UNDERDAMPED LANGEVIN FLOW AND HAMILTONIAN MCMC

ALEXEY IZMAILOV

MAY 5, 2025

ABSTRACT. Early in 2822G, the overdamped Langevin equation was presented as an example of a probability flow. In this expository note, we discuss the probability flow governed by the closely related underdamped Langevin equation, its relation to Hamiltonian dynamics, and the design of sampling methods based on the time integration of the dynamics. Our emphasis is on overarching proof strategies used to analyze underdamped Langevin samplers and properties of the flow which lead to high-order integration schemes. Along the way, many connections with Hamiltonian dynamics and the widely-adopted Hamiltonian Monte Carlo are discussed. Finally, numerical examples illustrate asymptotic convergence rates and properties of several sampling schemes. Implementations of the algorithms are located at https://github.com/alizma/langevin-mcmc.

1 Introduction

Molecular dynamics simulations have persistently been at the forefront of mathematical innovation in computational science. Some of the major challenges in modelling at the microscale involve sampling configurational distributions of particle systems and calculation of expected values of a system's quantities of interest with respect to high dimensional distributions. The microscopic view considers a system composed of N atoms/particles as being fully determined by the positions $q = \langle q_1, \ldots, q_N \rangle$ and momenta $p = \langle p_1, \ldots, p_N \rangle$ of every particle. However, for any meaningful macroscale insight, N must be astronomically large; a single mole of a substance contains $N \sim 10^{23}$ molecules. Yet the objective of molecular dynamics to bridge the micro-macroscale gap has been pursued ceaselessly since the earliest days of computing.

The physical content of a system is captured by the potential function $U(q)$, which describes the energy of N interacting particles. The total energy of a microscopic system is quantified by the Hamiltonian

$$H(q,p) = \frac{1}{2}|p|^2 + U(q)$$

which combines the particle's kinetic and interaction energies. For the purposes of this expository note, our focus is on computational aspects which do not rely on particle masses (e.g. preconditioning), hence the inherent assumption in this definition that all particles are of identical unit mass. Macroscopic properties of a system are determined as averages of certain quantities with respect to a thermodynamic ensemble, which is a probability measure on the configuration space $\mathcal{E}$ of the system at equilibrium. The canonical ensemble refers to a system

in which the number of particles, temperature, and domain of the system all remain fixed. In the setting considered in this note, this will always be the Boltzmann-Gibbs distribution

$$\mu(dq\,dp) = Z_\mu^{-1} e^{-\beta H(q,p)} dqdp, \quad Z_\mu = \int_{\mathcal{E}} e^{-\beta H}$$

for the Hamiltonian H. The normalization constant Z_μ is assumed to be finite (determined by integrating the distribution over its domain of definition, which is in principle not a tractable problem) and $\beta = (k_B T)^{-1} > 0$ is the inverse termperature. For a given observable $\varphi(q,p)$, one would like to estimate $\mathbb{E}_\mu(\varphi)$ (either with respect to the full measure or a marginal) in order to determine a macroscopic quantity of interest, which involves integration against a probability distribution whose normalization is unknown and unknowable. The numerical difficulty of computational modelling in molecular dynamics is compounded by the features of physically relevant models (e.g. metastable potentials), the need to compute dynamical properties such as transition (or exit) times, time scales for dynamics that are on the order of 10^{-15} seconds, and long time horizons.

This note focuses on numerical methods for sampling the canonical ensemble of a system by following Langevin dynamics, whose equations are some of many physically significant models used for the purposes of sampling. Owing to its physical and geometric properties, Langevin dynamics are the origin of many modern sampling algorithms with wide-ranging applications in molecular dynamics [Leimkuhler & Matthews, 2015], uncertainty quantification in machine learning [Bai & Chandra, 2023], optimization [Wibisono, 2018], among many others.

2 Langevin Dynamics

There are several choices to describe microscopic dynamics. Systems for which the entire environment can be simulated, deterministic Hamiltonian dynamics describe the time evolution. However, for settings in which the environment is assumed so large as to keep the temperature fixed (i.e. it is placed in a thermal bath), the system and its ambient environment typically cannot be simulated at once. Consequently, the environment's average effects are replaced by a fluctuation-dissipation process that results in ergodic dynamics with respect to the canonical measure. Langevin dynamics is one of the paradigms used to describe these processes by SDEs.

2.1 Continuous Formulation

The Langevin equation

$$\ddot{q} = -\nabla U(q) - \gamma \dot{q} + \sqrt{2\gamma\beta^{-1}}\dot{W}$$

is a second-order SPDE for the time-dependent position $q_t = q(t)$, describing the time evolution of particles subject to stochastic forcing by standard Brownian motion W and linear dissipation due to friction governed by the friction parameter $\gamma > 0$. This equation can be viewed as Newton's equation of motion with two additional terms. The potential $U(q)$ has several typical assumptions, which will be assumed for the remainder of this note unless otherwise noted.

- **Lipschitz Continuous Gradient**: there exists $L > 0$ such that for all $x, y \in \mathbb{R}^N$,
$$||\nabla V(x) - \nabla V(y)||_2 \leq L||x - y||_2.$$
- **Strong Convexity**: there exists a constant $m > 0$ such that for all $x, y \in \mathbb{R}^N$,
$$V(y) - V(x) \geq \langle \nabla V(x), y - x\rangle + \frac{m}{2}||x - y||_2^2.$$

Under these assumptions, $mI_{N\times N} \preccurlyeq \nabla^2 V(x) \preccurlyeq LI_{N\times N}$. These conditions are sufficient, but not necessary, to establish existence and uniqueness of solutions among other properties.

The full Langevin equation can be rewritten as a system of first-order SDEs posed on the phase space of the system. Denoting the time-dependent momentum $p_t = \dot{q}_t$, the phase space for N particles are the position-momenta pairs $(q, p) \in \mathbb{R}^{2N}$. As a first-order system, the Langevin equation reads
$$\begin{cases} dq_t &= p_t dt \\ dp_t &= -\nabla U(q_t)dt - \gamma p_t dt + \sqrt{2\gamma\beta^{-1}}dW_t. \end{cases}$$
The initial position and momenta are assumed to be distributed according to a probability distribution $\rho_0(p, q)$. For deterministic initial momentum and position p_0, q_0 prescribed at given values, $\rho_0(q, p) \equiv \delta(q - q_0)\delta(p - p_0)$.

The process (p_t, q_t) defines a Markov process with infinitesimal generator
$$\mathcal{L} = \underbrace{p \cdot \nabla_q}_{\text{Hamiltonian drift}} + \underbrace{(-\nabla U(q) - \gamma p) \cdot \nabla_p}_{\text{Friction}} + \underbrace{\gamma \nabla_p \cdot \nabla_p}_{\text{Diffusion}}$$
for which the L^2-adjoint $\mathcal{L}^*$ determines the Fokker-Planck equation
$$\begin{cases} \partial_t \rho &= -p \cdot \nabla_q \rho + \nabla_q V \cdot \nabla_p \rho + \gamma(\nabla_p \cdot (p\rho) + \beta^{-1}\Delta_p \rho) \\ \rho(q, p, 0) &= \rho(q, p) \end{cases}$$
This equation governs the evolution of the probability distribution function ρ for the process $\{q_t, p_t\}$ in the system's phase space $\mathbb{R}^{2N}$. As such, the Langevin dynamics at finite time have an interpretation as a transport map for the initial distribution of position and momenta to a distribution on these variables at a later time.

The Liouville operator
$$B := -p \cdot \nabla q + \nabla_q U \cdot \nabla_p$$
determines the Hamiltonian vector field $b(q, p) = (p, -\nabla V)$ for the Hamiltonian $H(p, q)$. Under Hamiltonian dynamics,
$$\begin{cases} \dot{q} = \partial_p H = p \\ \dot{p} = -\partial_q H = -\nabla U(q), \end{cases}$$
the Hamiltonian energy is conserved. The null space of B contains arbitrary smooth functions $f(H)$ satisfying $Bf(H) = 0$, which leads to many possible invariant distributions for Hamiltonian systems. However, the noise and dissipation terms in the Langevin equation determine the unique invariant distribution for the process.

Proposition 2.1 (Invariant Distribution for Langevin Flow)**.** Given a smooth potential $U(x)$, the Markov process with infinitesimal generator $\mathcal{L}$ above has the unique invariant distribution

$$\rho_\beta(p,q) \propto e^{-\beta H(p,q)}.$$

The invariant distribution (which can be interpreted as the infinite-time distribution to which the Fokker-Planck equation transports an initial distribution on position-momenta) of Langevin dynamics is therefore independent of the friction of the system. Establishing this result requires non-standard tools, owing to the fact that the operator of the Fokker-Planck equation is not uniformly elliptic, which will be discussed later.

2.2 Friction Limits

The Langevin equation results in two different dynamics when considering limits of the friction parameter γ. For rigorous justifications of these limiting dynamics, we refer to Section 6.5 of [Pavliotis, 2014]. In both cases, the resulting dynamics remain Markovian while corresponding to different distribution flows.

