ACCELERATED REGULARIZED WASSERSTEIN PROXIMAL SAMPLING ALGORITHMS

HONG YE TAN*, STANLEY OSHER, AND WUCHEN LI

ABSTRACT. We consider sampling from a Gibbs distribution by evolving a finite number of particles using a particular score estimator rather than Brownian motion. To accelerate the particles, we consider a second-order score-based ODE, similar to Nesterov acceleration. In contrast to traditional kernel density score estimation, we use the recently proposed regularized Wasserstein proximal method, yielding the Accelerated Regularized Wasserstein Proximal method (ARWP). We provide a detailed analysis of continuous- and discrete-time non-asymptotic and asymptotic mixing rates for Gaussian initial and target distributions, using techniques from Euclidean acceleration and accelerated information gradients. Compared with the kinetic Langevin sampling algorithm, the proposed algorithm exhibits a higher contraction rate in the asymptotic time regime. Numerical experiments are conducted across various low-dimensional experiments, including multi-modal Gaussian mixtures and ill-conditioned Rosenbrock distributions. ARWP exhibits structured and convergent particles, accelerated discrete-time mixing, and faster tail exploration than the non-accelerated regularized Wasserstein proximal method and kinetic Langevin methods. Additionally, ARWP particles exhibit better generalization properties for some non-log-concave Bayesian neural network tasks.

1. INTRODUCTION

Let $V : \mathbb{R}^d \to \mathbb{R}$ be a known $\mathcal{C}^1$ potential function that typically satisfies some growth condition at infinity. The problem is to design an algorithm for sampling from target Gibbs distributions with densities

$$\pi(x) \propto \exp(-\beta V(x)),$$

where $\beta > 0$ is some diffusion/temperature parameter. Such tasks occur frequently in data science, such as uncertainty quantification and physical modeling [1], or more recently in generative modeling using diffusion models [13].

Traditional methods, such as Markov chain Monte Carlo (MCMC) methods, apply Markov chains with an invariant distribution π. MCMC methods usually arise from discretizations of stochastic differential equations (SDEs), which evolves a density according to a Fokker–Planck equation

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\nabla V(x)\rho) + \beta^{-1} \Delta \rho, \quad \rho(x, 0) = \rho_0(x). \tag{1}$$

Standard examples include the unadjusted Langevin algorithm (ULA) and the Metropolis-adjusted Langevin algorithm (MALA) [32, 14]. These algorithms arise from particular discretizations or approximations of the (overdamped) Langevin equation

$$\mathrm{d}X_t = -\nabla V(X)\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}W\,. \tag{2}$$

It can be shown that the density of particles evolving under (2) satisfies the Fokker–Planck equation (1). Traditionally, accelerating overdamped Langevin dynamics yields the underdamped Langevin equation, sometimes known as kinetic Langevin dynamics. While the invariant distribution of the overdamped Langevin equation is the target Gibbs distribution π, the invariant distribution of the underdamped dynamics is a separable joint density in a position-momentum phase space. Some methods arising from discretizing the underdamped Langevin dynamics include the variational

acceleration flow [8], inertial Langevin algorithm [16], kinetic Langevin Monte Carlo [12, 10], or some splitting methods, such as the OBA or BAOAB methods [21, 22]. Generalizations of kinetic Langevin dynamics include the Hessian-free high resolution dynamics [23], and primal-dual damping stochastic dynamics [42]. Convergence results for the kinetic Langevin equation are well studied, including non-asymptotic Wasserstein-2 contraction in continuous time and under different time discretizations [12, 22, 10, 27]. Recently, [5] demonstrates L^2 convergence under a Poincaré inequality, and provides an optimal friction coefficient. Related flows for sampling from phase-space include Hamiltonian Monte Carlo methods [9].

As opposed to using a discretized SDE, another sampling paradigm approximates the Fokker–Planck equation by evolving a finite collection of particles through the score-based ODE

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = -\nabla V(X) - \beta^{-1}\nabla \log \rho_t(X),\tag{3}$$

where ρ_t is the density of X_t at time t. From the continuity equation, the density of particles evolving according to this ODE (3) also follows the Fokker–Planck equation (1). However, the score function $\nabla \log \rho_t$ is often intractable and requires estimation from empirical distributions. To derive algorithms that work with finitely many particles, (3) needs to be modified with kernel functions. Examples of score-based sampling methods using kernels include Stein variational gradient descent (SVGD) [25], which performs steepest descent with respect to the KL divergence with respect to a Wasserstein-type metric structure on the space of distributions induced by the Stein operator with a kernel function. Another example is the blob method [7, 11], which considers Wasserstein gradient flows with particular kernel regularizations of the energy functional.

To accelerate the score-based flow (3), one may consider adding a momentum variable, similar to Nesterov acceleration methods for classical optimization problems [29, 35]. One possible acceleration to sample from a distribution comes from a particular Hamiltonian evolution [41, 36, 8, 6]. In the particular case of the Wasserstein-2 metric, the accelerated flow to minimize the KL divergence can be written as coupled ODEs in the density ρ_t and its momentum variable. By adding a damping term into the momentum equation, [41, 36] derive the *accelerated information gradient flow*. The particle evolutions take the following form, which can be viewed as a second-order dynamics of the original score-based ODE (3):

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} P \\ -aP - \nabla V(X) - \nabla \log \rho_t(X) \end{bmatrix}.\tag{4}$$

We study an equation-level modification of equation (4). In particular, following [37], we approximate the intractable score $\nabla \log \rho_t$ with a tractable approximation $\nabla \log \mathrm{WProx}\,\rho_t$, where WProx is the *regularized Wasserstein proximal operator* (RWPO), defined in Definition 2 in the following section as the solution of a mean field control problem. This allows us to compute the score approximation without relying on selecting an appropriate kernel, such as the Gaussian kernel density estimator used in [36, 41] to approximate $\log \rho_t$. The algorithm takes the form of a particular discretization of the particle evolution

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} P \\ -aP - \nabla V(X) - \nabla \log \mathrm{WProx}\,\rho_t(X) \end{bmatrix}.\tag{5}$$

1.1. **Contributions.** Focusing on the accelerated flow (5), this work is organized as follows:

- Section 2 details the regularized Wasserstein proximal operator of [24], and its links to regularizing the Fokker–Planck equation (1). Moreover, we recall the space-varying kernel representation of this operator, which will be used in future computations.
- Section 3 introduces the *accelerated regularized Wasserstein proximal (ARWP) method*. By performing a particular symplectic time discretization of the flow (5), we derive a discrete-time interacting particle algorithm to evolve the positions and velocities using the RWPO.
- Using the Gaussian closure property of the RWPO, Section 4 analyzes the convergence of the ARWP method in the case of Gaussian initial and target distributions. Using closed-form updates, we provide an asymptotic continuous and discrete-time mixing analysis through linearization. Moreover, a detailed Lyapunov analysis demonstrates convergence

in continuous time, where the damping parameter is sufficiently large to satisfy standard assumptions (up to modification by the regularization parameter).

- Section 5 verifies the convergence analysis, and tests the ARWP method against various baselines. This includes a Rosenbrock distribution to identify tail exploration, a multi-modal Gaussian mixture to test mixing times, and a Bayesian neural network example for simulations in higher dimensions.

Additional background on accelerated probability flows, proofs of Section 4, and additional qualitative/quantitative results, including hyperparameter ablations and choices, are also provided in the appendix.

2. REGULARIZED WASSERSTEIN PROXIMAL

We begin with the definition of the Wasserstein-2 distance and the Wasserstein proximal.

Definition 1 ([34, 2]). *Let $\mathcal{P}_2(\mathbb{R}^d)$ be the set of probability densities with finite second moment. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the* Wasserstein-2 *distance $\mathcal{W}_2(\mu, \nu)$ is*

$$\mathcal{W}_2(\mu, \nu) = \inf_{\pi \in \Gamma(\mu, \nu)} \int \|x - y\|^2 \mathrm{d}\pi(x, y), \tag{6}$$

where the infimum is taken over couplings $\pi \in \Gamma(\mu, \nu)$, i.e. probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ satisfying

$$\int_{\mathbb{R}^d} \pi(x, y) \, \mathrm{d}y = \mu(x), \quad \int_{\mathbb{R}^d} \pi(x, y) \, \mathrm{d}x = \nu(y). \tag{7}$$

Consider a probability density $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and $V \in \mathcal{C}^1(\mathbb{R}^d)$ be a lower-bounded potential function. For a scalar $T > 0$, the Wasserstein proximal *of ρ_0 is defined as*

$$\mathrm{WProx}_{T, V}(\rho_0) := \arg\min_{q \in \mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} V(x) q(x) \, \mathrm{d}x + \frac{\mathcal{W}(\rho_0, q)^2}{2T}. \tag{8}$$

A recent line of work considers a principled score estimator called the *backwards regularized Wasserstein proximal* (BRWP) method [37]. This method utilizes the fact that the time-discretized score-based particle update (3) corresponds to a time evolution of a modified Fokker–Planck equation, which can be computed using a kernel formula. This has been extended to utilize tensor train score approximations [18] and splitting methods [17]. A recent followup work incorporates a preconditioning matrix into the underlying Fokker–Planck approximation, resulting in a modified mean-field control problem and different kernel [38]. In this section, we recall the definition of the regularized Wasserstein proximal, which is used in the proposed scheme from discretizing (5).

