious extension on the notation from (1.7),

$$\begin{aligned}\int \eta_k(\phi_{i,\epsilon}-\phi_i)f_i d\tilde{\pi} &= -\int \eta_k(\phi_{i,\epsilon}-\phi_i)\mathcal{L}_\epsilon^*\tilde{\phi}_{i,\epsilon}d\tilde{\pi}\\ &= -\int (\mathcal{L}_\epsilon\eta_k(\phi_{i,\epsilon}-\phi_i)+2\nabla_p\eta_k(\Gamma+\epsilon\delta\Gamma)\nabla_p(\phi_{i,\epsilon}-\phi_i)\\ &\qquad -\epsilon\eta_k\mathcal{S}\phi_i)\tilde{\phi}_{i,\epsilon}d\tilde{\pi}\\ &= -\int(\mathcal{L}_\epsilon\eta_k(\phi_{i,\epsilon}-\phi_i)+2\nabla_p\eta_k(\Gamma+\epsilon\delta\Gamma)\nabla_p(\phi_{i,\epsilon}-\phi_i))\tilde{\phi}_{i,\epsilon}d\tilde{\pi}\\ &\quad -\int \epsilon(\eta_k\nabla_p\phi_i^\top\delta\Gamma\nabla_p\tilde{\phi}_{i,\epsilon}+\tilde{\phi}_{i,\epsilon}\nabla_p\phi_i^\top\delta\Gamma\nabla_p\eta_k)d\tilde{\pi}. \end{aligned} \tag{6.20}$$

By Lemma 6.1 and 3.2, the terms involving gradients of η_k converge to zero as $k\to\infty$, so that taking $k\to\infty$, then $i\to\infty$,

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

holds, where (6.17), (6.18) and (6.19) have been used. Plugging into (6.16), the directional derivative becomes

$$\frac{1}{2}d\sigma^2.\delta\Gamma = -\lim_{\epsilon\to 0}\int \nabla_p\phi^\top\delta\Gamma\nabla_p(\tilde{\phi}+\delta\tilde{\phi}_\epsilon)d\tilde{\pi}. \tag{6.21}$$

From here, for any $\epsilon>0$, the unwanted term under the limit can be controlled by approximating again with a truncation and f_i, $i\in\mathbb{N}$,

$$\begin{aligned}\lambda_m\int \eta_k\Big|\nabla_p(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)\Big|^2 d\tilde{\pi}&\\ &\le \int \eta_k\nabla_p(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)^\top(\Gamma+\epsilon\delta\Gamma)\nabla_p(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)d\tilde{\pi}\\ &= -\int(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)\nabla_p\eta_k^\top(\Gamma+\epsilon\delta\Gamma)\nabla_p(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)d\tilde{\pi}\\ &\quad -\int \eta_k(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)(-M^{-1}p+\nabla_p)^\top((\Gamma+\epsilon\delta\Gamma)\nabla_p(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)d\tilde{\pi}, \end{aligned} \tag{6.22}$$

where

$$\lambda_m = \inf_{0<\epsilon\le\epsilon'}\lambda_m^\epsilon$$

and λ_m^ϵ is the smallest eigenvalue of $\Gamma+\epsilon\delta\Gamma$. The first term on the right hand side is negligible as $k\to\infty$ because of Lemmata 6.1 and 3.2. The remaining term is

$$\begin{aligned}-\int \eta_k(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)(-M^{-1}p+\nabla_p)^\top((\Gamma+\epsilon\delta\Gamma)\nabla_p(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)d\tilde{\pi}&\\ &= -\int \eta_k(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)\mathcal{L}_\epsilon^*(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)d\tilde{\pi}\\ &\quad -\int \eta_k(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)(p^\top M^{-1}\nabla_p-\nabla U^\top\nabla_p)(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)d\tilde{\pi}, \end{aligned} \tag{6.23}$$

where the last term, after integrating by parts, gives

$$\begin{aligned}\int \eta_k(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)(p^\top M^{-1}\nabla_p-\nabla U^\top\nabla_p)(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)d\tilde{\pi}&\\ &= -\frac{1}{2}\int(p^\top M^{-1}\nabla_q-\nabla U^\top\nabla_p)\eta_k(\tilde{\phi}_{i,\epsilon}-\tilde{\phi}_i)^2 d\tilde{\pi}, \end{aligned} \tag{6.24}$$

which is negligible as $k \to \infty$ again due to Lemma 6.1 and (2.2). On the other hand, the first term on the right hand side of (6.23) is

$$\begin{aligned} -\int \eta_k(\tilde{\phi}_{i,\epsilon} - \tilde{\phi}_i)\mathcal{L}_\epsilon^*(\tilde{\phi}_{i,\epsilon} - \tilde{\phi}_i)d\tilde{\pi} &= \epsilon \int \eta_k(\tilde{\phi}_{i,\epsilon} - \tilde{\phi}_i)(M^{-1}p - \nabla_p)^\top \delta\Gamma \nabla_p \tilde{\phi}_i d\tilde{\pi} \\ &= \epsilon \int \nabla_p \eta_k(\tilde{\phi}_{i,\epsilon} - \tilde{\phi}_i)^\top \delta\Gamma \nabla_p \tilde{\phi}_i d\tilde{\pi} \\ &\quad + \epsilon \int \eta_k \nabla_p(\tilde{\phi}_{i,\epsilon} - \tilde{\phi}_i)^\top \delta\Gamma \nabla_p \tilde{\phi}_i d\tilde{\pi}, \end{aligned} \tag{6.25}$$

where again the term involving $\nabla_p \eta_k$ is negligible as $k \to \infty$, so that putting together (6.22), (6.23), (6.24) and (6.25), then taking $k \to \infty$ and $i \to \infty$ with (6.19) gives

$$\begin{aligned} \lambda_m \int \left|\nabla_p \delta\tilde{\phi}_\epsilon\right|^2 d\tilde{\pi} &\le \epsilon \int \nabla_p \delta\tilde{\phi}_\epsilon^\top \delta\Gamma \nabla_p \tilde{\phi} d\tilde{\pi} \\ &\le \frac{\epsilon\lambda_M}{\epsilon'} \int \left(\left|\nabla_p \delta\tilde{\phi}_\epsilon\right|^2 + \left|\nabla_p \tilde{\phi}\right|^2 \right) d\tilde{\pi}, \end{aligned}$$

where $\lambda_M = \inf_{0<\epsilon\le\epsilon'} \lambda_M^\epsilon$, λ_M^ϵ is the largest eigenvalue of $\Gamma + \epsilon\delta\Gamma$ and $\delta\Gamma = \frac{\Gamma + \epsilon'\delta\Gamma - \Gamma}{\epsilon'}$ has been used. Therefore

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

holds for small enough ϵ and putting into (6.21) concludes the proof.