Definition 2.2 (Overdamped Langevin Equation)**.** In the limit $\gamma \to +\infty$ of the Langevin equation, the resulting equations read

$$dq_t = -\nabla U(q_t)dt + \sqrt{2\beta^{-1}}dW_t.$$

The overdamped dynamics only evolve in an N-dimensional phase space, on account of the fact that there is no equation for momentum evolution. As such the invariant distribution for these dynamics is proportional to $\exp(-U(x))$. As a result, the dynamics typically result in slow, diffusive exploration and samples generated by following these dynamics sample only from the position component of the Gibbs distribution. The analysis of overdamped Langevin dynamics is extensive (see for example [Vempala & Wibisono, 2023]) and will not be discussed further since the focus is on the other limiting dynamics

Definition (Underdamped Langevin Equation)**.** In the limit $\gamma \to 0^+$ of the Langevin equation, the resulting equations read

$$\begin{cases} dp_t &= q_t dt \\ dq_t &= -\gamma q_t dt - \beta \nabla U(p_t)dt + \sqrt{2\gamma\beta}dB_t. \end{cases}$$

The underdamped dynamics retain the inertial effects of the full Langevin dynamics and explore the full position-momenta phase space. Samples generated along these dynamics sample the canonical ensemble of the system. These equations have a richer structure owing to the presence of both position and momenta which lead to a somewhat specialized numerical integrator theory as compared to techniques typically used for the overdamped equations. Additionally, a line of recent progress in the area applies higher-order integrators for the underdamped equations in high-friction regimes and analyzes their behaviors for $\gamma \to +\infty$, a

different friction regime than the equations are derived for. Surprisingly some schemes for the underdamped equations remain consistent in the high-friction regime while offering additional favorable properties (e.g. superconvergence). The focus of the remainder of this report is on sampling methods based on the underdamped Langevin equations.

The continuous underdamped Langevin process is explicitly related to simpler dynamics that are well understood and have closed expressions for the invariant distribution. We will use the following result in later numerical illustrations because the analytical result simplifies checking many assumptions for schemes' properties and simplifies error computations.

Example 2.3 (Gaussian Potential and Ornstein-Uhlenbeck Process)**.** In the case off a Gaussian potential, the underdamped Langevin dynamics admit an explicit connection to the Ornstein-Uhlenbeck processes. the fidelity of sampling for implementations. For the Gaussian potential $V(p) = \frac{1}{2}(p-m)^\top C^{-1}(p-m)$, the invariant distribution of the underdamped Langevin process (obtained by considering the corresponding Hamiltonian) is the Gibbs distribution

$$p(p,q) \propto \exp\left(-\frac{1}{2}(p-m)^\top C^{-1}(p-m) - \frac{1}{2}q^\top q\right).$$

This can be rewritten as a multivariate Gaussian over $\mathbb{R}^{2N}$

$$p(z) \propto \exp\left(-\frac{1}{2}(z-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(z-\boldsymbol{\mu})\right),$$

where $z = (p,q)^\top$, $\boldsymbol{\mu} = (m, \mathbf{0})^T$, and

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} C^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

This result agrees with the invariant distribution of a a multivariate OU process with state vector (p_t, q_t) for the given mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{\Sigma}$,

$$\begin{cases} dp_t &= q_t dt \\ dq_t &= -\gamma q_t dt - \boldsymbol{C}^{-1}(p_t - \boldsymbol{m})dt + \sqrt{2\gamma}dB_t. \end{cases}$$

by directly evaluating the gradient of the Gaussian potential.

Remark 2.4 (Hamiltonian Dynamics)**.** While this report primarily discusses sampling based on discretizations of the underdamped Langevin dynamics, there are several connections with Hamiltonian dynamics whose equations are the basis for one of today's foremost sampling algorithms, Hamiltonian Monte Carlo (HMC) [Betancourt, 2018]. The equations of motion for a Hamiltonian $H(p,q)$,

$$\begin{cases} \dot{q} = \partial_p H \\ \dot{p} = -\partial_q H \end{cases}$$

have the same position-momenta interpretations. Indeed, these equations are exactly the deterministic components of the underdamped Langevin equations when $H(p,q) = \frac{1}{2}|p|^2 + U(q)$.

These equations present several desirable properties in describing a system: the Hamiltonian is strictly conserved over time, there is no dissipation due to external forces, the equations are time-reversible, and admit a symplectic structure that preserve the phase space volume of the system. There is also no inherent equilibrium as orbits (e.g. due to Poincaré recurrence) may indefinitely continue motion. In a crude yet not totally dishonest sense, the underdamped Langevin equations are merely those of Hamiltonian dynamics with the addition of friction and noise. Indeed in the years following HMC's wide adoption, ostensibly new sampling methods that operated by Hamilton's equations with these additions were published [Chen et al., 2014].

2.3 The Wasserstein Metric

The convergence analysis of underdamped Langevin dynamics at both discrete and continuous levels requires a common notion of distance. Most convergence results and rates of convergence for different sampling schemes are stated in the Wasserstein metric, which is defined on the space of probability distributions with finite moments up to a prescribed order. This metric is particularly well-suited for interpreting results in relation to the measure transport by the Langevin flow. The Wasserstein metric quantifies the cost of transporting the mass of the initial distribution for a Langevin flow to the distribution at a later time. Its properties make it an attractive choice over alternative choices of distance between probability measures, such as KL divergence because of its interpretations in terms of optimal transport (and that the arguments to the metric are not required to have the same domain, which is important for comparing empirical discrete distributions to their continuous counterparts). We briefly summarize the definitions as presented in [Figalli & Glaudo, 2023].

Definition 2.5 (Transport Plan)**.** Given two probability measures μ, ν, we call $\gamma \in \mathcal{P}(X \times Y)$ a transport plan (synonymously, a coupling or transference plan) if

$$(\pi_X)_{\#}\gamma = \mu \quad \text{and} \quad (\pi_Y)_{\#}\gamma = \nu$$

where

$$\pi_X(x, y) = x, \ \pi_Y(x, y) = y \quad \forall (x, y) \in X \times Y.$$

Let $\Gamma(\nu, \mu)$ be the set of couplings of ν and μ. The condition above equivalently states

$$\Gamma(\nu, \mu) := \{\text{laws on } X \times Y \text{ whose first marginal is } \mu \text{ and second marginal is } \nu\}.$$

More explicitly, this is equivalent to requiring all Borel and bounded $\varphi, \psi : X \to \mathbb{R}$,

$$\int_{X\times Y} \varphi(x) d\gamma(x, y) = \int_{X\times Y} \varphi \circ \pi_X(x, y) d\gamma(x, y) = \int_X \varphi(x) d\mu(x)$$

and

$$\int_{X\times Y} \psi(x) d\gamma(x, y) = \int_{X\times Y} \psi \circ \pi_Y(x, y) d\gamma(x, y) = \int_Y \psi(x) d\nu(x).$$

Definition 2.6 (2-Wasserstein Distance)**.** For two probability measures ν and μ with finite second moment, the 2-Wasserstein distance is defined as

$$W_2(\nu,\mu) := \left(\inf_{\pi\in\Gamma(\nu,\mu)} \mathbb{E}_{(x,y)\sim\pi}||x-y||^2\right)^{1/2}$$

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

Computing the Wasserstein distance between two arbitrary distributions is in general a difficult computational problem. Nevertheless, there are explicit forms in some special cases that are useful for illustrating properties of sampling schemes.

Example 2.7 (2-Wasserstein between Gaussians)**.** For two Gaussian distributions p and q with means μ_p, μ_q and covariances Σ_p, Σ_q, respectively,

$$W_2(p,q)^2 = \frac{1}{2}\left(||\mu_p-\mu_q||_2^2 + \text{trace}(\Sigma_p+\Sigma_q-2\sqrt{\Sigma_p\Sigma_q})\right).$$

Moreover, when the covariance matrices commute (both are diagonalizable),

$$W_2(p,q)^2 = ||\mu_p-\mu_q||^2 + \left|\left|\Sigma_p^{1/2}-\Sigma_q^{1/2}\right|\right|_F^2,$$

where the second term is the Frobenius norm.

In later examples, this expression will be used to confirm theoretical properties of sampling schemes when applied to Gaussian potentials by considering the empirical distribution (i.e. the discrete "invariant distribution" from the sampling chain) after many samples are computed as approximately Gaussian. With this simplification, one can compute the empirical mean and covariance of generated samples and then use the foregoing expression against the true Gaussian distribution.

Wasserstein spaces have an extensive geometric theory that has been the subject of much recent work by both categories of researchers interested in either the continuous-time processes or their discrete counterparts. As one of many directions of work, we refer to [Wibisono, 2018], wherein the sampling problem of overdamped Langevin was related to classical optimization problems posed in Wasserstein space.

2.4 Continuous Exponential Contraction

The continuous-time underdamped Langevin process exponentially converges to its invariant distribution at a rate determined by properties of the potential. However, merely having a convergence result for Langevin dynamics is not sufficient for analyzing sampling algorithms; the convergence of a discrete process must be compared against that of the continuous-time process. To this end, results about the continuous dynamics show that solutions of the Langevin SDEs enjoy an exponential contraction in time to the invariant distribution, when measured in the Wasserstein metric. Studying the proofs of this result reveals common proof patterns

that transfer to the analysis of numerical schemes, namely synchronous coupling arguments as well as recasting contraction as a matrix problem. Early general results, as well as alternative conditions under which the underdamped Langevin process has these properties date to [Roberts & Tweedie, 1996], but we present the comparatively simpler analysis from [Cheng et al., 2018].

This continuous-time process admits an exponential convergence to the stationary distribution governed by the rate $\kappa = L/m$ from properties of the potential.