The regularized Wasserstein proximal was first proposed as an approximation to a mean field control problem, obtained from the Wasserstein proximal through the Benamou–Brenier formulation [24, 4]. In particular, the variational form (8) has two equivalent formulations. One of them is the mean field control (MFC) formulation, where the Wasserstein proximal is given by the terminal time solution ρ_T to the following MFC:

$$\inf_{\rho, v, q} \int_0^T \int_{\mathbb{R}^d} \frac{1}{2} \|v(t, x)\|^2 \rho(t, x) \, \mathrm{d}x \, \mathrm{d}t + \int_{\mathbb{R}^d} V(x) q(x) \, \mathrm{d}x, \tag{9a}$$

$$\partial_t \rho(t, x) + \nabla \cdot (\rho(t, x) v(t, x)) = 0, \quad \rho(0, x) = \rho_0(x), \rho(T, x) = q(x). \tag{9b}$$

The minimization (9a) is taken jointly with respect to a time-varying density function ρ, a velocity field v, and the terminal density function q. The conditions in (9b) are a continuity equation and some boundary conditions, coupling the density and the velocity field. The Benamou–Brenier formulation then states that the solution to this minimization problem is equivalently given by the terminal time solution $\rho(T, x)$ of the following set of coupled PDEs,

$$\begin{cases} \partial_t \rho(t, x) + \nabla_x \cdot (\rho(t, x) \nabla_x \Phi(t, x)) = 0, \\ \partial_t \Phi(t, x) + \frac{1}{2} \|\nabla_x \Phi(t, x)\|^2 = 0, \\ \rho(0, x) = \rho_0(x), \quad \Phi(T, x) = -V(x). \end{cases} \tag{10}$$

We can now define the regularized Wasserstein proximal by adding some appropriate diffusive terms to the MFC, with an equivalent definition in terms of a terminal time solution to some coupled PDEs.

Definition 2. *For a probability distribution $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and a diffusion parameter $\beta > 0$, the regularized Wasserstein proximal* $\mathrm{WProx}_{T,V}(\rho_0)$ *is given by the solution to the following MFC problem:*

$$\inf_{\rho,v,q} \int_0^T \int_{\mathbb{R}^d} \frac{1}{2}\|v(t,x)\|^2 \rho(t,x)\,\mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}^d} V(x)q(x)\,\mathrm{d}x\,, \tag{11a}$$

$$\partial_t \rho(t,x) + \nabla \cdot (\rho(t,x)v(t,x)) = \beta^{-1}\Delta\rho(t,x),$$
$$\rho(0,x) = \rho_0(x),\ \rho(T,x) = q(x), \tag{11b}$$

The RWPO is defined as follows:

$$\mathrm{WProx}_{T,V}\,\rho_0 := \rho_T.$$

The above MFC problem (11a) *is a modification of* (9a)*, where the constraint* (11b) *is with a Laplacian term in the continuity equation* (9b).

Solving the optimality conditions yields that the regularized Wasserstein proximal satisfies a similar regularized Benamou–Brenier formulation, given by Laplacian regularization in both the forward-time Fokker–Planck and backward-time Hamilton–Jacobi equations,

$$\begin{cases} \partial_t \rho(t,x) + \nabla_x \cdot (\rho(t,x)\nabla_x\Phi(t,x)) = \beta^{-1}\Delta_x\rho(t,x), & \text{(12a)} \\[2mm] \partial_t \Phi(t,x) + \frac{1}{2}\|\nabla_x\Phi(t,x)\|^2 = -\beta^{-1}\Delta_x\Phi(t,x), & \text{(12b)} \\[2mm] \rho(0,x) = \rho_0(x), \quad \Phi(T,x) = -V(x). & \text{(12c)} \end{cases}$$

Using a particular Cole–Hopf formula connects these coupled PDEs with a set of forward-backward coupled heat equations, which has an exact solution based on a kernel formula:

$$\rho_T(x) = \rho(T,x) = \int_{\mathbb{R}^d} K(x,y)\rho_0(y)\,\mathrm{d}y\,, \tag{13a}$$

$$K(x,y) = \frac{\exp\left(-\frac{\beta}{2}(V(x) + \frac{\|x-y\|^2}{2T})\right)}{\int_{\mathbb{R}^d} \exp\left(-\frac{\beta}{2}(V(z) + \frac{\|z-y\|^2}{2T})\right)\mathrm{d}z}. \tag{13b}$$

Motivated by the approximation of the Wasserstein proximal and the iterative component of the JKO scheme [20], [37] propose the *Backwards Regularized Wasserstein Proximal* (BRWP) method. This is a semi-implicit discretization of the regularized Fokker–Planck equation. In particular, by the continuity equation, the Fokker–Planck equation (12a)

$$\partial_t \rho(t,x) + \nabla \cdot \big(\rho(t,x)\nabla\Phi(t,x) - \beta^{-1}\rho\nabla(\log\rho)(t,x)\big) = 0, \tag{14}$$

corresponds to particles evolving as

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \nabla\Phi(t,X) - \beta^{-1}\nabla\log\rho(t,X). \tag{15}$$

By discretizing the particle-based updates (15) using the backward Euler method, the dual function Φ simply becomes an update in V due to the boundary conditions of the MFC (12c). Moreover, the score term $\nabla\log\rho(T,x)$ is precisely given by the score of the regularized Wasserstein proximal, which is computable for an empirical distribution given by a collection of particles. The BRWP iterations can be written as

$$X_{k+1} = X_k - \eta\Big(\nabla V(X_k) + \beta^{-1}\nabla\log\mathrm{WProx}_{T,V}\,\rho_k(X_k)\Big). \tag{16}$$

This allows for an adaptive kernel-based modification of the score-based update (3), replacing an arbitrary choice of kernel function (such as Gaussian or Matérn), with a choice of regularization parameter T.

3. ACCELERATED REGULARIZED WASSERSTEIN PROXIMAL METHOD

The BRWP method replaces the log density $\log \rho$ with the log density of the regularized Wasserstein proximal $\log \mathrm{WProx}_{T,V} \rho$, and provides a discrete time update. We propose the *accelerated regularized Wasserstein proximal* (ARWP) method in continuous and discrete time, which arises from using the approximation $\log \rho \approx \log \mathrm{WProx}_{T,V} \rho$ within (4). For a particle with position X and momentum P, we recall the iteration (5):

$$\begin{cases} \dfrac{\mathrm{d}X}{\mathrm{d}t} = P, & \text{(17a)} \\[2mm] \dfrac{\mathrm{d}P}{\mathrm{d}t} = -aP - \nabla V(X) - \beta^{-1} \nabla \log \mathrm{WProx}_{T,V} \rho(t, X), & \text{(17b)} \end{cases}$$

where $\rho(t, \cdot)$ denotes the distribution of particles at time t. The discrete time update for a finite set of particles is given as follows. In the finite particle setting with positions and momentums $\{(\mathbf{x}_i^{(k)}, \mathbf{p}_i^{(k)})\}_{i=1}^N$, the semi-implicit scheme is given by setting ρ to be the empirical distribution at each iteration. The discrete-time ARWP algorithm for a step-size $\eta > 0$ and possibly-varying damping parameters $a_k > 0$ is defined by updating the particle positions and momenta using the symplectic (semi-implicit) Euler discretization:

$$\begin{cases} \mathbf{p}_i^{(k+1)} = (1 - \eta a_k)\mathbf{p}_i^{(k)} - \eta \nabla V(\mathbf{x}_i^{(k)}) - \eta \beta^{-1} \nabla \log \mathrm{WProx}_{T,V} \rho_k(\mathbf{x}_i^{(k)}), \\ \mathbf{x}_i^{(k+1)} = \mathbf{x}_i^{(k)} + \eta \mathbf{p}_i^{(k+1)}. \end{cases} \quad \text{(18)}$$