□

For the proof of Proposition 4.9, some notation is introduced. For $\tilde{k} \in \mathbb{N}$, let the tridiagonal matrix $M_{\tilde{k}} \in \mathbb{R}^{\tilde{k}+1 \times \tilde{k}+1}$ be given by its elements

$$(M_{\tilde{k}})_{i,j} = \begin{cases} i & \text{if } i+1 = j, \\ (i-1)\gamma & \text{if } i = j, \\ i - \tilde{k} - 2 & \text{if } i-1 = j, \\ 0 & \text{otherwise} \end{cases} \tag{6.26}$$

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

Lemma 6.2. *Let $m \in \mathbb{N}$. Any tridiagonal matrix $\tilde{M} \in \mathbb{R}^{m \times m}$ of the form*

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

for constants $b_i, b_i', b_i'' \in \mathbb{R}$, has an order γ determinant as $\gamma \to 0$ if m is odd and a determinant that is bounded away from zero as $\gamma \to 0$ if m is even.

Lemma 6.2 is straightforwardly proved by repeatedly taking Laplace expansions. An explicit proof is not given here.

Proof. (of Proposition 4.9) Only a standard Gaussian and $M = 1$ is considered, the arguments for the general centered Gaussian case are the same. First consider the observable

$$f(q) = q^k \tag{6.27}$$

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

$$\phi(q,p) = \sum_{i,j=0}^{k} a_{i,j} q^i p^j \tag{6.28}$$

for $a_{i,j} \in \mathbb{R}$ and $\Gamma = \gamma > 0$. It will be shown that arbitrarily small asymptotic variance is achieved in the $\gamma \to 0$ 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,

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

where

$$a_{i,j} = 0 \quad \forall i,j < 0 \text{ and } \forall i,j > k. \tag{6.29}$$

Comparing coefficients in (1.5),

$$-(i+1) a_{i+1,j-1} + (j+1) a_{i-1,j+1} + \gamma j a_{i,j} - \gamma (j+2)(j+1) a_{i,j+2} = 0 \tag{6.30}$$

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

$$a_{i'+j',k+1-j'} = 0 \quad \forall i', j' \geq 0 \tag{6.31}$$

because of the following. The base case $j' = 0$ follows by (6.29), the induction step follows by taking $(i,j) = (i'+j'-1, k+2-j')$ for $i' \geq 0$ in (6.30) and again using (6.29) where necessary. Comparing coefficients in the Poisson equation (1.5) for $(i,j) = (k,0)$ and using (6.29), (6.31) yields[4]

$$a_{k-1,1} = 1. \tag{6.32}$$

Combining (6.32) with setting $(i,j) = (j'-1, k+1-j')$ for $j' = 1, \ldots, k$ in (6.30), the entries $a_{j',k-j'}$ satisfy the linear system

$$M_k (a_{k,0}, a_{k-1,1}, \ldots, a_{0,k})^\top = (1, 0, \ldots, 0)^\top, \tag{6.33}$$

where $M_k \in \mathbb{R}^{k+1 \times k+1}$ is the tridiagonal matrix given in (6.26). In order to find the order in γ as $\gamma \to 0$ of the elements of $(a_{k,0}, \ldots, a_{0,k})^\top$ appearing in (6.33), 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 ith 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 6.2 for this $(k+1-i)$-by-$(k+1-i)$ matrix gives that C_i is of order γ as $\gamma \to 0$ for odd i. Furthermore, the determinant of M_k is bounded away from zero as $\gamma \to 0$ by Lemma 6.2. Therefore the elements of $(a_{k,0}, \ldots, a_{0,k})$ in the left hand side of (6.33) with an odd index, that is $a_{k-j,j}$ for even j, have order γ and at most order 1 otherwise as $\gamma \to 0$. These elements with odd indices are exactly those from the vector $(a_{k,0}, \ldots, a_{0,k})^\top$ that make a contribution to the asymptotic variance. The 'next' set of contributions come from the vector $(a_{k-2,0}, a_{k-3,1} \ldots, a_{0,k-2})$. Using again (6.29) and (6.30), the vector satisfies

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

for some vector v_{k-2} (from the last term on the left hand side of (6.30)) of order γ as $\gamma \to 0$ and since the determinant of M_{k-2} is of order 1 (by Lemma 6.2), the contributions here to the asymptotic variance are again of

[4] It is illustrative to imagine a grid of coefficients and the relations (6.30) and (6.32) as L-shaped chains on the grid, where (6.31) and (6.29) leave only a triangular area of nonzero coefficients.

order γ. Continuing for $(a_{k-2j,0}, a_{k-2j-1,1} \ldots, a_{0,k-2j})^\top$, $j \in \mathbb{N}$, it follows that all contributions are of order γ as $\gamma \to 0$. 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 (4.24), 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 (4.24), the proof concludes. □

Proof. (of Proposition 4.10) Take the polynomial ansatz

$$\phi(q,p) = \sum_{i,j=0}^{4} a_{i,j} q^i p^j \tag{6.34}$$

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 (1.5) and using the same strong induction argument as in the proof of Proposition 4.9 leads to (6.30) for all $(i,j) \neq (4,0), (0,0)$ and equation (6.31). Taking $(i,j) = (j'-1, 5-j')$ for $1 \leq j' \leq 4$ in (6.30) and comparing the q^4 coefficients in the Poisson equation, it holds that

$$M_4(a_{4,0}, a_{3,1}, a_{2,2}, a_{1,3}, a_{0,4})^\top = (1, 0, \ldots, 0)^\top \tag{6.35}$$

and taking $(i,j) = (j'-1, 3-j')$ for $j' \geq 1$ in (6.30) yields

$$M_2(a_{2,0}, a_{1,1}, a_{0,2})^\top = \gamma(2a_{2,2}, 6a_{1,3}, 12a_{0,4})^\top. \tag{6.36}$$