Theorem 2.8 (W_2 Exponential Convergence)**.** Let ρ_0 be arbitrary distribution with $(p_0, q_0) \sim \rho_0$. Let π_0 and $\Phi_t \pi_0$ be the distributions of $(p_0, p_0 + q_0)$ and $(p_t, p_t + q_t)$, respectively, and π^* be the unique invariant distribution of the process for $(p_t, p_t + q_t)$. Then

$$W_2 \left(\Phi_t \pi_0, \pi^*\right) \leq e^{-t/2\kappa} W_2 \left(\pi_0, \pi^*\right).$$

Remark 2.9. Exponential contraction results for underdamped Langevin flow can be proven under considerably weaker assumptions on the potential. For instance, the Lipschitz gradient and strong convexity assumptions can be loosened to a Poincaré inequality on the potential: supposing

$$\lim_{|x| \to \infty} \frac{|\nabla U(x)|^2}{2} - \Delta U(x) = +\infty,$$

there exists a constant $\lambda > 0$ such that under the associated Gibbs measure $\mu(dx) := Z^{-1} e^{-\beta U(x)}$, for every $f \in C^1(\mathbb{R}^N) \cap L^2(\mu)$ with average value 0,

$$\lambda ||f||^2_{L^2(\mu)} \leq ||\nabla f||^2_{L^2(\mu)}.$$

A convexity condition $\nabla^2 U \geq \lambda I$ (otherwise known as the Bakry-Emery criterion) guarantees a Poincaré inequality with the same constant. Under this assumption, it can be shown that the continuous process admits an exponential contraction result with rate $e^{-\lambda t}$, as presented in Section 4.4 of [Pavliotis, 2014]. The proof for exponential contraction under a Poincaré inequality is not suggestive of the proof pattern frequently used for results on sampling schemes, and as such is not presented here.

The proof for exponential convergence in our setting depends crucially on the following contraction result for solutions of the SDE.

Theorem 2.10 (Flow Contraction)**.** Let $(x_0, v_0), (y_0, w_0) \in \mathbb{R}^{2N}$, p_0 the Dirac delta distribution at (x_0, v_0) and let p_0' the Dirac delta distribution at (y_0, w_0). Let $\beta = 1/L$ and $\gamma = 2$ in the underdamped Langevin equation (chosen to simplify constant factors). Then for every $t > 0$, there exists a coupling $\zeta_t \left((x_0, v_0), (y_0, w_0)\right) \in \Gamma \left(\Phi_t p_0, \Phi_t p_0'\right)$ such that

$$\begin{aligned}
&\mathbb{E}_{((x_t, v_t), (y_t, w_t)) \sim \zeta_t((x_0, v_0), (y_0, w_0))} \left[\|x_t - y_t\|_2^2 + \|(x_t + v_t) - (y_t + w_t)\|_2^2 \right] \\
&\leq e^{-t/\kappa} \left\{ \|x_0 - y_0\|_2^2 + \|(x_0 + v_0) - (y_0 + w_0)\|_2^2 \right\}.
\end{aligned}$$

The proof for this result is illustrative of many proofs in this area. It relies on a synchronous coupling argument by independent p_0, p_0' and shared Brownian noise W_t. The continuous process implies

$$\frac{d}{dt}\left[(x_t + v_t) - (y_t + w_t)\right] = -\left(\gamma - 1\right) v_t - \beta \nabla f(x_t) - \left\{-(\gamma - 1) w_t - \beta \nabla f(y_t)\right\}.$$

where the synchronous coupling results in cancellation of the Brownian motion terms. Denoting $z_t := x_t - y_t, \psi_t := v_t - w_t$ and

$$H_t := \int_0^1 \nabla^2 U(x_t + h(y_t - x_t)) dh,$$

the mean value theorem (assuming second-order differentiability) implies

$$\nabla U(x_t) - \nabla U(y_t) = H_t z_t.$$

Moreover, expanding definitions,

$$\frac{d}{dt}[z_t + \psi_t] = -((\gamma - 1)\psi_t + \beta H_t z_t), \quad \frac{d}{dt}[z_t] = \psi_t.$$

From these results,

$$\begin{aligned}
&\frac{d}{dt}\left[\|z_t + \psi_t\|_2^2 + \|z_t\|_2^2\right] \\
&= -2\begin{bmatrix} z_t + \psi_t & z_t \end{bmatrix} \underbrace{\begin{bmatrix} (\gamma - 1) I_{d\times d} & \beta H_t - (\gamma - 1) I_{d\times d} \\ -I_{d\times d} & I_{d\times d} \end{bmatrix}}_{S_t} \begin{bmatrix} z_t + \psi_t \\ z_t \end{bmatrix}.
\end{aligned}$$

The aim is to bound this quantity in time as it is directly in the argument of the expectation in the definition of the Wasserstein metric. Bounding its change through time in terms of its value at t indicates the convergence of the Langevin flow's contraction in time along a coupling of an initial distribution to the invariant distribution. For $x \in \mathbb{R}^{2N}$, $x^T S_t x = x^T Q_t x$ where $Q_t := (S_t + S_t^T)/2$, so from earlier,

$$\frac{d}{dt}\left[\|z_t + \psi_t\|_2^2 + \|z_t\|_2^2\right] = -2[z_t + \psi_t, z_t]^T S_t [z_t + \psi_t, z_t] = -2[z_t + \psi_t, z_t]^T Q_t [z_t + \psi_t, z_t].$$

So to show the claimed result, it suffices to determine a lower bound on the eigenvalues of Q_t. *This is a recurring situation in many proofs in this area for both continuous and discrete time processes. Essentially, an overarching proof technique converts the problem of establishing a contraction for a process to that of showing properties of matrices (e.g. positive definiteness) and estimating their eigenvalues.* The eigenvalue problem for Q_t reads

$$\det \begin{bmatrix} (\gamma - 1 - \lambda) I_{N\times N} & \frac{\beta H_t - \gamma I_{N\times N}}{2} \\ \frac{\beta H_t - \gamma I_{N\times N}}{2} & (1 - \lambda) I_{N\times N} \end{bmatrix} = 0$$

which by properties of block matrices is equivalent to solving for λ in

$$\det\left((\gamma - 1 - \lambda)(1 - \lambda) I_{N\times N} - \frac{1}{4}(\beta H_t - \gamma I_{N\times N})^2\right) = 0.$$

Recall that H_t is a matrix in its own right (as an integral of $\nabla^2 U$ along a line). The strong convexity and Lipschitz smoothness assumptions ensure that the eigenvalues of H_t satisfy $0 < m \leq \Lambda_j < L$. From a diagonalization of H_t, the characteristic equation for Q_t results in N eigenvalue equations

$$(\gamma - 1 - \lambda)(1 - \lambda) - \frac{1}{4}(\beta\Lambda_j - \gamma)^2 = 0.$$

With the choice of γ, β,

$$\lambda_j := 1 \pm \left(1 - \frac{\Lambda_j}{2L}\right)$$

are solutions to the characteristic equation of Q_t. Substituting $\kappa = L/m$ and using the estimates on Λ_j, it follows that the eigenvalues of Q_t are bounded from below by $1/2\kappa$. Substituting this result into the earlier computations,

$$\frac{d}{dt}\left[\|z_t + \psi_t\|_2^2 + \|z_t\|_2^2\right] = -2[z_t + \psi_t, z_t]^T Q_t [z_t + \psi_t, z_t] \geq -\frac{1}{\kappa}[||z_t + \psi_t||_2^2 + ||z_t||_2^2].$$

The result follows by Gronwall's inequality applied to this differential inequality where the corresponding differential equation evolving from $t = 0$.

This result immediately leads to the aforementioned exponential contraction result as measured in W_2. Namely, choosing an optimal coupling $\zeta \in \Gamma(\pi_0, \pi^*)$ such that

$$\mathbb{E}_{\zeta_0}[||x_0 - y_0||_2^2 + \|x_0 - y_0 + v_0 - w_0\|_2^2] = W_2^2(\pi_0, \pi^*),$$

the definition of the Wasserstein metric and tower property of expectation implies that, with the coupling ζ_t from the foregoing result,

$$W_2^2(\Phi_t \pi_0, \pi^*) \leq \mathbb{E}_{(x_0,v_0,y_0,w_0)\sim\zeta_0}\left[\mathbb{E}_{(x_t,v_t,y_t,w_t)\sim\zeta_t(x_0,v_0,y_0,w_0)}\left[\|x_t - y_t\|_2^2 + \|x_t - y_t + v_t - w_t\|_2^2 \middle| x_0, y_0, v_0, w_0\right]\right]$$

Applying the preceding result, this implies

$$W_2^2(\Phi_t \pi_0, \pi^*) \leq \mathbb{E}_{(x_0,v_0,y_0,w_0)\sim\zeta_0}\left[e^{-t/\kappa}\left(\|x_0 - y_0\|_2^2 + \|x_0 - y_0 + v_0 - w_0\|_2^2\right)\right].$$

Finally since ζ_0 is taken as the optimal coupling,

$$W_2^2(\Phi_t \pi_0, \pi^*) \leq e^{-t/\kappa} W_2^2(\pi_0, \pi^*),$$

which is precisely the exponential contraction property of the continuous flow.

The proof techniques employed here are used in many other scenarios, in particular both for showing exponential contraction of discrete-time flows and for estimating their distance against the continuous flow. The ideas are also not constrained to the analysis of the underdamped Langevin flows, as synchronous coupling is used in an analogous fashion for analyzing overdamped Langevin flows.