For each point, we can compute the score function of the RWPO of the empirical distribution $\rho_k = \frac{1}{N} \sum_{i=1}^N \delta(\mathbf{x}_i^{(k)})$ using the kernel formula (13) [37], where $\delta(\mathbf{x})$ denotes the Dirac delta at the point $\mathbf{x}$. Temporarily dropping the iteration k subscripts and superscripts, the RWPO of the empirical distribution can be computed at each point as follows:

$$\mathrm{WProx}_{T,V} \rho(\mathbf{x}_i) = \frac{1}{N} \sum_{j=1}^N K(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{N} \sum_{j=1}^N \frac{\exp\left[-\frac{\beta}{2}\left(V(\mathbf{x}_i) + \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2T}\right)\right]}{\mathcal{Z}(\mathbf{x}_j)},$$

$$\nabla \mathrm{WProx}_{T,V} \rho(\mathbf{x}_i) = \frac{1}{N} \sum_{j=1}^N \frac{\left(-\frac{\beta}{2}\left(\nabla V(\mathbf{x}_i) + \frac{\mathbf{x}_i - \mathbf{x}_j}{T}\right)\right)\exp\left[-\frac{\beta}{2}\left(V(\mathbf{x}_i) + \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2T}\right)\right]}{\mathcal{Z}(\mathbf{x}_j)},$$

$$\mathcal{Z}(\mathbf{x}_j) := \int_{\mathbb{R}^d} e^{-\frac{\beta}{2}(V(z) + \frac{\|z - \mathbf{x}_j\|^2}{2T})} \, \mathrm{d}z. \quad \text{(19)}$$

Using these expressions, the score function of the RWPO of $\rho = \rho_k$ can be computed with the simple identity

$$\nabla \log \mathrm{WProx}_{T,V} \rho(\mathbf{x}_i) = \frac{\nabla \mathrm{WProx}_{T,V} \rho}{\mathrm{WProx}_{T,V} \rho}(\mathbf{x}_i).$$

3.1. Computational Considerations.

A parallelization similar to [38] may be employed by concatenating the position and momentum variables into a matrix, and utilizing the structure of $\mathrm{WProx}_{T,V} \rho$ as a sum of exponentials. Recall that the softmax function of a vector $v \in \mathbb{R}^N$ is defined as

$$\mathrm{softmax}(v) = \left(\frac{\exp(v_i)}{\sum_{j=1}^N \exp(v_j)}\right)_{i=1,\dots,N},$$

satisfying $\sum_j \mathrm{softmax}(v)_j = 1$. The score approximation $\nabla \log \mathrm{WProx}_{T,V} \rho_k$ may then be written in terms of a softmax matrix:

$$\nabla \log \mathrm{WProx}_{T,V} \rho_k(\mathbf{x}_i) = -\frac{\beta \nabla V(\mathbf{x}_i)}{2} - \frac{\beta}{2T} \mathbf{x}_i + \frac{\beta}{2T} \sum_{j=1}^N \mathrm{softmax}(W_{i,\cdot})_j \mathbf{x}_j, \quad \text{(20)}$$

where $W_{i,j}$ is an interaction matrix defined as

$$W_{i,j} := -\beta \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{4T} - \log \mathcal{Z}(\mathbf{x}_j). \quad \text{(21)}$$

Notice that the reformulation (20) additionally contains a $\nabla V(\mathbf{x}_i)$ term within the score, which can be combined with the ∇V term within the momentum update (17b). In particular, the discrete-time momentum update (18) can be rewritten as

$$
\mathbf{p}_i^{(k+1)} = (1 - a_k\eta)\mathbf{p}_i^{(k)} + \eta\left(-\nabla V(\mathbf{x}_i^{(k)}) - \beta^{-1}\nabla\log\mathrm{WProx}_{T,V}\,\rho_k(\mathbf{x}_i^{(k)})\right)
$$

$$
= (1 - a_k\eta)\mathbf{p}_i^{(k)} - \frac{\eta}{2}\nabla V(\mathbf{x}_i^{(k)}) + \frac{\eta}{2T}\left(\sum_{j=1}^{N}\mathrm{softmax}(W_{i,\cdot}^{(k)})_j(\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)})\right). \qquad (22)
$$

To compute the normalization constant (19), we can use a Monte Carlo integral applied to

$$
\mathcal{Z}(\mathbf{x}_j) = (4\pi T\beta^{-1})^{d/2}\mathbb{E}_{z\sim\mathcal{N}(\mathbf{x}_j, 2T\beta^{-1})}\left[\exp\left(-\frac{\beta V(z)}{2}\right)\right]. \qquad (23)
$$

We note that the constant $(4\pi T\beta^{-1})^{d/2}$ is canceled out by the logarithm and softmax operators, and can be ignored during computation.

Remark 1. *An alternative to Monte Carlo integration in high dimensions for small T is to use the Laplace approximation [38, 39]. This reads*

$$
\mathcal{Z}(\mathbf{x}_j) = \int_{\mathbb{R}^d} e^{-\frac{\beta}{2}\left(V(z) + \frac{\|z - \mathbf{x}_j\|^2}{2T}\right)}\,\mathrm{d}z
$$

$$
\approx e^{-\frac{\beta}{2}V(\mathbf{x}_j)}C(\beta, T),
$$

where $C(\beta, T)$ is a constant independent of $\mathbf{x}_i$ which also cancels out under the logarithm and the softmax operations.

To finish the parallelization, we combine the position and momentum vectors into matrices

$$
\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_N\end{bmatrix} \in \mathbb{R}^{d\times N}, \quad \mathbf{P} = \begin{bmatrix}\mathbf{p}_1 & \dots & \mathbf{p}_N\end{bmatrix} \in \mathbb{R}^{d\times N}.
$$

The ARWP update (18) can be combined with (22) to be written in matrix form:

$$
\mathbf{P}^{(k+1)} = (1 - a_k\eta)\mathbf{P}^{(k)} - \frac{\eta}{2}\nabla V(\mathbf{X}^{(k)}) + \frac{\eta}{2T}\left(\mathbf{X}^{(k)} - \mathbf{X}^{(k)}\,\mathrm{softmax}(W^{(k)})^\top\right),
$$

$$
\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} + \eta\mathbf{P}^{(k+1)},
$$

where $W^{(k)}$ is the interaction matrix (21) at iteration k. This is summarized in Algorithm 1.

Algorithm 1: ARWP: Accelerated Regularized Wasserstein Proximal Method

Data: Initial points $\mathbf{x}_1^{(1)}, ..., \mathbf{x}_N^{(1)} \in \mathbb{R}^d$, potential $V : \mathbb{R}^d \to \mathbb{R}$, regularization parameter $T > 0$, diffusion $\beta > 0$, step-size $\eta > 0$, iteration count K, damping parameters $a_k > 0$.

Result: $\mathbf{X}^{(K)} = \begin{bmatrix}\mathbf{x}_1^{(K)} & \dots & \mathbf{x}_N^{(K)}\end{bmatrix}$ sampling from $\exp(-\beta V)$.

1 Initialize $\mathbf{X}^{(1)} = \begin{bmatrix}\mathbf{x}_1^{(1)} & \dots & \mathbf{x}_N^{(1)}\end{bmatrix} \in \mathbb{R}^{d\times N}$, initialize $\mathbf{P}^{(1)} = \mathbf{0}_{d\times N}$;

2 **for** $k = 1, ..., K$ **do**

3 $\quad$ Approximate normalizing constants $\mathcal{Z}(\mathbf{x}_i^{(k)})$, $i = 1, ..., N$ using Monte Carlo/Laplace method;

4 $\quad$ Compute interaction matrix $W_{i,j} = -\beta\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{4T} - \log\mathcal{Z}(\mathbf{x}_j)$;

5 $\quad$ Compute row-wise softmax interaction matrix $\mathrm{softmax}(W)_{i,j} = \mathrm{softmax}(W_{i,\cdot})_j$;

6 $\quad$ Evolve momentum matrix

$\qquad \mathbf{P}^{(k+1)} = (1 - a_k\eta)\mathbf{P}^{(k)} - \frac{\eta}{2}\nabla V(\mathbf{X}^{(k)}) + \frac{\eta}{2T}\left(\mathbf{X}^{(k)} - \mathbf{X}^{(k)}\,\mathrm{softmax}(W^{(k)})^\top\right)$;

7 $\quad$ Evolve particle positions $\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} + \eta\mathbf{P}^{(k+1)}$;

8 **end**

Remark 2. *This approximation is used in contrast to the kernel density estimation of [41] or "diffusion map approximation" in [36]. From a computational perspective, one needs to approximate the log-score from an empirical distribution. The regularized Wasserstein proximal allows for this, interpretable as a potential-aware modified kernel method.*

The proposed ARWP method differs from the unregularized accelerated information gradient flow in its choice of score approximation. In particular, [36, 41] both consider using some variant of Gaussian kernel density estimation, for which the interplay between the bias and convergence is not characterized. In the following section, we use properties of the RWPO operator to characterize the asymptotic and non-asymptotic convergence behavior for quadratic potentials.