Equations (6.35), (6.36) can be solved explicitly and the asymptotic variance is a weighted sum of the resulting coefficients. Those in (6.34) 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 $\gamma \to 0$ or $\gamma \to \infty$. Comparing constant terms in the Poisson equation yields

$$a_{0,2} = \frac{1}{2\Gamma} \int q^4 \frac{e^{\frac{q^2}{2}}}{\sqrt{2\pi}} dq = \frac{3}{2\Gamma},$$

which turns out to be satisfied by the solution for $a_{0,2}$, so that (6.34) is indeed a solution; note that the coefficients associated to M_3 and M_1 are zero by a similar procedure as above. □

7 Discussion

7.1 Relation to previous methodologies

The infinite time integral (1.9) has been used for the calculation of transport coefficients in molecular dynamics [52, 65] and the derivative of the expectation appearing in (1.9) with respect to initial conditions is a problem considered when calculating the 'greeks' in mathematical finance [33]. On the topic of the latter and in contrast to [33], there is previous work dealing with cases of degenerate noise in the system, but the formulae derived were done so under different motivations and do not seem to improve upon (1.10) in our situation; some of these references are given in Remark 5.2.
Taking $\Gamma \to \infty$ together with a time rescaling, the dynamics (1.1) become the overdamped Langevin equation [66]. An analogous result holds [46] when $\Gamma = \Gamma(q)$ is position dependent, where a preconditioner for the corresponding overdamped dynamics appears in terms of Γ^{-1}; see Section 7.3 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 [1].
On the application of underdamped Langevin dynamics with (variance reduced) stochastic gradients alongside the related Hamiltonian Monte Carlo method, [85] presents a comparison with convergence rates for the latter. In [17], convergence guarantees are provided for variance reduced gradients in the overdamped case and the control variate

stochastic gradients in the underdamped case, 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 [22].
Variance reduction by modifying the observable instead of changing the dynamics has been considered for example in [3, 5, 77]. The methods there are incompatible with the framework in the present work due to the improved observable being unknown before the simulation of the Markov chain. Although useful, their applicability are limited in large n cases due to storage requirements[3], not to mention either escalating computational cost for improvements in the observable or requirement of a priori knowledge[77].

7.2 The nonconvex case

In the case where U is nonconvex, the Monte Carlo procedure in Section 5.1.3 may continue to be used as presented, however the first variation process could easily stray from the case of exponential decay as in Theorem 3.5. 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 occuring (as if the case of Corollary 4.8), 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, the Galerkin method presented in Appendix B tends to give good convergence for Γ in such cases.

7.3 Position-dependent friction

It is possible to adapt the formula (3.2) to the case of position-dependent gradient direction in Γ given a Feynman-Kac representation formula and the corresponding existence result, which will be the aim of future work. The gradient direction is the same as (1.8) 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 5.1.3 is not as well justified, or rather, does not happen in the execution of Algorithm 2 due to (5.11) not being satisfied.

7.4 Metropolisation

Throughout Section 5, the implementation has not involved accept-reject steps. Metropolisation of discretisations of the underdamped Langevin dynamics was given in [45], see also Section 2.2.3.2 in [53] and [58, 74]. The systematic discretisation error is removed with the inclusion of this step but the momentum is reversed upon rejection (to avoid high rejection rates [74]), 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.

7.5 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 Corollary 4.8). This is consistent with the numerical conclusion from the proposed 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 (1.6) and (1.10) can be used to guide Γ in arbitrarily high dimension

by extrapolation; that is, 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 that can be extrapolated to the original problem. In particular, for the Bayesian inference problem as formulated in Section 5.2.3, 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 includes well-posedness of the optimisation in Γ, extension to higher-order Langevin samplers methods as in [16, 59] and gradient formulae in the discrete time case analogous to Theorem 3.4.

Acknowledgements

M.C. was funded under a EPSRC studentship. G.A.P. was partially supported by the EPSRC through grants EP/P031587/1, EP/L024926/1, and EP/L020564/1. N.K. and G.A.P. were funded in part by JPMorgan Chase & Co under a J.P. Morgan A.I. Research Awards 2019. Any views or opinions expressed herein are solely those of the authors listed, and may differ from the views and opinions expressed by JPMorgan Chase & Co. or its affiliates. This material is not a product of the Research Department of J.P. Morgan Securities LLC. This material does not constitute a solicitation or offer in any jurisdiction. 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] H. AlRachid, L. Mones, and C. Ortner. Some remarks on preconditioning molecular dynamics. *SMAI J. Comput. Math.*, 4:57–80, 2018.

[2] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Stat. Comput.*, 18(4):343–373, 2008.

[3] J. Baker, P. Fearnhead, E. B. Fox, and C. Nemeth. Control variates for stochastic gradient MCMC. *Stat. Comput.*, 29(3):599–615, 2019.

[4] R. Bardenet, A. Doucet, and C. Holmes. On Markov chain Monte Carlo methods for tall data. *J. Mach. Learn. Res.*, 18:Paper No. 47, 43, 2017.

[5] D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov, and S. Samsonov. Variance reduction for Markov chains with application to MCMC. *Stat. Comput.*, 30(4):973–997, 2020.

[6] A. Beskos, G. Roberts, A. Stuart, and J. Voss. MCMC methods for diffusion bridges. *Stoch. Dyn.*, 8(3):319–350, 2008.

[7] A. Beskos and A. Stuart. MCMC methods for sampling function space. In *ICIAM 07—6th International Congress on Industrial and Applied Mathematics*, pages 337–364. Eur. Math. Soc., Zürich, 2009.

[8] M. Betancourt, S. Byrne, S. Livingstone, and M. Girolami. The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli*, 23(4A):2257–2298, 2017.

[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(2):185–201, 1982.

[10] J. Bierkens, P. Fearnhead, and G. Roberts. The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *Ann. Statist.*, 47(3):1288–1320, 2019.

[11] 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(5):867–884, 2010.

[12] N. Bou-Rabee and E. Vanden-Eijnden. Pathwise accuracy and ergodicity of metropolized integrators for SDEs. *Comm. Pure Appl. Math.*, 63(5):655–696, 2010.

[13] A. Bouchard-Côté, S. J. Vollmer, and A. Doucet. The bouncy particle sampler: a nonreversible rejection-free Markov chain Monte Carlo method. *J. Amer. Statist. Assoc.*, 113(522):855–867, 2018.

[14] G. Bussi and M. Parrinello. Accurate sampling using Langevin dynamics. *Phys. Rev. E*, 75:056707, May 2007.

[15] 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(2):337–382, 2012.

[16] M. Chak, N. Kantas, and G. A. Pavliotis. On the Generalised Langevin Equation for Simulated Annealing, 2021. arXiv: 2003.06448.