2.5 Hypoellipticity and Contraction

Establishing the existence, uniqueness, and exponential convergence to equilibrium of the underdamped Langevin equations requires new techniques because the infinitesimal generator of the process is not elliptic. Consequently, classical theorems which require a uniform ellipticity assumption (e.g. Section 4.1 in [Pavliotis, 2014]) are not applicable to the corresponding Fokker-Planck equation. Before elaborating on the technical consequences, this issue can be seen heuristically. The Fokker-Planck operator is not uniformly elliptic since it contains second-derivatives only with respect to p and not q. One can alternatively reason about this by noticing that the Langevin equation implies an anisotropic energy dissipation in position and momentum: the friction term only directly dampens the momentum variable p and moreover the Laplacian Δ_p introduces diffusion only in the momentum variable. Hence, energy is only directly removed from p and only indirectly from q through the equations' coupling. Considering the $\mathbb{R}^{2N}$ Euclidean norm, $||(p,q)||_2^2 = ||p||_2^2 + \|q\|_2^2$ the two variables of the system are treated symmetrically. So while there is direct dampening in p, which implies roughly that there is exponential contraction in that variable, the q-component is only stabilized via the coupling and thus suffers weaker, delayed dissipation. The implications for this is that it is not possible to prove convergence with respect to the standard Euclidean norm. A subtle point about the proof from [Cheng et al., 2018] presented in the previous section is that it implicitly leverages a weighted Euclidean norm; couplings are considered for expressions of the form $\|x_t - y_t\|_2^2 + \|(x_t + v_t) - (y_t + w_t)\|_2^2$ which is distinct from $\|(x_t - y_t, v_t - w_t)\|_2^2$.

Showing that contraction results in the standard Euclidean norm are not possible to prove requires some manipulation of the infinitesimal generator. Let $A_i := (\beta)^{-1/2}\partial_{p_i}$, for which the $L^2(\mu)$ adjoint reads $A_i^* = \beta^{1/2}p_i = (\beta)^{-1/2}\partial_{p_i}$. The infinitesimal generator $\mathcal{L}$ can be rewritten in the sum of squares form

$$\mathcal{L} = -B - \gamma \sum_{i=1}^{N} A_i^* A_i$$

where B is the Liouville operator $B := -p \cdot \nabla q + \nabla_q V \cdot \nabla_p$. The Fokker-Planck equation is then the evolution equation

$$\partial_t \rho + (A^* A - B)\rho = 0.$$

The Liouville operator is antisymmetric ($B^* = -B$) under the L^2 inner product, which follows by expanding definitions and integrating by parts with respect to p. From this antisymmetry,

$$\frac{1}{2}\frac{d}{dt}||\rho||^2 = -||A\rho||^2 \leq 0$$

and thus $\langle -(B - A^*A)f, f\rangle \geq 0$ for all admissible f. But this is not sufficient to ensure exponential convergence to the invariant distribution. Section 6.2.1 of [Pavliotis, 2014] gives a precise exposition and introduction to hypocoercivity analysis of the Fokker-Planck equation.

Contraction results for underdamped Langevin dynamics are proven in a weighted Euclidean norm. Following [Monmarché, 2021], the following norm is used in contraction results considered further in this report.

Definition 2.11 (Modified Euclidean Norm)**.** For $z = (x, v) \in \mathbb{R}^{2d}$ and $a, b > 0$, the norm

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

This is equivalent to the Euclidean norm on $\mathbb{R}^{2d}$ as long as $b^2 < a$, and moreover, under the condition $b^2 < a/4$,

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

This modified norm induces a corresponding change for the Wasserstein metric

Definition 2.12 (Modified Wasserstein Metric)**.** For two probability measure μ, ν over $\mathbb{R}^{2N}$ with finite second moments,

$$W_{2,a,b}(\nu, \mu) := \left(\inf_{\pi \in \Gamma(\nu,\mu)} \mathbb{E}_{(x,y)\sim\pi}||x - y||_{a,b}^2\right)^{1/2}.$$

3 Sampling Methods for Underdamped Langevin Dynamics

The purpose of this note is to illustrate numerical sampling schemes for Langevin dynamics and the tools used in their analysis. As with standard timesteppers, the analysis is primarily concerned with rates of convergence with respect to discretization parameters, stability, scalability, and robustness. Answering these questions and designing better sampling algorithms requires a structural understanding of the equations.

3.1 Prelude: HMC

HMC is a widely-used sampling method whose structure is similar to that of underdamped Langevin dynamics and from which many ideas have been applied to designing and analyzing Langevin-based sampling approaches. As seen earlier, the Hamiltonian equations at the core of HMC are the deterministic components of underdamped Langevin flow. Because of the many conserved quantities, symplectic integrators which preserve properties of the Hamiltonian were developed for this purpose. Their defining characteristic is that each discrete time-step exactly preserves the symplectic structure of the phase space to match the continuous dynamics and moreover have bounded energy errors. Traditional numerical integrators, such as Runge-Kutta methods, cannot be applied naively as they do not preserve this symplectic structure: the discrete system may artifically gain or lose energy, the phase volume is distorted, and long-time dynamics are qualitatively incorrect and disagree with the true equilibrium distribution. The most common integrator used in practice is the velocity-Verlet or leapfrog integrator. The position and momenta after a single time step of size $h > 0$ are updated by the sequence

$$\begin{aligned} q_{1/2} &= q_0' - h/2\nabla U\left(p_0\right) \\ p_1 &= p_0 + hq_{1/2} \\ q_1' &= q_{1/2} - h/2\nabla U\left(p_1\right). \end{aligned}$$

These symplectic integration steps generate samples in HMC. For a sampling chain at position $p_n(0)$, a random Gaussian momentum $q_n(0) \in \mathcal{N}(0, Id)$ is first generated. Then, the particles are integrated forward in time by L steps by the leapfrog algorithm to time Lh to compute $(x_n(Lh), p_n(Lh))$. At this stage, a Metropolis-Hastings step can be used to determine the transition to the next sample.

3.2 Analysis of Sampling Methods

The analysis of different sampling schemes is principally concerned with proving long-time convergence estimates of discrete trajectories generated along from the underdamped Langevin dynamics. Many works base their proof techniques on first establishing long-time convergence estimates for the continuous-time process corresponding to a discretization and then analyze the discretization error (e.g. [Cheng et al., 2018]). Others directly prove a Wasserstein contraction for the discrete-time Markov chain, after which the bias of the equilibrium/invariant distribution is analyzed (e.g. [Monmarché, 2021]). Results following the latter strategy have the additional benefit of providing non-asymptotic results including, but not limited to, confidence intervals for metropolized schemes which we do not discuss. Overall, the aims of asymptotic rates is to determine the dependency of sampling error on the ambient dimension d of the dynamics and the time step size $h > 0$. Estimates for the number of required iterations to guarantee an error to a tolerance $\varepsilon > 0$ are of additional interest.

Remark 3.1 (Measures of Convergence)**.** The convergence of numerical integrators for SDEs is not classically analyzed with respect to the Wasserstein metric, which with traditional senses of convergence, address distinct aspects of approximation quality. These types of convergence in principle should not be compared because of their different meanings. In brief, we summarize other choices seen in SDE integrator analysis and why convergence results with respect to Wasserstein are particularly well-suited for analyzing the quality of sampling schemes.

- **Strong sense convergence** measures path-wise proximity between the numerical solution Y_T and the true solution X_T at time T. A method has strong order γ if there exists a constant $C > 0$ such that
$$\mathbb{E}\left[|X_T - Y_T|\right] \leq Ch^\gamma,$$
This sense of convergence is most useful when the accuracy of individual trajectories generated from the SDE are most important.

- **Weak sense convergence** quantifies the approximation of expectations of some function(al)s of the numerical solution and true solution. A method has weak order γ if, for sufficiently smooth test functions f, there exists a constant $C > 0$ such that
$$|\mathbb{E}[f(X_T)] - \mathbb{E}[f(Y_T)]| \leq Ch^\gamma.$$

For the sake of determining the accuracy of approximating a *distribution*, neither of the above are satisfactory. The Wasserstein metric is special for its ability to measure distance in the

space of probability distributions in a way that is not agnostic to the inherent geometric information (e.g. ensures that a sampled distribution must have approximately the same modes). It also accounts for both the bias incurred from approximation error and the noise in the sampling (unlike strong convergence). Finally, Wasserstein convergence typically yields dimension-dependent rates, which enables an analysis of the scalability of any particular method.

3.3 Synchronous Coupling and Contraction

The proofs for many convergence results of sampling schemes have a common pattern mentioned earlier as part of the proof for exponential contraction of the continuous-time Langevin dynamics. The overall idea is to first establish contraction results of two discrete trajectories via a synchronous coupling through a shared stochastic noise. As discussed earlier, these results are necessarily proven in a weighted Euclidean norm because the corresponding Fokker-Planck operator is not uniformly elliptic. The following result, proven as Corollary 20 in [Monmarché, 2021], then asserts that contraction in the weighted Euclidean norm implies contraction in Wasserstein distance.

Theorem 3.2. Suppose a numerical scheme with discrete transition kernel P_h is applied to Langevin dynamics with an m-strongly convex, $M-\nabla$Lipschitz potential U. Let $(x_n, v_n), (\tilde{x}_n, \tilde{v}_n)$ be synchronously coupled chains satisfying the contraction property

$$\|(x_n - \tilde{x}_n, v_n - \tilde{v}_n)\|_{a,b}^2 \leq C(1 - c(h))^n \|(x_0 - \tilde{x}_0, v_0 - \tilde{v}_0)\|_{a,b}^2$$

where $\gamma^2 \geq C_\gamma M, h \leq C_h(\gamma, \sqrt{M})$ are constants determined by properties of the Langevin dynamics, $a, b > 0$ are constants satisfying $b^2 < a/4$. Then for all such γ, h and all probability distributions $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^{2d})$, for all $n \in \mathbb{N}$,

$$W_2^2(P_h^n \nu, P_h^n \mu) \leq 3C \max\{a, a^{-1}\}(1 - c(h))^n W_2^2(\mu, \nu)$$

where $C > 0$ is a constant. Moreover, P_h has a unique invariant measure $\pi_h \in \mathcal{P}_2(\mathbb{R}^{2d})$ determined by the step size.