ARWP does not incur a significant computational increase over the non-accelerated BRWP (16). The main computational cost requirement comes from the interaction term in Algorithm 1, which requires constructing (rows of) an $N \times N$ matrix, which is also required in BRWP. Since updating each particle requires only one call of ∇V, having to track an additional momentum parameter per particle incurs only a constant memory factor increase, which can range from double to negligible depending on the level of parallelization employed.

4. Convergence for Quadratic Potentials

This section analyzes the convergence of the ARWP method, in the case of Gaussian distributions. [37] utilizes a closed-form update for the RWPO for Gaussian distributions, demonstrating that the update in the BRWP method with quadratic potential stays in Gaussian distributions. A similar argument shows that the ARWP method updates Gaussian distributions to Gaussian distributions. In other words, if the target distribution is Gaussian and the initial distribution is Gaussian, then the discrete-time particle updates (17)'s density function ρ_k also follows a Gaussian distribution.

Fix a covariance matrix $\Lambda \in \mathrm{Sym}_{++}(\mathbb{R}^d)$, and define the quadratic potential function $V(x) = x^\top \Lambda^{-1} x / 2$. We additionally assume that all matrices commute so that we may work in a common eigenbasis, and fix $\beta = 1$ without loss of generality.

We show the closure within Gaussian distributions by considering the particle-wise update. Suppose that the distribution at iteration k is Gaussian. We have the following lemmas characterizing the effect of the regularized Wasserstein proximal on a covariance matrix.

Lemma 1. *[37, 38] For a covariance matrix Σ, if $T < \lambda_{\min}(\Lambda)$, then the regularized Wasserstein proximal of the Gaussian distribution $\mathcal{N}(0, \Sigma)$ is also a Gaussian distribution $\mathcal{N}(0, \tilde{\Sigma})$, whose covariance takes the form:*

$$\mathrm{WProx}_{T,V} \mathcal{N}(0, \Sigma) = \mathcal{N}(0, \tilde{\Sigma}),$$

$$\tilde{\Sigma} := 2\beta^{-1} T \left(I + T\Lambda^{-1} \right)^{-1} + \left(I + T\Lambda^{-1} \right)^{-1} \Sigma \left(I + T\Lambda^{-1} \right)^{-1}.$$

Moreover, the inverse operator of the regularized Wasserstein proximal satisfies

$$\mathrm{WProx}_{T,V}^{-1}(\mathcal{N}(0, \Sigma)) = \mathcal{N}(0, (1 + T\Lambda^{-1})\Sigma(1 - T\Lambda^{-1})). \tag{24}$$

As a shorthand, we will use the tilde notation to denote the regularized Wasserstein proximal of a covariance matrix throughout. It is shown in [37] that the terminal distribution under the BRWP update is $\mathcal{N}(0, \Sigma_\infty)$, where Σ_∞ is such that $\mathrm{WProx}_{T,V}(\mathcal{N}(0, \Sigma_\infty)) = \mathcal{N}(0, \Lambda)$, i.e. $\tilde{\Sigma}_\infty = \Lambda$. Moreover, the bias of the stationary distribution under BRWP (and ARWP) depends only on T, in particular, is independent of the step-size η.

We will also find it convenient to define the following two matrix expressions $K_\pm$:

$$K_+ := I + T\Lambda^{-1}, \quad K_- := I - T\Lambda^{-1}. \tag{25}$$

In this section, we analyze the accelerated backward-regularized Wasserstein proximal method for the special case of Gaussian distributions under different approximations of (17).

- Section 4.1 converts the ARWP update (17) into a pair of coupled ODEs in covariance and a dual term.
- Section 4.2 treats the linearized continuous time case, showing the corresponding coupled ODEs observe an asymptotic $\mathcal{O}\left(t \exp\left(-\sqrt{2\lambda^{-1} \frac{1 - T\lambda^{-1}}{1 + T\lambda^{-1}}} t \right) \right)$ error convergence in each eigendirection, where λ is the corresponding eigenvalue of the target covariance matrix Λ.

- While the convergence rate of the corresponding ODE is slightly slower than the unregularized case $T = 0$, Section 4.3 shows that the step-size can be taken to be larger, resulting in a discrete-time iteration speedup by a constant factor of $\frac{1+\sqrt{2}}{2}$. This gives the asymptotic mixing rate of the proposed ARWP method.
- We compare the discrete-time rates with the kinetic Langevin algorithm in Section 4.4, demonstrating a faster rate arising from the regularization.
- In the non-asymptotic case, we show in Section 4.5 that the continuous-time coupled ODEs converge linearly to the target distribution, and show that the damping condition $a > \lambda^{-1/2}$ is sufficient for convergence. This is done using a particular Lyapunov analysis, splitting the ODEs into underdamped and overdamped cases. This allows for the analysis in Section 4.2 to apply over a large time.

4.1. **Continuous Time Covariance Update of ARWP.** Since the regularized Wasserstein proximal of a Gaussian distribution is a Gaussian distribution, the $\mathrm{d}P/\mathrm{d}t$ in (17) is linear in X. Therefore, we may use the ansatz $P_t = G_t X_t$, where $G : \mathbb{R}_{\geq 0} \times \mathbb{R}^d \to \mathbb{R}^d$ is a time varying linear map. Let Σ_t denote the covariance of X_t. After a change of variables, the update (17b) satisifies

$$\frac{\mathrm{d}P}{\mathrm{d}t} = -aP - \Lambda^{-1}X + \tilde{\Sigma}_t^{-1}X = \frac{\mathrm{d}G}{\mathrm{d}t}X + G\frac{\mathrm{d}X}{\mathrm{d}t} = \frac{\mathrm{d}G}{\mathrm{d}t}X + G^2 X.$$

Moreover, the update (17a) turns $\mathrm{d}X/\mathrm{d}t = P = GX$ into $\dot{\Sigma}_t = G_t \Sigma_t + \Sigma_t G_t$. Rearranging yields the following coupled ODE system

$$\begin{cases} \dot{\Sigma}_t = G_t \Sigma_t + \Sigma_t G_t, \\ \dot{G}_t = -aG_t - G_t^2 - \Lambda^{-1} + \tilde{\Sigma}_t^{-1}, \end{cases} \tag{26}$$

where $\tilde{\Sigma}_t$ is the covariance of the regularized Wasserstein proximal applied to $\mathcal{N}(0, \Sigma_t)$. Observe that $(\Sigma, G) = (\tilde{\Sigma}_\infty, \mathbf{0}_{d \times d})$ is a stationary point of (26), where $\Sigma_\infty = (1 + T\Lambda^{-1})\Lambda(1 - T\Lambda^{-1})$ is such that $\tilde{\Sigma}_\infty = \Lambda$. The following sections find convergence rates to this point.