[17] N. S. Chatterji, N. Flammarion, Y. Ma, Y. Ma, P. Bartlett, and M. I. Jordan. On the theory of variance reduction for stochastic gradient monte carlo. *ArXiv*, abs/1802.05431, 2018.

[18] X. Chen, S. Liu, R. Sun, and M. Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. 2019. 7th International Conference on Learning Representations, ICLR 2019 ; Conference date: 06-05-2019 Through 09-05-2019.

[19] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR, 06–09 Jul 2018.

[20] A. S. Dalalyan and L. Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

[21] B. Delyon and Y. Hu. Simulation of conditioned diffusion and application to parameter estimation. *Stochastic Process. Appl.*, 116(11):1660–1675, 2006.

[22] Z. Ding, Q. Li, J. Lu, and S. J. Wright. Random Coordinate Underdamped Langevin Monte Carlo, 2020. arXiv: 2010.11366.

[23] Z. Dong and X. Peng. Malliavin matrix of degenerate SDE and gradient estimate. *Electron. J. Probab.*, 19:no. 73, 26, 2014.

[24] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195(2):216–222, 1987.

[25] C. DuBois, A. Korattikara, M. Welling, and P. Smyth. Approximate slice sampling for bayesian posterior inference. In *Artificial Intelligence and Statistics*, pages 185–193. PMLR, 2014.

[26] A. B. Duncan, T. Lelièvre, and G. A. Pavliotis. Variance reduction using nonreversible Langevin samplers. *J. Stat. Phys.*, 163(3):457–491, 2016.

[27] A. B. Duncan, N. Nüsken, and G. A. Pavliotis. Using perturbed underdamped Langevin dynamics to efficiently sample from probability distributions. *J. Stat. Phys.*, 169(6):1098–1131, 2017.

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

[29] A. Durmus and E. Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm, 2018. arXiv: 1605.01559.

[30] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *J. Mach. Learn. Res.*, 20:Paper No. 183, 42, 2019.

[31] J. C. M. Fok, B. Guo, and T. Tang. Combined Hermite spectral-finite difference method for the Fokker-Planck equation. *Math. Comp.*, 71(240):1497–1528, 2002.

[32] J. Foster, T. Lyons, and H. Oberhauser. The shifted ODE method for underdamped Langevin MCMC. *ArXiv*, abs/2101.03446, 2021.

[33] E. Fournié, J.-M. Lasry, J. Lebuchoux, P.-L. Lions, and N. Touzi. Applications of Malliavin calculus to Monte Carlo methods in finance. *Finance Stoch.*, 3(4):391–412, 1999.

[34] A. Friedman. *Stochastic differential equations and applications. Vol. 1.* Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Probability and Mathematical Statistics, Vol. 28.

[35] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European Control Conference (ECC)*, pages 310–315, 2015.

[36] A. Guillin and F.-Y. Wang. Degenerate Fokker-Planck equations: Bismut formula, gradient estimate and Harnack inequality. *J. Differential Equations*, 253(1):20–40, 2012.

[37] B.-Y. Guo. *Spectral methods and their applications.* World Scientific Publishing Co., Inc., River Edge, NJ, 1998.

[38] M. Hairer, A. M. Stuart, and J. Voss. Analysis of SPDEs arising in path sampling. II. The nonlinear case. *Ann. Appl. Probab.*, 17(5-6):1657–1706, 2007.

[39] M. Hairer, A. M. Stuart, and J. Voss. Sampling conditioned hypoelliptic diffusions. *Ann. Appl. Probab.*, 21(2):669–698, 2011.

[40] 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(4):587–603, 2005.

[41] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[42] Y. He, K. Balasubramanian, and M. A. Erdogdu. On the Ergodicity, Bias and Asymptotic Normality of Randomized Midpoint Sampling Method, 2020. arXiv: 2011.03176.

[43] L. Hörmander. Hypoelliptic second order differential equations. *Acta Math.*, 119:147–171, 1967.

[44] A. M. Horowitz. The second order Langevin equation and numerical simulations. *Nuclear Physics B*, 280:510–522, 1987.

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

[46] 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(3):1259–1283, 2015.

[47] A. Kavalur, V. Guduguntla, and W. K. Kim. Effects of Langevin friction and time steps in the molecular dynamics simulation of nanoindentation. *Molecular Simulation*, 46(12):911–922, 2020.

[48] T. Komorowski, C. Landim, and S. Olla. *Fluctuations in Markov processes*, volume 345 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Heidelberg, 2012. Time symmetry and martingale approximation.

[49] N. V. Krylov, M. Röckner, and J. Zabczyk. *Stochastic PDE's and Kolmogorov equations in infinite dimensions*, volume 1715 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin; Centro Internazionale Matematico Estivo (C.I.M.E.), Florence, 1999. Lectures given at the 2nd C.I.M.E. Session held in Cetraro, August 24–September 1, 1998, Edited by G. Da Prato, Fondazione CIME/CIME Foundation Subseries.

[50] B. Leimkuhler and C. Matthews. Rational construction of stochastic numerical methods for molecular sampling. *Appl. Math. Res. Express. AMRX*, (1):34–56, 2013.

[51] B. Leimkuhler and C. Matthews. Robust and efficient configurational molecular sampling via langevin dynamics. *The Journal of Chemical Physics*, 138(17):174102, 2013.

[52] B. Leimkuhler, C. Matthews, and G. Stoltz. The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics. *IMA J. Numer. Anal.*, 36(1):13–79, 2016.

[53] T. Lelièvre, M. Rousset, and G. Stoltz. *Free energy computations*. Imperial College Press, London, 2010. A mathematical perspective.

[54] T. Lelièvre and G. Stoltz. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numer.*, 25:681–880, 2016.

[55] Y.-A. Ma, N. S. Chatterji, X. Cheng, N. Flammarion, P. L. Bartlett, and M. I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3):1942–1992, 2021.

[56] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[57] P. Monmarché. Almost sure contraction for diffusions on $\setminus^d$. Application to generalised Langevin diffusions, 2021. arXiv: 2009.10828.

[58] P. Monmarché. High-dimensional MCMC with a standard splitting scheme for the underdamped Langevin diffusion. *Electron. J. Stat.*, 15(2):4117–4166, 2021.

[59] 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:Paper No. 42, 41, 2021.

[60] R. Neal. Bayesian learning via stochastic dynamics. In S. Hanson, J. Cowan, and C. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5. Morgan-Kaufmann, 1993.