Following [Leimkuhler et al., 2024], contraction and convergence in the Wasserstein distance rely on a relation of the contraction estimate to a problem regarding properties of certain matrices. For two Markov chains in $\mathbb{R}^{2N}$ generated by the same discretization scheme, $z_n = (p_n, q_n), \tilde{z}_n = (\tilde{p}_n, \tilde{q}_n)$, one wishes to choose $a, b > 0, b^2 < a/4$ and determine a function $c(h)$ that satisfy the estimate

$$\|\tilde{z}_{n+1} - z_{n+1}\|_{a,b}^2 \leq (1 - c(h))\|\tilde{z}_n - z_n\|_{a,b}^2.$$

The choice of norm parameters a, b also have consequences for the step size and friction under which a given scheme is applicable. Generalizing the proof structure observed in [Cheng et al., 2018], this contraction property is equivalent to showing that

$$\overline{z}_n^T \left((1 - c(h))G - P^T G P\right) \overline{z}_n \geq 0, \quad G := \begin{bmatrix} I_N & bI_N \\ bI_N & aI_N \end{bmatrix}$$

where $\overline{z}_n := \tilde{z}_n - z_n$ and subsequent iterates are related by $\overline{z}_{n+1} = P\overline{z}_n$. Proving contraction is then equivalent to showing that the matrix $H := (1 - c(h))G - P^T GP$ is positive definite. This proof strategy can and has been generically applied with considerable success to show contraction for a variety of numerical schemes (see [Leimkuhler et al., 2024] for applications to several classes of sampling methods). The benefit of this proof strategy is that the eigenvalues of P determine admissible h and has the flexibility for variable step sizes, which many of the early proof strategies cannot simply account for. We provide examples of this proof technique for specific schemes in later sections.

As the final step in these analyses, a sandwich argument relates the Wasserstein distances of the discrete distributions to the invariant distribution of the continuous process. Recalling π_0 is the distribution of $(p_0, p_0 + q_0)$ where $(p_0, q_0) \sim \rho_0$, $\Phi_t\pi_0$ the distribution of $(p_t, p_t + q_t)$ with associated invariant distribution π^*, and the distribution ρ^* being the true invariant distribution of the process, [Cheng et al., 2018] prove the following.

Lemma 3.3 (Sandwich Inequality)**.** For distributions as labelled above,

$$\frac{1}{2}W_2(\rho_t, \rho^*) \leq W_2(\pi_t, \pi^*) \leq 2W_2(\rho_t, \rho^*).$$

Once the parameters a, b of the weighted Euclidean norm are determined, the equivalence of norms ensures similar properties for the associated weighted Wasserstein metric.

3.4 Operator Splitting

We describe a general framework by which new sampling schemes have been designed and analyzed. The standard ULMC algorithm, which relies on an Euler-Maruyama discretization of the undnedamped Langevin equations, will emerge as a simple case. Part of the motivation is that the error in sampling (and thus in computing ensemble averages) comes from two sources: bias on the invariant measure due to the discretization of the continuous dynamics, and statistical errors. The latter are governed by typical statistical results (e.g. CLT), but the former can be controlled by adapting the sampling method to the dynamics. The results and schemes presented in this note have a tie to Trotter-Strang splitting (which has a long history in deterministic PDE numerics, see [McLachlan & Quispel, 2002]) for the Langevin semigroup. The overall goal is to design numerical schemes which adhere the continuous dynamics and minimize the error incurred by discretization.

Recall the generator for the Markov process (p_t, q_t) associated with underdamped Langevin is the differential operator

$$\mathcal{L} = q \cdot \nabla_p - \nabla_p \cdot V\nabla_q + \gamma(-q\nabla_q + \Delta_q)$$

which can be additively split into the components

$$\mathcal{L} = \mathcal{L}_\mathcal{A} + \mathcal{L}_\mathcal{B} + \gamma\mathcal{L}_\mathcal{O}$$

with the operators

$$\mathcal{L}_{\mathcal{A}} = \langle q, \nabla_x \rangle, \quad \mathcal{L}_{\mathcal{B}} = -\langle \nabla U, \nabla_q \rangle, \quad \mathcal{L}_{\mathcal{O}} = -\langle q, \nabla_q \rangle + \Delta_v.$$

These terms can be interpreted in terms of the equation's linear positional drift, impulse from a spatially-dependent external force, and the dissipative-stochastic term as in an OU process. For a time step-size $h > 0$, these terms correspond to discrete mappings:

$$\begin{aligned}
&\mathcal{A} : (p, q) \to (p + hq, q) \\
&\mathcal{B} : (p, q) \to (p, q - h\nabla U(p)), \\
&\mathcal{O} : (p, q) \to \left(p, \eta q + \sqrt{1 - \eta^2} G\right)
\end{aligned}$$

where $\eta := \exp(-\gamma h)$ and $G \sim \mathcal{N}(0, I_d)$. These operators can be interpreted as distinct components of the dynamics: $\mathcal{L}_{\mathcal{A}}$ governs the change of position subject to the velocity; $\mathcal{L}_{\mathcal{B}}$ governs acceleration due to the system's potential; and $\mathcal{L}_{\mathcal{O}}$ governs the friction and stochastic noise in the velocity, corresponding to an exact OU process. By separating $\mathcal{L}$ into these components, the deterministic and stochastic dynamics can be analyzed separately. Moreover, each individual discrete step for these operators can be integrated exactly: $\mathcal{A}$ becomes a linear translation because the velocity v is kept constant with no external forces or noise; $\mathcal{B}$ becomes a linear velocity change because $\nabla U(p)$ is evaluated at a fixed position without position drift; and $\mathcal{O}$ is an OU process, for which the discrete mapping can be seen as an evaluation of the analytical solution. Before considering convergence results, it's first necessary to show that such operator splittings preserve the invariant measure of the underdamped Langevin dynamics at the continuous level.

Recall that the definition of a Markov semigroup gives the transition function P_t in terms of the infinitesimal generator $\mathcal{L}$ of the Markov process by the relation $P_t = e^{t\mathcal{L}}$. For a fixed discretization parameter $h > 0$, the invariant distribution for the Langevin dynamics remains unchanged under operators $e^{h\gamma\mathcal{L}_{\mathcal{O}}}, e^{\frac{h}{2}\mathcal{L}_{\mathcal{A}}}$ and $e^{\frac{h}{2}\mathcal{L}_{\mathcal{B}}}$ because the latter correspond to Hamiltonian propagation steps and $\mathcal{L}_{\mathcal{O}}$ introduces the combined effect of friction and noise in the system. Thus, splittings in terms of the above operators do not introduce any additional sampling error. More precisely, recall that the invariant distribution μ must satisfy $\mathcal{L}^*\mu = 0$ in terms of the adjoint of the infinitesimal generator. Under mild assumptions, it can be justified that

$$(\mathcal{L}_{\mathcal{A}} + \mathcal{L}_{\mathcal{B}})^* = -(\mathcal{L}_{\mathcal{A}} + \mathcal{L}_{\mathcal{B}}), \quad \mathcal{L}_{\mathcal{O}}^* = \mathcal{L}_{\mathcal{O}}$$

in an appropriate sense, implying the formal adjoint reads

$$\mathcal{L}^* = -(\mathcal{L}_{\mathcal{A}} + \mathcal{L}_{\mathcal{B}}) + \gamma\mathcal{L}_{\mathcal{O}}$$

which can be interpreted as an illustration of the reversal of the momentum in the Hamiltonian term with respect to the invariant measure. One can then either check the definition or note that $P_h^* = P_h^{-1}$, which ensures that the invariant measure is preserved at the continuous level. Splitting-based sampling schemes compose the maps $\mathcal{A}, \mathcal{B}, \mathcal{O}$ in different orders which determine their names, e.g. ABO, which is a scheme that computes the $\mathcal{A}, \mathcal{B}, \mathcal{O}$ propagators in

sequence. As will be discussed, there are many possible orderings that correspond to sampling methods with distinct properties. There is seemingly no consensus (at least in the academic literature) regarding the optimal choice, as instead each sampling method comes with its own benefits and drawbacks.

3.5 First-Order Methods

First-order sampling methods based on operator splitting require no additional assumptions besides the strong convexity and Lipschitz gradient of the potential. These schemes all consist of some permutation of the sequence ABO. Each of the discrete propagator equations are integrated exactly, which in principle should yield some improvement in error bounds as compared with a generic Euler-Maruyama discretization, but in practice require some additional assumptions for any improvement when comparing to a Euler-Maruyama discretization, which is the generic way of approximating a first-order SDE. The Euler-Maruyama discretization for underdamped Langevin reads

$$\begin{cases} p_{k+1} &= p_k + h_{k+1} q_k \\ q_{k+1} &= q_k - h_{k+1}(\nabla U(p_k) + \gamma q_k) + \sqrt{2\gamma h_{k+1}} Z_{k+1} \end{cases}$$

for $\{Z_k\}_{k\in\mathbb{N}} \sim \mathcal{N}(0, I)$. [Cheng et al., 2018] established the canonical results for the Euler-Maruyama discretization under the earlier assumptions on the potential.