4.2. **Continuous Time Asymptotic Convergence Rate.** To check the convergence behavior of (26) near zero, we can linearize near the terminal state. As we assume that all covariances commute, let us work in one dimension, where our expressions are written in lower case. Then, the continuous time update in 1D is given by

$$\begin{cases} \dot{\sigma}_t = 2g_t \sigma_t, \\ \dot{g}_t = -ag_t - g_t^2 - \lambda^{-1} + \tilde{\sigma}_t^{-1}. \end{cases} \tag{27}$$

Linearizing about the stationary point $(\sigma_\infty, 0)$, where $\tilde{\sigma}_\infty = \lambda$, let us consider the ansatz $\sigma_t = \sigma_\infty + \varepsilon_t$. The first order approximation to the g_t update becomes

$$\lambda^{-1} - \tilde{\sigma}_t^{-1} = \lambda^{-2}(1 + T\lambda^{-1})^{-2}\varepsilon_t + \mathcal{O}(\varepsilon_t^2).$$

The linearized system (in phase space) near the stationary point becomes

$$\begin{cases} \dot{\varepsilon}_t = 2g_t(1 + T\lambda^{-1})(1 - T\lambda^{-1})\lambda, \\ \dot{g}_t = -ag_t - \lambda^{-2}(1 + T\lambda^{-1})^{-2}\varepsilon. \end{cases} \tag{28}$$

The corresponding second order equation is

$$\ddot{\varepsilon} = -a\dot{\varepsilon} - 2\lambda^{-1}k_+^{-1}k_-\varepsilon, \tag{29}$$

where $k_\pm$ are defined as the one-dimensional counterparts of (25). Using the ansatz $\varepsilon_t = e^{-rt}$, the rate satisfies

$$r_\pm = \frac{a \pm \sqrt{a^2 - 8\lambda^{-1}k_+^{-1}k_-}}{2}. \tag{30}$$

The convergence rate is then given by the smaller root in case both roots are real, or the real part in case both roots are complex. In one dimension, the convergence rate is fastest when $a = \sqrt{8\lambda^{-1}k_+^{-1}k_-}$.

In this case, the rate is given by

$$\|(\sigma_t, g_t) - (\sigma_\infty, 0)\|_2 = \mathcal{O}\left(t \exp\left(-\sqrt{2\lambda^{-1}k_+^{-1}k_-}\,t\right)\right).$$

The continuous time rate with regularization is thus slightly slower than the unregularized case $T = 0$ by a factor of $\sqrt{k_+^{-1}k_-} = \sqrt{(1 - T\lambda^{-1})/(1 + T\lambda^{-1})} < 1$. This analysis can be extended into multiple dimensions to find the asymptotic convergence rate of covariance in the trace norm.

Proposition 1. *Let $V(x) = x^\top \Lambda^{-1} x/2$, where the smallest and largest eigenvalues of Λ are $\lambda_{\min}, \lambda_{\max}$ respectively. Let $a > 0$ be some damping parameter, $T \in [0, \lambda_{\min})$ be the regularization parameter, and suppose the initial distribution is $\mathcal{N}(0, \Sigma_0)$. If Σ_t is the continuous-time evolution of ARWP and Σ_t converges to its stationary distribution's covariance $\Sigma_\infty = \mathrm{WProx}_{T,V}^{-1}(\Lambda)$, then the asymptotic convergence rate is*

$$\mathrm{Tr}(\Sigma_t - \Sigma_\infty) = \mathcal{O}(\exp(-rt)), \tag{31}$$

where

$$r = \frac{1}{2}\left[a - \sqrt{\max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]}\left(a^2 - 8\lambda^{-1}\frac{1 - T\lambda^{-1}}{1 + T\lambda^{-1}}, 0\right)}\right]. \tag{32}$$

Proof. The overall convergence rate is given by the smallest rate over each component. The rate corresponding to an eigenvalue $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ is given from (30) as

$$\begin{cases} \frac{a}{2}, & \text{if } a^2 \leq 8\lambda^{-1}\frac{1 - T\lambda^{-1}}{1 + T\lambda^{-1}}; \\ \frac{a - \sqrt{a^2 - 8\lambda^{-1}k_+^{-1}k_-}}{2}, & \text{if } a^2 \geq 8\lambda^{-1}\frac{1 - T\lambda^{-1}}{1 + T\lambda^{-1}}. \end{cases} \tag{33}$$

This is equivalent to (32). $\square$

Remark 3. *Since the regularized Wasserstein proximal performs an affine transformation on the covariance matrix of a Gaussian distribution, the asymptotic convergence rate of the covariance Σ and the RWPO covariances $\tilde{\Sigma}$ are identical.*

While this proposition appears to indicate that a smaller value of a leads to better convergence rates, a necessary condition is that the evolution converges to the stationary point. In Section 4.5, we show that this holds if $a > \lambda^{-1/2}$, similar to existing convergence results for the kinetic Langevin diffusion. This corresponds to requiring a sufficient amount of damping in order for the iterations to converge.

In the next section, we consider the discrete-time analog of the linearized system (28). We show the main advantage of regularizing: we can take a larger step-size if T is nonzero, which increases the discrete-time asymptotic mixing rate. This arises since the condition number of the regularized system is lower than that of the unregularized system.

4.3. **Linearized Discrete-Time Convergence Rate.** In this subsection, we consider the convergence of the one-dimensional RWPO covariances $\tilde{\sigma}_t$ to their stationary distributions $\mathcal{N}(0, \lambda)$, similarly to [37, 38]. Under the change of variables

$$\tilde{\sigma}_t = 2Tk_+^{-1} + k_+^{-2}\sigma_t,$$

as well as the ansatz $\tilde{\sigma}_t = \lambda + \delta_t$, the linearization of (27) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} \delta_t \\ g_t \end{pmatrix} = \begin{pmatrix} [2\lambda - 4Tk_+^{-1}]g_t \\ -ag_t - \lambda^{-2}\delta_t \end{pmatrix} = \underbrace{\begin{bmatrix} 0 & 2\lambda - 4Tk_+^{-1} \\ -\lambda^{-2} & -a \end{bmatrix}}_{=:A}\begin{pmatrix} \delta_t \\ g_t \end{pmatrix}. \tag{34}$$

Denoting the matrix as $A = A(\lambda, T, a)$, the eigenvalues $\chi_\pm$ of A are given by

$$\chi_\pm = \frac{1}{2}[\mathrm{Tr} \pm \sqrt{\mathrm{Tr} - 4\det}]$$

$$= \frac{1}{2}\left[-a \pm \sqrt{a^2 - 4\lambda^{-2}(2\lambda - 4TK_+^{-1})}\right].$$

In the continuous-time case, the stability condition near $(\delta, g) = (0, 0)$ is that all eigenvalues have real component less than 0. This is the case for all $a > 0$.

In the discrete case with step-size $\eta > 0$, the (symplectic Euler) update becomes

$$\begin{pmatrix} \delta_{n+1} \\ g_{n+1} \end{pmatrix} = [I + \eta A] \begin{pmatrix} \delta_n \\ g_n \end{pmatrix}. \tag{35}$$

The stability condition for this update is that $I + \eta A$ has to have (both) eigenvalues lying in the open disk $\{|z| < 1 \mid z \in \mathbb{C}\}$ [19]. Moreover, the convergence rate in this direction is given by $\mathcal{O}(\max\{|1 + \eta\chi_+|, |1 + \eta\chi_-|\}^n)$. In particular, we have the following two special cases, corresponding to two different critical damping parameters. The step-size is controlled by the Lipschitz constant of V, which is $\lambda_{\min}^{-1}$. We demonstrate that the maximal step-size can be taken to be larger than the unregularized version, which corresponds to a faster discrete-time mixing rate, i.e., the rate at which $(\tilde{\Sigma}_k, G_k) \to (\Lambda, 0)$.

Proposition 2. *Suppose the eigenvalues of the covariance matrix Λ are $0 < \lambda_{\min} \leq ... \leq \lambda_{\max}$. Let the update be given by the discrete time update (35) with step-size $\eta > 0$ and regularization $T \in [0, (1 + \sqrt{2})^{-1}\lambda_{\min}]$.*

(a) If the momentum and step-size are chosen to be

$$a = 2\sqrt{2}\lambda_{\max}^{-1/2}\sqrt{\frac{\lambda_{\max} - T}{\lambda_{\max} + T}}, \quad \eta = \frac{\lambda_{\max}^{-1/2}\sqrt{\frac{\lambda_{\max} - T}{\lambda_{\max} + T}}}{\sqrt{2}\lambda_{\min}^{-1}\frac{\lambda_{\min} - T}{\lambda_{\min} + T}}, \tag{36}$$

then the mixing rate of the linearized system is

$$\sqrt{1 - \kappa^{-1}\frac{\lambda_{\max} - T}{\lambda_{\max} + T}\frac{\lambda_{\min} + T}{\lambda_{\min} - T}}. \tag{37}$$