[61] R. M. Neal. Slice sampling. *Ann. Statist.*, 31(3):705–767, 2003. With discussions and a rejoinder by the author.

[62] R. M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 113–162. CRC Press, Boca Raton, FL, 2011.

[63] C. Nemeth and P. Fearnhead. Stochastic Gradient Markov Chain Monte Carlo. *J. Amer. Statist. Assoc.*, 116(533):433–450, 2021.

[64] A. B. Owen. Statistically efficient thinning of a Markov chain sampler. *J. Comput. Graph. Statist.*, 26(3):738–744, 2017.

[65] G. A. Pavliotis. Asymptotic analysis of the Green-Kubo formula. *IMA J. Appl. Math.*, 75(6):951–967, 2010.

[66] G. A. Pavliotis. *Stochastic processes and applications*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, 2014. Diffusion processes, the Fokker-Planck and Langevin equations.

[67] B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[68] C. Prévôt and M. Röckner. *A concise course on stochastic partial differential equations*, volume 1905 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

[69] P. E. Protter. *Stochastic integration and differential equations*, volume 21 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, second edition, 2004. Stochastic Modelling and Applied Probability.

[70] S. Redon, G. Stoltz, and Z. Trstanova. Error analysis of modified Langevin dynamics. *J. Stat. Phys.*, 164(4):735–771, 2016.

[71] J. Roussel and G. Stoltz. Spectral methods for Langevin dynamics and associated error estimates. *ESAIM Math. Model. Numer. Anal.*, 52(3):1051–1083, 2018.

[72] M. Sachs, B. Leimkuhler, and V. Danos. Langevin dynamics with variable coefficients and nonconservative forces: From stationary states to numerical methods. *Entropy*, 19(12), 2017.

[73] J. M. Sanz-Serna and K. Zygalakis. Wasserstein distance estimates for the distributions of numerical approximations to ergodic stochastic differential equations. *ArXiv*, abs/2104.12384, 2021.

[74] A. Scemama, T. Lelièvre, G. Stoltz, E. Cancès, and M. Caffarel. An efficient sampling algorithm for variational monte carlo. *The Journal of Chemical Physics*, 125(11):114105, 2006.

[75] R. Shen and Y. T. Lee. The randomized midpoint method for log-concave sampling. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[76] R. D. Skeel and C. Hartmann. Choice of Damping Coefficient in Langevin Dynamics, 2021. arXiv: 2106.11597.

[77] L. F. South, C. J. Oates, A. Mira, and C. Drovandi. Regularised Zero-Variance Control Variates for High-Dimensional Variance Reduction, 2020. arXiv: 1811.05073.

[78] G. Stoltz and Z. Trstanova. Langevin dynamics with general kinetic energies. *Multiscale Model. Simul.*, 16(2):777–806, 2018.

[79] J. Teichmann. Calculating the Greeks by cubature formulae. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 462(2066):647–670, 2006.

[80] Z. Trstanova and S. Redon. Estimating the speed-up of adaptively restrained Langevin dynamics. *J. Comput. Phys.*, 336:412–428, 2017.

[81] P. Vanetti, A. Bouchard-Côté, G. Deligiannidis, and A. Doucet. Piecewise-Deterministic Markov Chain Monte Carlo, 2018. arXiv: 1707.05296.

[82] S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh. Exploration of the (non-)asymptotic bias and variance of stochastic gradient Langevin dynamics. *J. Mach. Learn. Res.*, 17:Paper No. 159, 45, 2016.

[83] M. Welling and Y. W. Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 681–688, Madison, WI, USA, 2011. Omnipress.

[84] J. Yang, G. O. Roberts, and J. S. Rosenthal. Optimal scaling of random-walk metropolis algorithms on general target distributions. *Stochastic Process. Appl.*, 130(10):6094–6132, 2020.

[85] D. Zou and Q. Gu. On the Convergence of Hamiltonian Monte Carlo with Stochastic Gradients. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 13012–13022. PMLR, 18–24 Jul 2021.

Appendix A Preliminaries

Theorem A.1. *Let Assumption 1 hold. For any $\mathcal{F}_0$-measurable $z_0 : \Omega \to \mathbb{R}^{2n}$, there exists a unique almost surely continuous in t solution $(q_t, p_t) = z_t : \Omega \to \mathbb{R}^{2n}$ to* (1.1) *that is $\mathcal{F}_t$-adapted and satisfies*

$$\mathbb{E}|z_t|^2 \leq e^{\kappa t}(1 + \mathbb{E}|z_0|^2) \tag{A.1}$$

for a constant $\kappa \in \mathbb{R}$ and all $t \geq 0$. Furthermore, for any $z \in \mathbb{R}^{2n}$, $t \geq 0$, let p_t^z be the probability measure given by

$$p_t^z(A) = \mathbb{P}(z_t^z \in A) \tag{A.2}$$

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

1. *is a transition probability in the sense that*
   (a) $(t,z) \mapsto p_t^z(A)$ *is Borel measurable on* $(0,\infty) \times \mathbb{R}^{2n}$,
   (b) *the Chapman-Kolmogorov relation [34] holds and*
2. *admits a density denoted* $p(z,\cdot,t) : \mathbb{R}^{2n} \to \mathbb{R}$ *with respect to the Lebesgue measure on* $\mathbb{R}^{2n}$ *at every* $(t,z) \in (0,\infty) \times \mathbb{R}^{2n}$ *such that* p *is a measurable function satisfying for every* $z \in \mathbb{R}^{2n}$,

$$p(z,\cdot,\cdot) \in C^\infty(\mathbb{R}^{2n} \times (0,\infty)). \tag{A.3}$$

Proof. Theorem 3.1.1 in [68] yields existence and uniqueness of the solution to (1.1) together with (A.1). Theorem 3.1 and 3.6 in Section 5 of [34] give that $p_t^z(A)$ given by (A.2) is a probability kernel, that is, $p_t^z(A)$ 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 $(t,z) \mapsto p_t^z(A)$ for fixed A, consider $\hat{z}_t^z$ given by

$$\hat{z}_t^z(\omega) = \begin{cases} z_t^z(\omega) & \text{if } \omega : z_\bullet^z(\omega) \in C([0,\infty)), \\ 0 & \text{otherwise.} \end{cases} \tag{A.4}$$