Theorem 3.4. Let Φ_t and $\tilde{\Phi}_t$ be the flow maps corresponding to the continuous-time and discrete-time processes, respectively. Let π_0 be a common initial distribution for $(p_t, p_t + q_t)$. Let the step size $0 < h \leq 1$ and $\beta = 1/L$ and $\gamma = 2$. Then the distance between the continuous-time process and the discrete-time process satisfies

$$W_2\left(\Phi_h \pi_0, \tilde{\Phi}_h \pi_0\right) \leq h^2 \sqrt{\frac{2\mathcal{E}_K}{5}} \in \mathcal{O}(h^2\sqrt{d})$$

where $\mathcal{E}_K$ is the kinetic energy/second moment of velocity. Consequently, after $\ell \in \mathbb{N}$ iterations, the long-time error $W_2(\Phi_h^n \rho_0, \rho^*) \in \mathcal{O}(h\sqrt{d})$.

Contraction for Euler-Maruyama has a comparatively simple proof in the framework described earlier, which in many ways generalizes the proof of the above result.

Example 3.5 (Euler-Maruyama Contraction)**.** For two Markov chains in $\mathbb{R}^{2N}$ generated by the Euler-Maruyama discretization, $z_n = (p_n, q_n), \tilde{z}_n = (\tilde{p}_n, \tilde{q}_n)$, let $\overline{z}_n := \tilde{z}_n - z_n$. The Euler-Maruyama method has the update rule

$$\overline{p}_{n+1} = \overline{p}_n + h\overline{q}_n, \quad \overline{q}_{n+1} = \overline{q}_n - \gamma h \overline{q}_n - hQ\overline{p}_n,$$

where

$$Q := \int_0^1 \nabla U(\tilde{p}_n + t(p_n - \tilde{p}_n))dt$$

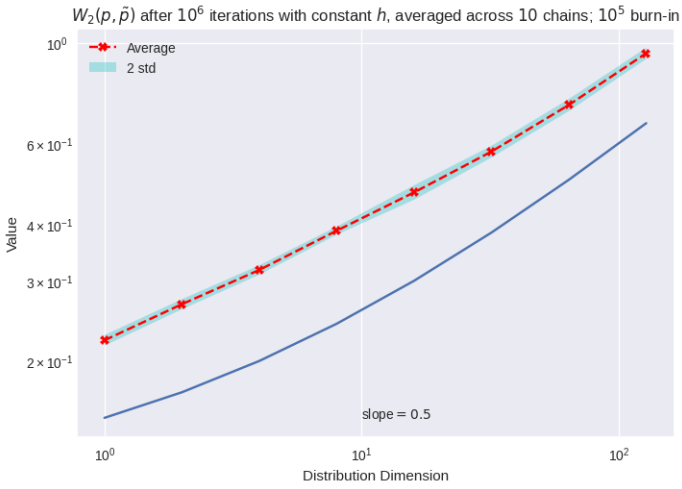
and $\nabla U(\tilde{p}_n) - \nabla U(p_n) = Q\overline{p}$. Then in the earlier notation, the contraction problem involves the matrix

$$P := \begin{bmatrix} I_N & hI_N \\ -hQ & (1-\gamma h)I_N \end{bmatrix}.$$

With the choices $h < 1/2\gamma, a = 1/L, b = 1/\gamma, c(h) = \frac{mh}{2\gamma}$, the contraction property then requires showing that $H := (1-c(h))G - P^T G P$ is positive definite, which is a discrete analogue of the matrix considered in the earlier proof for continuous contraction. The benefit of this proof strategy is that the eigenvalues of P determine the admissible range of h for which the discrete chains have a contraction property.

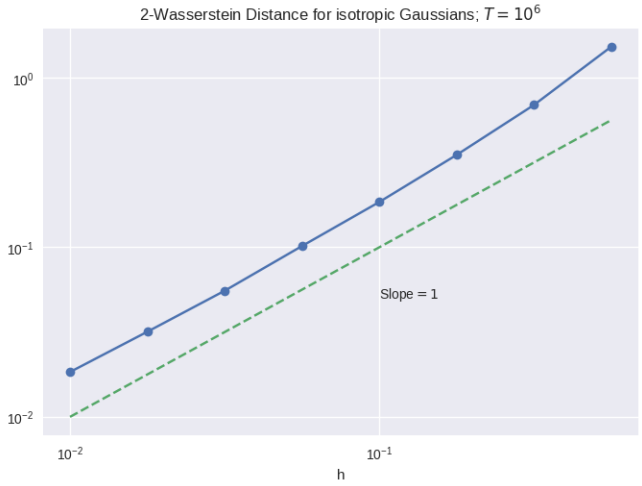
Example 3.6 (Sampling Gaussian Potentials)**.** The salient parts to verify numerically in the above asymptotic rates are the dependence on h and the dimension d. With the earlier explicit form for the Wasserstein distance between two Gaussians, we can compute the Wasserstein distance between a sampled distribution after many iterations against the true invariant distribution to confirm these rates with varying time-steps and dimension.

The first set of examples concerns isotropic Gaussian potentials, which evidently satisfy the conditions under which the earlier results are proven. The mean vector is initialized as the $2N$ dimensional vector $m = \langle 1, 1, \ldots, 0, 0, \ldots, 0\rangle$ half filled with 1 (corresponding to positions) and 0s otherwise (corresponding to momenta). The covariance matrix is $I_{2N\times 2N}$. This may be the simplest possible example, but served as validation that first implementations were correct. For the following results, the dimension N was varied with fixed $h = 5 \cdot 10^{-2}$. Initial momenta were sampled at random for 10 sampling chains, each of which computed 10^6 samples following the Euler-Maruyama discretization. Subsequently, the mean and covariance of the sampled empirical distribution were computed and the analytical expression for the Wasserstein distance was evaluated between the true Gaussian invariant distribution and the empirical distribution.



The corresponding figure shows the average Wasserstein distance two standard deviations across the 10 chains for each dimension $N = 2^k, k = 0, \ldots, 7$. We observe that the predicted

asymptotic convergence rate on the order of $\sqrt{N}$ is satisfied and that the variability in sampling error remains the same proportional to the mean which increases with the dimension, owing to the fact that all dimensions of the distribution are independent. We also verify the first-order step size asymptotic rate by sampling 4D isotropic Gaussians across a range of step sizes that generate chains with 10^6 samples each. All other parameters of the dynamics are the same as in the first example.



The Euler-Maruyama discretization corresponds to the same discrete updates as those of the ABO splitting scheme, but with the crucial difference that the OU step is integrated exactly for ABO by the exact solution formula

$$q \leftarrow q - \gamma q h + \sqrt{2\gamma h} G$$

performed by $\mathcal{O}$. Nevertheless, under the same assumptions on the potential, [Leimkuhler et al., 2024] show that not only ABO, but all schemes corresponding to the three-letter permutations, all have the same asymptotics in terms of h and d as the Euler-Maruyama discretization, though with different discrete contraction pre-constants that do not influence the asymptotic error against the invariant distribution.

3.6 Higher-Order Splitting Schemes

Higher-order sampling schemes share a common palindromic form. For a small $\delta > 0$, the second-order approximation of $e^{\delta(a+b)}$ results in many commutator terms involving the splitting operators. Symmetrizing the expansion, (e.g. the sequence of propagators OBABO) ensures that odd order terms vanish identically by repeated applications of the Jacobi identity. Heuristically, this implies that such symmetric schemes result in second-order errors. For instance, the OBABO expansion corresponds to

$$P_\delta = e^{\delta/2\mathcal{L}_\mathcal{O}} e^{\delta/2\mathcal{L}_\mathcal{B}} e^{\delta\mathcal{L}_\mathcal{A}} e^{\delta/2\mathcal{L}_\mathcal{B}} e^{\delta/2\mathcal{L}_\mathcal{O}} + o(\delta^2)$$

where the first term determines the transition kernel of the sampler. Discretely, this corresponds to the operations

$$\begin{aligned} v_0' &= \eta v_0 + \sqrt{1-\eta^2} G \\ v_{1/2} &= v_0' - \delta/2 \nabla U\left(x_0\right) \\ x_1 &= x_0 + \delta v_{1/2} \\ v_1' &= v_{1/2} - \delta/2 \nabla U\left(x_1\right) \\ v_1 &= \eta v_1' + \sqrt{1-\eta^2} G' \end{aligned}$$

where $\eta := \exp(-\delta\gamma/2)$. Formally, the the second-order transition kernel approximation transfers to a second-order approximation of the transition semigroup of the continuous process and may be expected to result in the same order of error of an empirical distribution against the continuous invariant measure. This is partly owing to the symmetry of the integration steps, which imply that schemes with palindromic names such as OBABO or BAOAB have a weak second order of accuracy, and thus the bias in long-time averages is $\mathcal{O}(\delta^2)$. That is, an expansion of the invariant distribution in terms of the small parameter δ results in the asymptotics

$$\rho_t \propto \rho_*(1 - \delta^2 f(p,q) + O(\delta^4)).$$

On the other hand, similar expansions show that first-order schemes such as ABO result in weak first-order approximations to the invariant measure. The components of higher-order splitting schemes also admit structural benefits. For instance, the sequence of propagators BAB correspond to the velocity Verlet integrator used for integrating the Hamiltonian equations. As such, the OBABO scheme is at its core a deterministic, time reversible symplectic integrator. As such, proofs of Wasserstein contraction are comparatively simpler than for other symmetric schemes such as BAOAB.