In particular, taking $T = (1 + \sqrt{2})^{-1}\lambda_{\min}$, the discrete time rate for $\kappa \gg 1$ and the linearized system is (up to first order)

$$1 - \frac{1 + \sqrt{2}}{2}\kappa^{-1}.$$

(b) If the momentum and step-size are chosen to be

$$a = 2\sqrt{2}\lambda_{\min}^{-1/2}\sqrt{\frac{\lambda_{\min} - T}{\lambda_{\min} + T}}, \quad \eta = 2a^{-1} = \frac{1}{\sqrt{2}}\lambda_{\min}^{1/2}\sqrt{\frac{\lambda_{\min} + T}{\lambda_{\min} - T}}, \tag{38}$$

then the mixing rate is

$$\sqrt{1 - \kappa^{-1}\frac{\lambda_{\max} - T}{\lambda_{\max} + T}\frac{\lambda_{\min} + T}{\lambda_{\min} - T}}.$$

In particular, taking $T = (1 + \sqrt{2})^{-1}\lambda_{\min}$, the discrete-time mixing rate for $\kappa \gg 1$ is (up to first order)

$$1 - \frac{1 + \sqrt{2}}{2}\kappa^{-1}. \tag{39}$$

Sketch. The momentum parameters are chosen to be optimal for the largest and smallest eigenvalues of Λ, respectively, and the step-sizes are chosen to be maximal such that the method converges. Moreover, the function $x \mapsto x^{-1}\frac{x-T}{x+T}$ is maximized at $(1 + \sqrt{2})T$ and is decreasing for $x > (1 + \sqrt{2})T$. The rates are obtained from a worst-case analysis over all possible eigenvalues for the given momentum and step sizes. A full derivation is given in Section B.1. $\square$

We note that the restriction on regularization $T \in [0, (1 + \sqrt{2})^{-1}\lambda_{\min}]$ is used only to provide a uniform worst-case bound, using the monotonicity of $\lambda \mapsto \lambda^{-1}\frac{\lambda - T}{\lambda + T}$, which is increasing over $[T, (1 + \sqrt{2})T]$ and decreasing over $[(1 + \sqrt{2})T, +\infty)$. This can be relaxed to $T \in [0, \lambda_{\min})$, by replacing all instances of $\lambda_{\min}^{-1}\frac{\lambda_{\min} - T}{\lambda_{\min} + T}$ and $\lambda_{\max}^{-1}\frac{\lambda_{\max} - T}{\lambda_{\max} + T}$ with $\max\{\lambda^{-1}\frac{\lambda - T}{\lambda + T} \mid \lambda \in \mathrm{Spec}(\Lambda) \subset [\lambda_{\min}, \lambda_{\max}]\}$ and $\min\{\lambda^{-1}\frac{\lambda - T}{\lambda + T} \mid \lambda \in \mathrm{Spec}(\Lambda) \subset [\lambda_{\min}, \lambda_{\max}]\}$ respectively. The constant acceleration factor arises as the condition number of eigenvalues decreases after applying the function $\lambda \mapsto \lambda^{-1}\frac{\lambda - T}{\lambda + T}$ if $T > 0$.

4.4. **Comparison with Kinetic Langevin Diffusion.** The kinetic/underdamped Langevin diffusion is given as a second-order version of the stochastic dynamics (2). In $\mathbb{R}^d$, if a particle position is X with momentum P, the kinetic Langevin update proceeds by adding a Brownian motion on the momentum parameter [12],

$$\begin{bmatrix} \mathrm{d}X \\ \mathrm{d}P \end{bmatrix} = \begin{bmatrix} P \\ -(aP + u\nabla V(X)) \end{bmatrix} \mathrm{d}t + \sqrt{2au} \begin{bmatrix} 0 \\ I \end{bmatrix} \mathrm{d}W, \tag{40}$$

where W is a $2d$-dimensional standard Brownian motion, $a > 0$ is a friction coefficient, and $u > 0$ is an inverse mass, which can be taken to be $u = 1$ without loss of generality. This converges to the phase-space stationary distribution $\rho(x,p) \propto \exp\left(-V(x) - \frac{1}{2u}\|p\|^2\right)$. A more detailed treatment is given in Section A. While more general convergence results are given in [10, 12], [42] specializes into the Gaussian setting, and we can compare the asymptotic convergence rates with the proposed ARWP method.

For now, consider the kinetic Langevin update in one dimension. For a target distribution $\mathcal{N}(0, \Lambda)$, the particle position and momentums $(X_t, P_t) \in \mathbb{R}^2$ follow a joint normal distribution

$$(X_t, P_t) \sim \mathcal{N}\left(0, \begin{pmatrix} \Sigma_{11}(t) & \Sigma_{12}(t) \\ \Sigma_{12}(t) & \Sigma_{22}(t) \end{pmatrix}\right),$$

with terminal values $(\Sigma_{11}, \Sigma_{12}, \Sigma_{22}) \to (\lambda, 0, 1)$ From [42, Cor. 3.3], the covariance update satisfies the following linear system[1]

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} \Sigma_{11} \\ \Sigma_{12} \\ \Sigma_{22} \end{bmatrix} = \begin{bmatrix} 0 & 2 & 0 \\ -\lambda^{-1} & -a & 1 \\ 0 & 2\lambda^{-1} & -2a \end{bmatrix} \begin{bmatrix} \Sigma_{11} - \lambda \\ \Sigma_{12} \\ \Sigma_{22} - 1 \end{bmatrix}. \tag{41}$$

In particular, the eigenvalues of the update matrix are given by

$$-a, \quad -a \pm \sqrt{a^2 - 4\lambda^{-1}}.$$

In multiple dimensions, standard numerical analysis gives that the convergence rate is given by the largest norm of $1 + \eta\chi$, where χ runs over the three eigenvalues of the update matrix in (41), and over the eigenvalues of Λ. It remains to compute the step-size that minimizes the maximum norm over all possible eigenvalues of Λ.

In the small momentum critical damping case, the optimal momentum is taken to be $a = 2\lambda_{\max}^{-1/2}$, which gives a continuous-time convergence rate of $\mathcal{O}(t^2 \exp(-at))$ in covariance. To compute the optimal step-size $\eta > 0$, one computes

$$|1 + \eta(-a \pm \sqrt{a^2 - 4\lambda_{\min}^{-1}})|^2 < 1$$
$$\Leftrightarrow 1 - 4\eta\lambda_{\max}^{-1/2} + 4\eta^2\lambda_{\min}^{-1} < 1.$$

The rate is minimized when the quadratic on the left is minimized, which occurs when $\eta = \lambda_{\min}\lambda_{\max}^{-1/2}/2$. The discrete time per-iteration contraction rate is then given by

$$\max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} |1 - \eta(-a \pm \sqrt{a^2 - 4\lambda^{-1}})|$$
$$= \sqrt{1 - 4\frac{\lambda_{\min}\lambda_{\max}^{-1/2}}{2}\lambda_{\max}^{-1} + 4\frac{\lambda_{\min}^2\lambda_{\max}^{-1}}{4}\lambda_{\min}^{-1}} = \sqrt{1 - \kappa^{-1}}.$$

This should be compared with Proposition 2(a). From (37), we have acceleration as the constant in front of κ^{-1} is $\frac{\lambda_{\max}-T}{\lambda_{\max}+T}\frac{\lambda_{\min}+T}{\lambda_{\min}-T} > 1$. The linearized system of the proposed ARWP method is therefore more well-behaved than the one in kinetic Langevin method.

A similar analysis can be performed in the high critical damping case, where $a = 2\lambda_{\min}^{-1/2}$. As in Section B.1.2, the optimal step-size is given by $\eta = 1/a$. The rate is similarly given by $\sqrt{1 - \kappa^{-1}}$, and we conclude the same conclusion as in the previous case.

---

[1] The result in the reference should be applied with $a = 0$ and $\mathbf{C} = I$ as denoted in their work.

4.5. **Non-Linearized Non-Asymptotic Continuous Time Convergence.** In order to apply the analysis of the previous three sections, we need to show that the system indeed converges to the stationary distribution. In this section, we demonstrate convergence for the non-linearized continuous time system (27), which converges for $a > \lambda^{-1/2}$.