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, p_t^z admits a density at every $(t,z) \in (0,\infty) \times \mathbb{R}^{2n}$ satisfying (A.3) due to Itô's rule and Hörmander's theorem [43]; measurability with respect to the starting point z and therefore jointly in all of the arguments follows by the strong Feller property given by Theorem 4.2 in [23], because $p(t,\cdot,\zeta)$ is the pointwise limit of the continuous functions $(\int \eta_k(\zeta - \zeta')p(t,\cdot,\zeta')d\zeta')_{k>0}$, where η_k denotes the standard scaled mollifiers. □

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

$$P_t(f) : z \mapsto \mathbb{E}(f(z_t^z)) = \mathbb{E}(f(z_t)|z_0 = z). \tag{A.5}$$

The family $(P_t)_{t\geq 0}$ forms a strongly continuous (Proposition 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

$$Lu = \lim_{t\to 0} \frac{P_t(u) - u}{t} \tag{A.6}$$

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 $\|f - g\|_{L^2(\tilde{\pi})} \leq \frac{\epsilon}{3}$. Writing

$$\|P_t f - f\|_{L^2(\tilde{\pi})} \leq \|P_t f - P_t g\|_{L^2(\tilde{\pi})} + \|f - g\|_{L^2(\tilde{\pi})} + \|P_t g - g\|_{L^2(\tilde{\pi})},$$

the right hand side is bounded by ϵ after Jensen's inequality, invariance of $\tilde{\pi}$ and Itô's rule together with a small enough t. □

Appendix B Solving the Poisson equation with a Galerkin method

Throughout this appendix, $M = I_n$ is assumed. In low dimensions, it is feasible to approximate $\nabla_p \phi$ and a change in Γ using Hermite polynomials. This approach gives an approximation in a finite subspace of $L^2(\tilde{\pi})$ at the level of ϕ, as opposed to estimates of $\nabla_p \phi$ at particular points in space as in the Monte Carlo approach in Section 5. Specifically, the polynomials given by

$$H_l(z) = \frac{(-1)^l}{\sqrt{l!}} e^{\frac{z^2}{2}} \frac{d^l}{dz^l}\left(e^{-\frac{z^2}{2}}\right)$$

for $l \in \mathbb{N}$ and their products in the multidimensional case

$$H_{\underline{l}}(p) = \prod_{k=1}^{n} H_{\underline{l}_k}(p_k), \qquad p = (p_1, \dots, p_n) \in \mathbb{R}^n$$

for multiindices $\underline{l} = (\underline{l}_1, \dots, \underline{l}_n) \in \mathbb{N}^n$ are considered in the weighted $L^2(\omega)$ space, where $\omega(p) = \frac{e^{-\frac{1}{2}|p|^2}}{(2\pi)^{-\frac{n}{2}}}$. A property of the Hermite polynomials that is repeatedly used here is that

$$\partial_z H_l(z) = \sqrt{l} H_{l-1}(z).$$

For the application of Hermite polynomials in solving the Poisson equation associated to Langevin dynamics (in the case of scalar friction), we refer to [71]. See also Chapter 5 in [37] for Hermite polynomials in the multidimensional setting. In the case of a non-quadratic potential U, the same polynomials are used here after a Gram-Schmidt procedure in $L^2(\pi)$, which are denoted $(\hat{H}_{\underline{l}})_{\underline{l} \in \mathbb{N}^n}$, so that

$$\hat{H}_{\underline{l}} = \sum_{|\underline{k}|_\infty \leq K} \alpha_{\underline{l}}^{\underline{k}} H_{\underline{k}},$$

where $|\underline{k}|_\infty = \max(\underline{k}_1, \dots, \underline{k}_n)$, $K \in \mathbb{N}$, for some constants $\alpha_{\underline{k}}^{\underline{l}} \in \mathbb{R}$ calculated numerically. Their products with $H_{\underline{l}}$ are considered on $L^2(\tilde{\pi})$. Similarly, Fourier approximations can be used in the case of an n-torus (in q).
The observable $f \in L^2_0(\pi)$ is approximated by the projection defined by

$$\Pi_K^q f := \sum_{|\underline{l}|_\infty \leq K} \hat{H}_{\underline{l}} \int f \hat{H}_{\underline{l}} d\pi = \sum_{|\underline{k}|_\infty, |\underline{l}|_\infty \leq K} \hat{H}_{\underline{l}} \alpha_{\underline{l}}^{\underline{k}} \int f H_{\underline{k}} d\pi. \tag{B.1}$$

Since the generator has the form

$$\mathcal{L} = \nabla_p^* \cdot \nabla_q - \nabla_q^* \cdot \nabla_p - (\nabla_p^*)^\top \Gamma \nabla_p,$$

where

$$\nabla_q^* = -\nabla_q + \nabla U, \qquad \nabla_p^* = -\nabla_p + p$$

are the respective formal $L^2(\tilde{\pi})$-adjoints of ∇_q and ∇_p, the negative of the generator in the Poisson equation applied on functions of the form (B.1) is the $(K+1)^{2n}$-by-$(K+1)^{2n}$ matrix given by

$$\begin{aligned}
L_{\underline{k},\underline{l},\hat{\underline{k}},\hat{\underline{l}}} &= \langle \hat{H}_{\underline{k}} H_{\underline{l}}, -\mathcal{L}(\hat{H}_{\hat{\underline{k}}} H_{\hat{\underline{l}}}) \rangle_{\tilde{\pi}} \\
&= -\langle \hat{H}_{\underline{k}} \nabla_p H_{\underline{l}}, \nabla_q \hat{H}_{\hat{\underline{k}}} H_{\hat{\underline{l}}} \rangle_{\tilde{\pi}} + \langle \nabla_q \hat{H}_{\underline{k}} H_{\underline{l}}, \hat{H}_{\hat{\underline{k}}} \nabla_p H_{\hat{\underline{l}}} \rangle_{\tilde{\pi}} + \langle \hat{H}_{\underline{k}} \nabla_p H_{\underline{l}}, \Gamma \hat{H}_{\hat{\underline{k}}} \nabla_p H_{\hat{\underline{l}}} \rangle_{\tilde{\pi}} \\
&= -\sum_i \langle \hat{H}_{\underline{k}}, \partial_{q_i} \hat{H}_{\hat{\underline{k}}} \rangle_\pi (\sqrt{\underline{l}_i} \delta_{\hat{\underline{l}}}^{\underline{l}-e_i}) + \sum_i \langle \partial_{q_i} \hat{H}_{\underline{k}}, \hat{H}_{\hat{\underline{k}}} \rangle_\pi (\sqrt{\hat{\underline{l}}_i} \delta_{\hat{\underline{l}}-e_i}^{\underline{l}}) \\
&\quad + \sum_{i,j} \delta_{\hat{\underline{k}}}^{\underline{k}} \delta_{\hat{\underline{l}}-e_i}^{\underline{l}-e_j} \sqrt{\underline{l}_j \hat{\underline{l}}_i} \Gamma_{i,j}
\end{aligned} \tag{B.2}$$