Under the same assumptions as earlier, however, the benefits of second-order approximation are not visible in the asymptotic rates. For instance, in [Monmarché, 2021], the OBABO scheme is shown to only have asymptotic convergence rates $W_2(\rho_t, \rho^*) \sim h\sqrt{d}$ and the asymptotic analysis of the bias for the BAOAB scheme in [Leimkuhler et al., 2024] with the same conditions results in the same asymptotic rates. Neither of these results are an improvement over the convergence guarantees of the Euler-Maruyama discretization in general despite the additional structure in these samplers. Conditions on higher derivatives of the potential are necessary, the simplest assumption being that the Hessian of the potential is Lipschitz. Yet [Monmarché, 2021] presents weaker assumptions which recover the expected second-order convergence guarantees.

Definition 3.7. The potential U is said to satisfy the $(\nabla^2\mathrm{pol}(\ell))$ condition if there exist $\ell \geqslant 2$, $\mathrm{L}_\ell > 0$ and $x_\star \in \mathbb{R}^d$ such that, for all $x, y \in \mathbb{R}^d$,

$$|\nabla^2 U(y) - \nabla^2 U(x)| \leqslant L_\ell |x-y| \left(2 + |x-x_\star|^{\ell-2} + |y-x_\star|^{\ell-2}\right)/4.$$

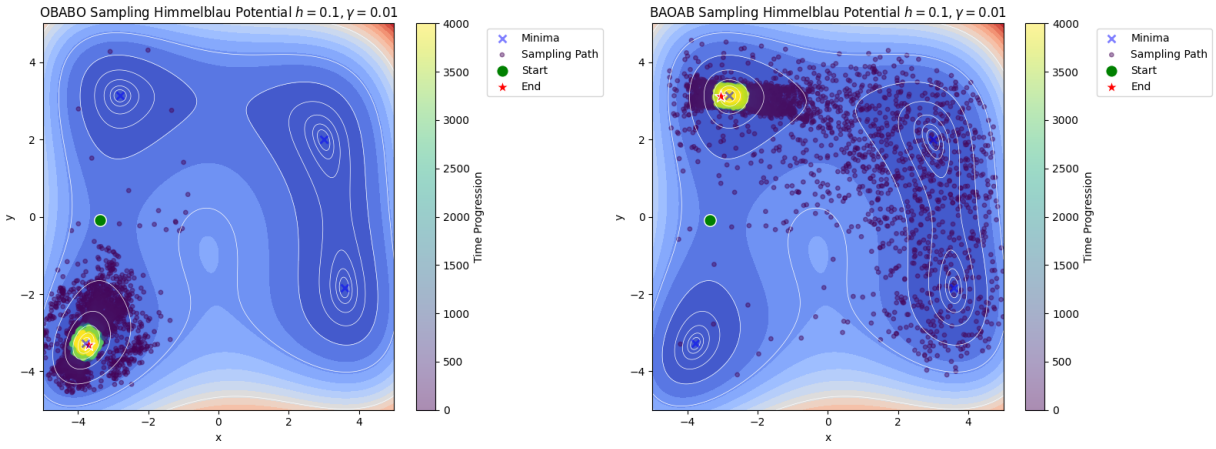
This condition always holds for Gaussian potentials under any parameters because the Hessian is constant. A direct computation shows that the inequality is sharp for the 2D

cubic potential $U(x) = \frac{1}{6}(x_1^3 + x_2^3)$ with $\ell = 2, L_\ell = 1$. This condition can be satisfied if $||\nabla^{(k)} U||_\infty \leq \infty$ for some $k \geq 3$ when considering a Taylor expansion. For potentials satisfying this assumption in addition to earlier ones, the asymptotic rates are improved to $W_2(\rho_t, \rho^*) \sim h^2 d^{\ell/2}$ in [Monmarché, 2021]. Besides the improved order of convergence, the distinct properties of each sampler are of interest to practitioners.

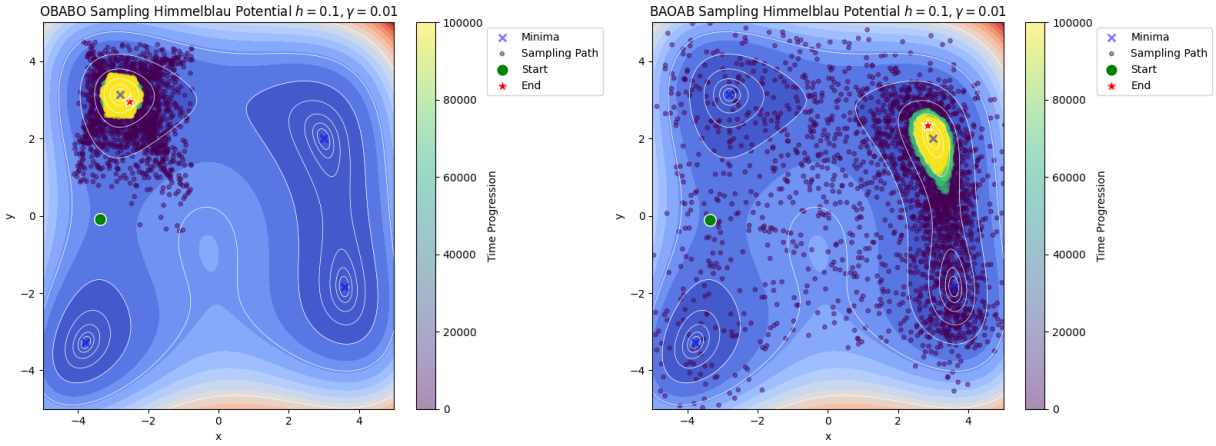
Example 3.8 (Transitions between Isolated/Metastable States)**.** The presence of the BAB component in OBABO implies the positive quality that the discrete dynamics are propagated in a Hamiltonian-like fashion; i.e. energy is conserved in the absence of the stochastic O steps. However, for potential landscapes which require barrier crossing (e.g. between minima connected by a narrow region), as frequently encountered in metastable systems, the placement of the O steps complicate transitions. Because the O steps add momentum noise before and after the force updates, each O step is not directly informed by the latest potential gradient, which reduces the ability for the sampling trajectory to leave a landscape's minima. We illustrate this intuition on the 2D Himmelblau potential

$$U(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2.$$

With $\gamma = 0.01, h = 0.1, T = 4000$, we visualize the sampling trajectories of OBABO and BAOAB as they progress.



We observe that the OBABO trajectories hardly explore the potential landscape. As instead, the trajectories remain trapped around a single minima which they do not leave. The situation is not improved by a more than tenfold increase in the length of the chain, as evidenced in the following figure.



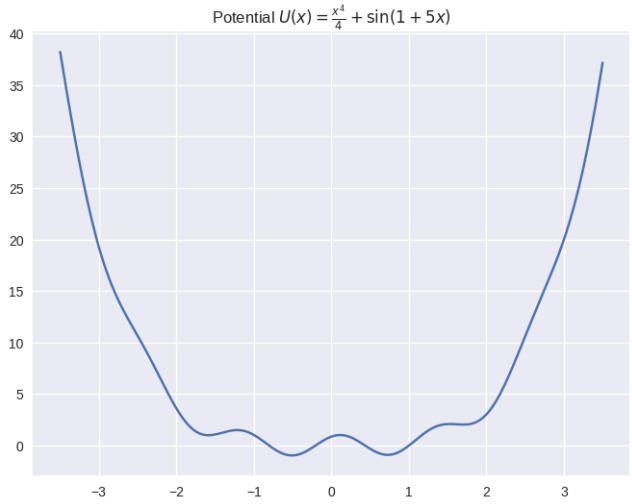
While BAOAB can be argued to also perform poorly, it's clear that the sampling trajectory explores considerably more of the distribution and can successfully transition through the narrow barrier between two minima at least at some points during sampling. Besides illustrating intuition regarding the OBABO sampler, this example is also suggestive of the fact that *a priori*, estimating the error of a sampled distribution should not be performed by measuring properties of the discrete chain between successive iterations (a "Cauchy-like" error estimate) as samples may cluster in a way that is not representative of the true distribution whatsoever.

Remark 3.9. The metastable potential in the preceding example is illustrative of one of many qualitative challenges faced by sampling problems in molecular dynamics. Namely, stochastic processes that govern microscopic behaviors can be metastable, meaning the solution trajectories may remain localized within a small region of the potential landscape for a long time before suddenly transitioning to a different metastable region. Predicting when and how this occurs is in general not possible and typically these transitions between metastable regions are rarely observed in numerical simulations.

3.7 Superconvergence of BAOAB

The BAOAB scheme is distinguished among second-order splitting schemes by its behavior in the high-friction limit $\gamma \to +\infty$. By viewing both δ and $\varepsilon := \gamma^{-1}$ as small parameters for which the invariant distribution is simultaneously expanded with cancellation of many terms owing to the ordering of the propagators, BAOAB is shown in [Leimkuhler & Matthews, 2012] to sample the invariant distribution at fourth-order in the high-friction regime across all step sizes. Cancellations of this type do not occur for any other higher-order splitting scheme.