We first show convergence of the non-linearized system in continuous time. In one dimension, the nonlinear coupled ODEs governing the covariance are given by

$$\begin{cases} \dot{\sigma}_t = 2g_t\sigma_t \\ \dot{g}_t = -ag_t - g_t^2 - \lambda^{-1} + \tilde{\sigma}_t^{-1}. \end{cases}$$

Nonlinearities arise from the introduction of the g_t^2 term, as well as the inverse covariance term $\tilde{\sigma}_t^{-1}$ in the g_t update. We may change this into a pair of coupled ODEs in $\dot{\tilde{\sigma}}_t$ using the change of variables

$$\tilde{\sigma}_t = 2Tk_+^{-1} + k_+^{-2}\sigma_t.$$

The change of variables becomes a forcing term in $\tilde{\sigma}_t$,

$$\begin{cases} \dot{\tilde{\sigma}}_t = 2g_t\tilde{\sigma}_t - 4g_tTk_+^{-1} \\ \dot{g}_t = -ag_t - g_t^2 - \lambda^{-1} + \tilde{\sigma}_t^{-1} \end{cases} \tag{42}$$

By selecting a particular Lyapunov function, we may show that this coupled ODE system converges to the stationary point $(\tilde{\sigma}_t, g_t) \to (\lambda, 0)$. This implies that for Gaussian distributions, the accelerated regularized Wasserstein proximal method in continuous time converges to the stationary distribution. We have three different results, corresponding to the underdamped case $a \in (\lambda^{-1/2}, 2\lambda^{-1/2}]$, a "critical" damping case $a = 2\lambda^{-1/2}$, and an overdamped case $a \geq 2\lambda^{-1/2}$. The underdamped and critical damping cases can use the same Lyapunov function, while the overdamped case requires a modified Lyapunov function. The results are summarized in the following two propositions.

Proposition 3. *Consider the quadratic potential in one dimension $V(x) = \lambda^{-1}x^2/2$ and diffusion parameter $\beta = 1$, and further let $T \in [0, \lambda)$ be a regularization parameter. Consider evolving a Gaussian distribution $\mathcal{N}(0, \Sigma_t)$ through the continuous-time ARWP system (42). Define a Lyapunov function as*

$$\mathcal{E}_t := (\tilde{\sigma}_t - 2Tk_+^{-1})[(\lambda^{-1/2} - \tilde{\sigma}_t^{-1/2}) + g_t]^2 + 2\mathrm{D}_{\mathrm{KL}}(\tilde{\sigma}_t, \lambda), \tag{43}$$

where we write the KL divergence between two variances to represent the KL divergence between the corresponding zero-mean Gaussian distributions. Then, the regularized Wasserstein proximal of the distributions $\mathrm{WProx}_{T,V}(\mathcal{N}(0, \sigma_t)) = \mathcal{N}(0, \tilde{\sigma}_t)$ converges to the terminal distribution $\mathcal{N}(0, \lambda)$. Furthermore, the convergence rate can be characterized as follows:

(1) (Critically damped) In one dimension, let the momentum parameter be taken as $a = 2\lambda^{-1/2}$. Furthermore, assume that the covariance satisfies $\tilde{\sigma}^2 \geq 2Tk_+^{-1}\lambda$. Then, the Lyapunov function satisfies the Lyapunov-like decay

$$\dot{\mathcal{E}}_t \leq -\lambda^{-1/2}(1 - 2Tk_+^{-1}\tilde{\sigma}_t^{-1})\mathcal{E}_t. \tag{44}$$

In particular, close to the terminal distribution $\tilde{\sigma}_t \approx \lambda$, the decay is $\mathcal{E}_t = \mathcal{O}(\exp(-rt))$, where the rate is

$$r = \left(\frac{\lambda - T}{\lambda + T}\right)\lambda^{-1/2}. \tag{45}$$

(2) (Underdamped) For $a \in (\lambda^{-1/2}, 2\lambda^{-1/2}]$, define

$$p = p_t := \lambda^{-1/2} + 2Tk_+^{-1}\tilde{\sigma}^{-3/2}, \quad b_+ := \lambda^{-1/2} + \tilde{\sigma}_t^{-1/2}.$$

Let r be the smallest positive root of the following (time-varying) quadratic equation:

$$p^2 - 4\left(-p + rb_+\frac{\tilde{\sigma} - 2Tk_+^{-1}}{\tilde{\sigma}(2\sqrt{\lambda}b_+ - 1)}\right)(-(1 - r)b_+) = 0.$$

Then r exists, and the rate is given by

$$\dot{\mathcal{E}}_t \leq -\frac{2rb_+(\tilde{\sigma} - 2Tk_+^{-1})}{\tilde{\sigma}(2\sqrt{\lambda}b_+ - 1)}\mathcal{E}_t.$$

Sketch. Differentiating the Lyapunov function gives a quadratic equation in g_t, which is upper-bounded over all possible g_t. The conditions arise from the requirement that the g_t^2 coefficient in the quadratic is negative. The full derivation is given in Section C; the first part is given in Section C.0.1 and the second part in Section C.0.2. □

Remark 4. *As seen in part 2 of the proposition, the assumption that the covariance is larger than a constant in part 1 is not strictly necessary. Moreover, the first case is a special case of the second. This proposition quantifies the observation in [37], that the convergence rate is a bit slower if the initial covariance is too small, but accelerates again close to the terminal distribution. This slowdown does not occur if the initial covariance is larger than the terminal covariance.*

This shows that in the underdamped and critically damped cases $a \in (\lambda^{-1/2}, 2\lambda^{-1/2}]$, the ODE system converges to the terminal solution $(\tilde{\sigma}_t, g_t) \to (\lambda, 0)$. We note that for $a \leq \lambda^{-1/2}$, the Lyapunov function may not necessarily decrease, and may lead to oscillation behaviors, similarly to [35]. A similar theoretical restriction arises in [12], which requires that the damping be greater than $m^{-1/2}$, where m is the strong convexity constant of V.

In the overdamped case, the Lyapunov function as defined in (43) does not necessarily decay. To show the convergence, we need to consider a modified Lyapunov function. This is given in the following proposition.

Proposition 4. *(Overdamped) Let $V(x) = \lambda^{-1}x^2/2$ and $T \in [0, \lambda)$ be as in the previous proposition. Suppose that the momentum damping parameter is $a \geq 2\lambda^{-1/2}$, and define $\zeta := a\lambda^{1/2}/2$. Define a modified Lyapunov function as*

$$\mathcal{F}_t = \zeta^{-1}(\tilde{\sigma}_t - 2Tk_+^{-1})[b_- + \zeta g_t]^2 + 2\zeta D_{\mathrm{KL}}(\tilde{\sigma}_t, \lambda). \tag{46}$$

Moreover, define the (time-varying) variables

$$p = p_t := a\zeta - \lambda^{-1/2} + 2Tk_+^{-1}\tilde{\sigma}_t^{-3/2}, \quad b_+ := \lambda^{-1/2} + \tilde{\sigma}_t^{-1/2}. \tag{47}$$

Let r be the smallest positive root of the (time-varying) quadratic equation

$$\zeta^{-2}p^2 + 4\left(-p + rb_+\frac{\tilde{\sigma}_t - 2Tk_+^{-1}}{(2\sqrt{\lambda}b_+ - 1)\tilde{\sigma}_t}\right)(1 - r)b_+ = 0.$$

Then r exists, and the modified Lyapunov function (46) decays as

$$\dot{\mathcal{F}}_t \leq -\frac{2rb_+(\tilde{\sigma} - 2Tk_+^{-1})}{\zeta(2\sqrt{\lambda}b_+ - 1)\tilde{\sigma}}\mathcal{F}_t. \tag{48}$$

Sketch. The form of the Lyapunov function comes from inspecting the previous Lyapunov function (43), and transferring the modifications of [35] to the linear convergence case. The definition of the rate being in terms of a quadratics' root is sufficient in order to guarantee that the Lyapunov function is decreasing. □

We remark that a more general sufficient condition is $\zeta \geq a\lambda^{-1/2}/2$. However, due to the presence of ζ^{-1} in the rate, it is not beneficial to take a larger ζ.

While the analysis presented so far is for the one-dimensional case, in the commuting case, we can extend this to higher dimensions simply by taking the trace over each eigendirection. For a damping parameter a to work for all eigenvalues, one should consider the overdamped case, i.e., extending Proposition 4. This can be summarized in the following corollary, in which the Lyapunov function is defined using a weighted KL divergence.