where δ denotes the Kronecker delta here, the dependences of $\hat{H}_{\underline{k}}$, $\hat{H}_{\hat{\underline{k}}}$ and $H_{\underline{l}}$, $H_{\hat{\underline{l}}}$ on q and p respectively have been suppressed, $\langle v, w \rangle$ denotes $\sum_i \langle v_i, w_i \rangle$ for $v = (v_1, \dots, v_n)$, $w = (w_1, \dots, w_n)$ and $\langle \cdot, \cdot \rangle$ denotes the inner product on $L^2(\tilde{\pi})$. Note further that

$$\langle \partial_{q_i} \hat{H}_{\underline{k}}, \hat{H}_{\hat{\underline{k}}} \rangle_\pi = \sum_{|\underline{l}|_\infty, |\hat{\underline{l}}|_\infty \leq K} \alpha_{\hat{\underline{k}}}^{\underline{l}} \sqrt{\underline{l}_i} \langle H_{\underline{l}-e_i}, H_{\hat{\underline{l}}} \rangle_\pi \alpha_{\hat{\underline{k}}}^{\hat{\underline{l}}},$$

so that since $\alpha_{\hat{\underline{k}}}^{\underline{l}}$ are derived from the inner products in $L^2(\pi)$ between the original Hermite polynomials $(H_{\underline{l}})_{\underline{l}}$, these inner products are the only values to be computed numerically other than those for the projection $\Pi_K^q f$ of

the observable onto the finite dimensional subspace of $L^2(\tilde{\pi})$ spanned by the first $K+1$ Hermite polynomials given by (B.1). Solving the Poisson equation then reduces to finding the coefficients $\phi_{\underline{k},\underline{l}} \in \mathbb{R}$ of

$$\Pi_K^{(q,p)}\phi = \sum_{|\underline{k}|_\infty, |\underline{l}|_\infty \le K} \phi_{\underline{k},\underline{l}} \hat{H}_{\underline{k}} H_{\underline{l}}$$

solving the linear system

$$\sum_{|\hat{\underline{k}}|_\infty, |\hat{\underline{l}}|_\infty \le K} L_{\underline{k},\underline{l},\hat{\underline{k}},\hat{\underline{l}}} \phi_{\hat{\underline{k}},\hat{\underline{l}}} = \Pi_K^{(q,p)} f = \begin{cases} \sum_{|\hat{\underline{k}}|_\infty \le K} \alpha_{\underline{k}}^{\hat{\underline{k}}} \int f H_{\hat{\underline{k}}} d\pi & \text{if } \underline{l} = \underline{0} \\ 0 & \text{otherwise,} \end{cases} \tag{B.3}$$

where note $L_{\underline{k},\underline{l},\underline{0},\underline{0}} = L_{\underline{0},\underline{0},\hat{\underline{k}},\hat{\underline{l}}} = 0$ so that only $\phi_{\underline{k},\underline{l}}$ for $(\underline{k},\underline{l}) \neq (\underline{0},\underline{0})$ are determined by (B.3) and $\phi_{\underline{0},\underline{0}} = 0$ is enforced independently. Finally, the gradient direction in Γ is given by

$$\begin{aligned}(\Delta\Gamma)_{i,j} &= \int \sum_{|\underline{k}|_\infty, |\underline{l}|_\infty \le K} \phi_{\underline{k},\underline{l}} \hat{H}_{\underline{k}} \sqrt{\underline{l}_i} H_{\underline{l}-e_i} \sum_{|\hat{\underline{k}}|_\infty, |\hat{\underline{l}}|_\infty \le K} \phi_{\hat{\underline{k}},\hat{\underline{l}}} (-1)^{|\hat{\underline{l}}|} \hat{H}_{\hat{\underline{k}}} \sqrt{\hat{\underline{l}}_j} H_{\hat{\underline{l}}-e_j} d\tilde{\pi} \\ &= \sum_{|\underline{k}|_\infty, |\underline{l}|_\infty, |\hat{\underline{k}}|_\infty, |\hat{\underline{l}}|_\infty \le K} \phi_{\underline{k},\underline{l}} \phi_{\hat{\underline{k}},\hat{\underline{l}}} \sqrt{\underline{l}_i \hat{\underline{l}}_j} (-1)^{|\hat{\underline{l}}|} \delta_{\hat{\underline{k}}}^{\underline{k}} \delta_{\hat{\underline{l}}-e_j}^{\underline{l}-e_i} \\ &= \sum_{|\underline{k}|_\infty, |\underline{l}|_\infty \le K} \phi_{\underline{k},\underline{l}} \phi_{\underline{k},\underline{l}-e_i+e_j} \sqrt{\underline{l}_i (\underline{l} - e_i + e_j)_j} (-1)^{|\underline{l}|}\end{aligned} \tag{B.4}$$

where $\left|\hat{\underline{l}}\right| = \hat{\underline{l}}_1 + \cdots + \hat{\underline{l}}_n$ and $\phi_{\underline{k},\underline{l}} = 0$ if there is some i such that $\underline{k}_i > K$ or $\underline{l}_i > K$. More robustly, the asymptotic variance can be discretised first, followed by taking the gradient direction with respect to the approximate asymptotic variance. Namely, (half of) the asymptotic variance $\int \nabla_p \phi^\top \Gamma \nabla_p \phi d\tilde{\pi}$ can be approximated by

$$\sum_{|\underline{k}|_\infty, |\underline{l}|_\infty, |\hat{\underline{k}}|_\infty, |\hat{\underline{l}}|_\infty \le K} \phi_{\underline{k},\underline{l}} L_{\underline{k},\underline{l},\hat{\underline{k}},\hat{\underline{l}}} \phi_{\hat{\underline{k}},\hat{\underline{l}}} \tag{B.5}$$