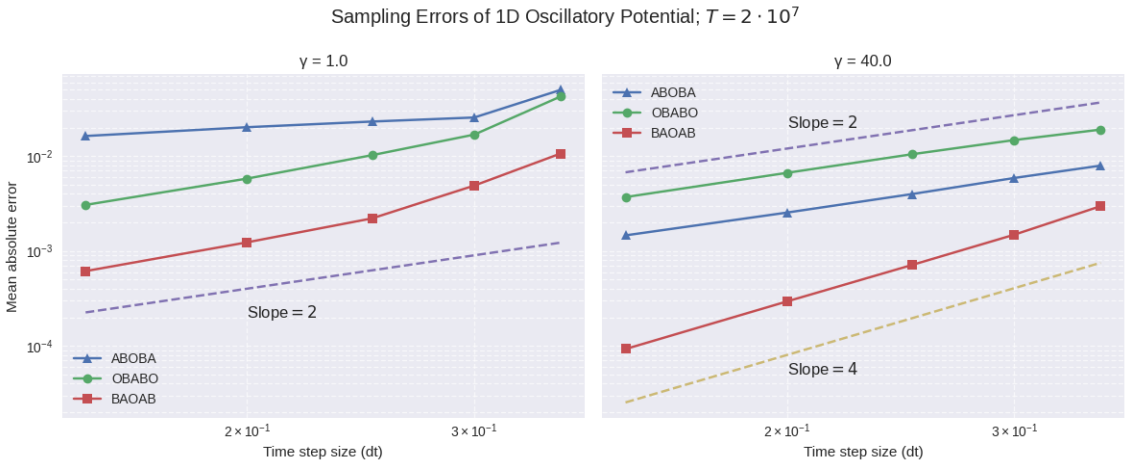
Example 3.10. To illustrate the fourth order superconvergence of the BAOAB scheme, consider the potential $U(x) = x^4/4 + \sin(1 + 5x)$, visualized below.



Owing to the fact that the Wasserstein distance cannot be computed for this potential, a cruder error estimate is used. M intervals of equal length are defined over the sampling domain, and the mean error is computed between the observed probability frequency compared to the exact expected frequency obtained by integrating the probability density function. Denoting f_i as the frequency in interval i and the expected frequency as $\hat{f}_i$,

$$\text{error} = \frac{1}{M} \sum_{i=1}^{M} |f_i - \hat{f}_i|.$$

In this 1D example, the potential is defined over the interval $x \in [-3.5, 3.5]$ which is covered by 20 equispaced intervals. The results of three schemes for a low friction regime with $\gamma = 1$ and a higher friction regime $\gamma = 40$ are illustrated in the following figure.



As predicted by the asymptotic results of the paper, BAOAB indeed enjoys a fourth order superconvergence in sampling error while other schemes retain the same second-order convergence. This result has practical implications for step size selection. In the high-friction limit,

the BAOAB equations become a scheme for sampling from the overdamped Langevin dynamics by the update rule

$$p_{n+1} = p_n - \frac{h^2}{2}\nabla U(p_n) + \frac{h}{2}(Z_n + Z_{n+1}).$$

While the contraction results in [Leimkuhler et al., 2024] place restrictions on the admissible step sizes h to ensure contraction in terms of the friction of the system for the underdamped dynamics, these are greatly simplified.

4 Conclusions and Outlook

This expository note only scratched the surface of the theory and applications of Langevin-based sampling methods. Along the way, the focus has been on established patterns which in one way or another leverage the structure of the Langevin equation to obtain the currently best known results, rather than on generic tools which may apply broadly but do not reveal anything more about the structure of these equations. The analysis of the continuous-time process for underdamped Langevin was shown to require specialized theory as a consequence of the infinitesimal generator of the Markov process not being uniformly elliptic. This fact motivated the definition of weighted Euclidean norms in which contraction results can be proven. For the purposes of sampling a distribution, we briefly presented a framework by which (1) contraction proofs are transformed to showing properties of certain matrices and (2) by which contraction between synchronously coupled discrete-time chains in the weighted Euclidean norm leads to rates of convergence for the Wasserstein distance between discrete and continuous processes' invariant distributions. These techniques work under a common set of assumptions on the potential, though additional assumptions are necessary to guarantee higher-order convergence for schemes competing with the generic Euler-Maruyama discretization. The design of these higher-order methods arises from exploiting the structure of the infinitesimal generator and splitting it into simpler operators for which the discrete counterparts correspond to exact integration. As the sampling error can primarily be controlled by reducing the systemic error induced by a discretization, these splitting schemes are the primary way by which sampling error can be reduced in a principled way. However, operator splitting results in a variety of admissible sampling schemes, each of which requires a separate analysis and has distinct properties. Some of these were illustrated by simple examples: the BAOAB sampler improves in high-friction regimes to fourth order, the OBABO sampler's structure, despite being at its core the deterministic velocity Verlet integrator used in HMC, is vulnerable to localization effects because of the delayed noise and velocity updates.

Several topics were regrettably excluded from this write-up. First, all of the presented results, proofs, and examples featured only constant step-size discretizations. In practice, however, one would hope that as with deterministic time-steppers that adapt the time step, integrators for the underdamped Langevin equations would do the same. But while there exists an extensive theory for adaptive time-stepping for deterministic dynamics, this is a relatively open area for research, especially for schemes based on operator splitting. The results

in [Leroy et al., 2024] concern adaptive time stepping for Euler-Maruyama discretizations by introducing a class of *monitor functions* that automatically increase the number of samples in a discrete chain in regions of high solution change by decreasing the time step while accelerating the process in other regions by increasing the time step. However, as the authors state themselves, convergence analysis is needed for the resulting schemes before they are more widely adopted.

While convergence results in terms of step size and dimension are useful and lead to estimates of the number of discrete time steps necessary to reach a prescribed tolerance, they give no practical avenue by which to assess the quality of a sampled distribution in practice. As mentioned in the metastable potential examples, merely estimating differences between results of successive time steps is not sufficient, as Langevin dynamics may end up localized in some regions of the potential landscape, then infrequently and unpredictably transition to other regions. To this end, one would like an estimator of the Wasserstein distance between a sampled empirical distribution and the true unknown distribution that can be evaluated efficiently. To this end, [Niles-Weed & Rigollet, 2022] presented such an estimator which cannot in practice be evaluated. Further results on the path towards an estimator that can be used in practice were proven in [Niles-Weed & Berthet, 2022]. The results of the second paper in particular confirm some of my suspicions that this question is related to width theory.

As alluded to at a few points in this note (e.g. a Poincaré inequality sufficing for exponential contraction), the assumptions on the potential under which the results were presented remained essentially unchanged throughout. Some recent research with deeper connections to functional inequalities, optimal transport, and the geometry of the process focuses on vastly different assumptions on the potential. Sinho Chewi's works are on my summer reading list.

References

[Bai & Chandra, 2023] Bai, G. & Chandra, R. (2023). Gradient boosting bayesian neural networks via langevin mcmc. *Neurocomputing*, 558, 126726.

[Betancourt, 2018] Betancourt, M. (2018). A conceptual introduction to hamiltonian monte carlo.

[Chen et al., 2014] Chen, T., Fox, E., & Guestrin, C. (2014). Stochastic gradient hamiltonian monte carlo. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research* (pp. 1683–1691). Bejing, China: PMLR.

[Cheng et al., 2018] Cheng, X., Chatterji, N. S., Bartlett, P. L., & Jordan, M. I. (2018). Underdamped langevin mcmc: A non-asymptotic analysis. In S. Bubeck, V. Perchet, & P. Rigollet (Eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research* (pp. 300–323).: PMLR.

[Figalli & Glaudo, 2023] Figalli, A. & Glaudo, F. (2023). *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows: Second Edition*. European Mathematical Society.

[Leimkuhler & Matthews, 2012] Leimkuhler, B. & Matthews, C. (2012). Rational construction of stochastic numerical methods for molecular sampling. *Applied Mathematics Research eXpress*.

[Leimkuhler & Matthews, 2015] Leimkuhler, B. & Matthews, C. (2015). *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*. Springer.

[Leimkuhler et al., 2024] Leimkuhler, B. J., Paulin, D., & Whalley, P. A. (2024). Contraction and convergence rates for discretized kinetic langevin dynamics. *SIAM Journal on Numerical Analysis*, 62(3), 1226–1258.

[Leroy et al., 2024] Leroy, A., Leimkuhler, B., Latz, J., & Higham, D. J. (2024). Adaptive stepsize algorithms for langevin dynamics. *SIAM Journal on Scientific Computing*, 46(6), A3574–A3598.

[McLachlan & Quispel, 2002] McLachlan, R. I. & Quispel, G. R. W. (2002). *Splitting methods*, (pp. 341–434). Cambridge University Press.

[Monmarché, 2021] Monmarché, P. (2021). High-dimensional mcmc with a standard splitting scheme for the underdamped langevin diffusion. *Electronic Journal of Statistics*, 15(2).

[Niles-Weed & Berthet, 2022] Niles-Weed, J. & Berthet, Q. (2022). Minimax estimation of smooth densities in wasserstein distance. *The Annals of Statistics*, 50(3).

[Niles-Weed & Rigollet, 2022] Niles-Weed, J. & Rigollet, P. (2022). Estimation of wasserstein distances in the spiked transport model. *Bernoulli*, 28(4).

[Pavliotis, 2014] Pavliotis, G. A. (2014). *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*. Springer.

[Roberts & Tweedie, 1996] Roberts, G. O. & Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4), 341.

[Vempala & Wibisono, 2023] Vempala, S. S. & Wibisono, A. (2023). *Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices*, (pp. 381–438). Springer International Publishing: Cham.

[Wibisono, 2018] Wibisono, A. (2018). Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In S. Bubeck, V. Perchet, & P. Rigollet (Eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research* (pp. 2093–3027).: PMLR.