(or simply the last term in (B.2) replacing $L_{\underline{k},\underline{l},\hat{\underline{k}},\hat{\underline{l}}}$), so that the derivative with respect to the entries $\Gamma_{i,j}$ can be taken as follows. With abuse of notation, let $L^{-1} \in \mathbb{R}^{(K+1)^{2n}-1 \times (K+1)^{2n}-1}$ be the inverse of the matrix depending on Γ given by (B.2) with the $L_{\underline{k},\underline{l},\underline{0},\underline{0}} = L_{\underline{0},\underline{0},\hat{\underline{k}},\hat{\underline{l}}} = 0$ entries removed. Let also $\underline{\phi} \in \mathbb{R}^{(K+1)^{2n}}$ be the vector made up of the coefficients $\phi_{\underline{k},\underline{l}}$ for $\underline{k} + \underline{l} \neq 0$ so that equation (B.3) can be rewritten as

$$\underline{\phi} = L^{-1}((\Pi_K^{(q,p)} f)_2, \ldots, (\Pi_K^{(q,p)} f)_{(K+1)^{2n}})^\top.$$

By (B.2), the derivative of $L_{\underline{k},\underline{l},\hat{\underline{k}},\hat{\underline{l}}}$ with respect to the entry $\Gamma_{i,j}$ is

$$\partial L^{i,j}_{\underline{k},\underline{l},\hat{\underline{k}},\hat{\underline{l}}} := \delta_{\hat{\underline{k}}}^{\underline{k}} \delta_{\hat{\underline{l}}-e_i}^{\underline{l}-e_j} \sqrt{\underline{l}_j \hat{\underline{l}}_i}.$$

Let $\partial L^{i,j} \in \mathbb{R}^{(K+1)^{2n}-1 \times (K+1)^{2n}-1}$ denote the matrix with entries $\partial L^{i,j}_{\underline{k},\underline{l},\hat{\underline{k}},\hat{\underline{l}}}$ except the $\partial L^{i,j}_{\underline{k},\underline{l},\underline{0},\underline{0}}$ and $\partial L^{i,j}_{\underline{0},\underline{0},\hat{\underline{k}},\hat{\underline{l}}}$ entries are deleted. The derivative of (B.5) with respect to the entry $\Gamma_{i,j}$ is then

$$\underline{\phi}^\top \partial L^{i,j} \underline{\phi} + \sum_{|\underline{k}|_\infty, |\underline{l}|_\infty, |\hat{\underline{k}}|_\infty, |\hat{\underline{l}}|_\infty \le K} 2(\partial\phi^{i,j})_{\underline{k},\underline{l}} L_{\underline{k},\underline{l},\hat{\underline{k}},\hat{\underline{l}}} \phi_{\hat{\underline{k}},\hat{\underline{l}}}, \tag{B.6}$$

where $\partial\phi^{i,j} \in \mathbb{R}^{(K+1)^{2n}}$ is the vector given by

$$(\partial\phi^{i,j})_k := \begin{cases} 0 & \text{if } k = 1 \\ -(L^{-1}\partial L^{i,j}\underline{\phi})_{k-1} & \text{otherwise,} \end{cases}$$

so that the gradient direction in Γ is given by the negative of (B.6).

It's also possible to approximate ϕ using a finite difference in q, Hermite projection in p approach in the case when the state space in q is the n-torus; we omit further descriptions but refer to [31] for this direction.

Appendix C Approximation of $\Delta\Gamma$ using independent realisations

One can use the endpoints of a number of independent realisations of (5.2) to approximate the integral with respect to π in (5.1) and, for each of those realisations, to use two additional sets of realisations of (5.2) and (5.4) to approximate each of the expectations under the integral in (5.1).
Fix a starting point (q, p); the first of the expectations in (5.1) (and similarly for the second) can be approximated at time $s = i\Delta t$ with

$$\frac{1}{K}\sum_{k=1}^{K}\nabla f(q^i_{(k)})^\top Dq^i_{(k)}$$

where $K \in \mathbb{N}$, $(q^i_{(k)}, p^i_{(k)})_{i\in\mathbb{N}}$ denotes the solution to (5.2) with initial condition (q, p), noise $\xi^i = \xi^i_{(k)}$ for all $i \in \mathbb{N}$ and where $\xi^i_{(k)}$ are independent as $k = 1, \ldots, K$, i varies and $(Dq^i_{(k)}, Dp^i_{(k)})$ is the corresponding solution to (5.4). Subsequently, introducing an additional population of independent realisations of (5.2) to draw from π after some burn-in period, the change (5.1) in Γ can be approximated by

$$-\frac{1}{L}\sum_{l=1}^{L}\Bigg(\sum_{i=B+1}^{B+T}\frac{\Delta t}{K}\sum_{k=1}^{K}\nabla f(q^i_{(l,k)})^\top Dq^i_{(l,k)}\Bigg)^\top\Bigg(\sum_{i=B+1}^{B+T}\frac{\Delta t}{K}\sum_{k=1}^{K}\nabla f(\tilde{q}^i_{(l,k)})^\top D\tilde{q}^i_{(l,k)}\Bigg)$$

where $B \in \mathbb{N}$ is some burn-in number of iterations, $T \in \mathbb{N}$ is some a posteriori number of iterations depending on whether the magnitude of the entries of $Dq^{B+T}_{(l,k)}$ are smaller than some fixed value for all k, l; furthermore $L \in \mathbb{N}$, $((q^i_{(l,k)}, p^i_{(l,k)}))_{i\in\mathbb{N}}$ denotes the solution to (5.2) with initial condition say $(0, 0)$, noise $\xi^i = \xi^i_{(l,k)}$ for all $i \in \mathbb{N}$ satisfying

$$\xi^i_{(l,k)} = \xi^i_{(l,k')} \qquad \forall i < B, 1 \le k, k' \le K$$

and are independent otherwise, $((\tilde{q}^i_{(l,k)}, \tilde{p}^i_{(l,k)}))_{i\ge B}$ denotes the solution to (5.2) with 'initial' condition

$$(\tilde{q}^B_{(l,k)}, \tilde{p}^B_{(l,k)}) = (q^B_{(l,k)}, -p^B_{(l,k)})$$

for all $1 \le k \le K$, $1 \le l \le L$, independent noise $\xi^i = \tilde{\xi}^i_{(l,k)}$ for $i \ge B$ independent also to $(\xi^i_{(l,k)})_{i\in\mathbb{N}}$. The notation $(Dq^i_{(l,k)}, Dp^i_{(l,k)})$, $(D\tilde{q}^i_{(l,k)}, D\tilde{p}^i_{(l,k)})$ represent the corresponding solutions to (5.4).