Corollary 1. *(Overdamped) Let $V(x) = x^\top \Lambda x/2$ and $T \in [0, \lambda_{\min})$, and assume Σ_0 commutes with Λ. Suppose that the momentum damping parameter is $a \geq 2\lambda_{\min}^{-1/2}$, and define $Z := a\Lambda^{1/2}/2$. Define a modified Lyapunov function as*

$$\mathcal{F}_t = \mathrm{Tr}\big(Z^{-1}(\Sigma_t - 2K_+^{-1}[B_- + ZG_t]^2)\big) + 2\sum_{i=1}^{d}\zeta_i D_{\mathrm{KL}}(\tilde{\sigma}_{t,i}, \lambda_i). \tag{49}$$

Moreover, define the (time-varying) variables in each eigendirection

$$p_i = p_{t,i} := a\zeta_i - \lambda_i^{-1/2} + 2Tk_{+,i}^{-1}\tilde{\sigma}_{t,i}^{-3/2}, \quad b_{+,i} := \lambda_i^{-1/2} + \tilde{\sigma}_{t,i}^{-1/2}, \quad i = 1,...,d. \qquad (50)$$

Let r_i, $i = 1,...,d$ be the smallest positive roots of the (time-varying) quadratic equations

$$\zeta^{-2}p_i^2 + 4\left(-p_i + rb_{+,i}\frac{\tilde{\sigma}_{t,i} - 2Tk_{+,i}^{-1}}{(2\sqrt{\lambda}b_{+,i} - 1)\tilde{\sigma}_{t,i}}\right)(1 - r)b_{+,i} = 0, \quad i = 1,...,d.$$

Then r exists, and the modified Lyapunov function (49) decays as

$$\dot{\mathcal{F}}_t \le -\min_{i=1,...,d}\left(\frac{2r_ib_{+,i}(\tilde{\sigma}_{t,i} - 2Tk_{+,i}^{-1})}{\zeta_i(2\sqrt{\lambda}b_{+,i} - 1)\tilde{\sigma}_{t,i}}\right)\mathcal{F}_t. \qquad (51)$$

Proof. By the assumption at the start of the section, the matrices Z, B_-, G_t, Σ_t all commute. Summing (46) over all eigendirections yields the multidimensional Lyapunov function (49). In each direction, the decay is given by (46); taking the smallest decay coefficient yields the uniform decay (49). $\qquad\square$

This section shows that the forced ODE system (42) converges to the desired stationary distribution. However, these decay rates are not sharp. Each result shows that if the covariance of the regularized Wasserstein proximal is $\tilde{\sigma} \approx 2Tk_+^{-1}$, corresponding to $\sigma \approx 0$, then the convergence rate is slow.

5. Experiments

In the following numerical experiments, we compare the performance of the proposed ARWP method with several classical sampling methods and the non-accelerated BRWP method. This is first done on Gaussian target distributions, directly in covariance space, then using particle evolutions. Some low-dimensional non-log-concave examples follow to demonstrate the sensitivity and effectiveness in exploring away from local potential wells, as well as a high-dimensional Bayesian neural network example. The compared parameters for each of the experiments are given in Section G.

5.1. **One-Dimensional Gaussian.** We first verify the analysis for convergence in distributions presented in Section 4. In discrete time, the ARWP update (18) has a closed-form update in covariances, given explicitly in Section D. We verify the results of the linearized discrete-time update in Section 4.3.

We consider the simplest case where the target variance is $\mathcal{N}(0, 1)$, or equivalently $V(x) = \|x\|^2/2$, and continue to fix $\beta = 1$. To demonstrate the optimal choice of damping parameter a and step-size η, we plot a contour plot of the (trace) norm $|\sigma_k - \sigma_\infty|$, where $\sigma_\infty = 1 - T^2$ is such that $\text{WProx}_{T,V}(\mathcal{N}(0, \sigma_\infty)) = \mathcal{N}(0, 1)$. This error is plotted against the damping parameter $a \in [10^0, 10^2]$ and step-size $\eta \in [10^{-4}, 10^{-0.5}]$, equally log-spaced with 200 points.

Figure 1 plots this error in the covariance when updating using the ARWP method, with fixed regularization parameter $T = 0.2$, and with the covariance starting either as $\sigma_0 = 10^{-3}$ or as $\sigma_0 = 4$. From Proposition 1 applied with $\lambda = 1$, we know that the optimal asymptotic rate is given when $a = 2\sqrt{2}\lambda^{-1/2}\sqrt{\frac{1-T\lambda^{-1}}{1+T\lambda^{-1}}}$, which qualitatively manifests as a cusp in the contour and is marked by a gray dashed line. Furthermore, the contour plot exhibits a "bouncing" phenomenon as the step-size increases for a fixed damping parameter a. As the number of iterations is fixed, the x-axis can be approximately interpreted as time, and this "bouncing" is the characteristic of Euclidean accelerated methods.

5.2. **Ill-Conditioned Gaussian.** We now consider a 2D Gaussian with $\Sigma = \text{diag}(0.1, 5)$, i.e. with condition number 50, applied with a finite number of particles $N = 100$. We compare with the standard Langevin algorithms ULA and MALA [15], and the non-accelerated regularized Wasserstein proximal method BRWP [37]. We additionally compare with two accelerated algorithms arising from discretizing the kinetic Langevin dynamics, namely the inertial Langevin algorithm (ILA) [16], and the kinetic Langevin Monte Carlo (KLMC) method [12]. These two algorithms are recalled in Section H.
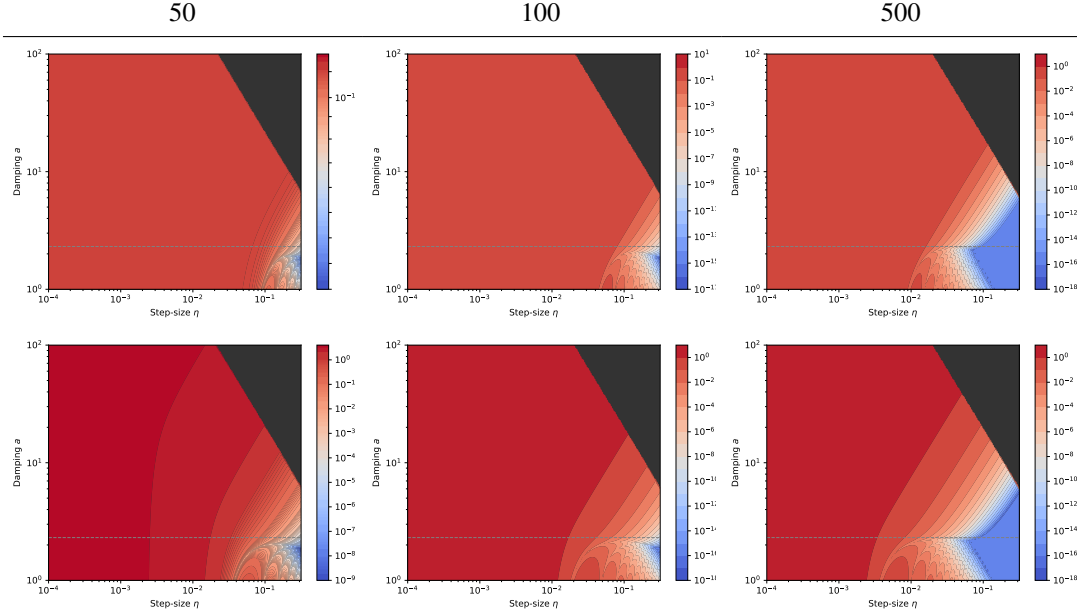
FIGURE 1. Contour plots of the covariance error of discrete-time ARWP (18), for target distribution $\Lambda = \mathcal{N}(0, I_1)$, with initialization $\mathcal{N}(0, 10^{-3})$ (top) and $\mathcal{N}(0, 4)$ (bottom). Error is plotted against the damping parameter $a \in [10^0, 10^2]$ and step-size $\eta \in [10^{-4}, 10^{-0.5}]$, and fixed $T = 0.2$. The gray line indicates the optimal damping parameter in continuous time, given by (38). The black region in the top right corner indicates (empirical) divergence, occurring when $a\eta > 2$.

We also consider another standard choice of acceleration, where the damping parameter is chosen as

$$1 - a_k\eta = \frac{k-1}{k+2}. \tag{52}$$

The case where the damping a is constant is denoted in future figures as "ARWP-Heavy-ball", while the variable case (52) is denoted "ARWP-Nesterov". This is in accordance with the classical optimization algorithms. We note that ARWP-Heavy-ball requires an additional choice of damping parameter a.

Figure 2 demonstrates the convergence in KL divergence of the proposed ARWP methods, as computed using a Gaussian KDE with bandwidth 0.05 and numerically integrated over $[-5, 5]^2$ with mesh size $\Delta x = 0.01$. It is compared with the unaccelerated BRWP method, classical Langevin methods ULA and MALA, as well as the accelerated Langevin methods ILA and KLMC. We observe acceleration of ILA and KLMC compared to ULA. In this simple case, MALA is able to perform similarly to the accelerated methods.

The deterministic methods ARWP and BRWP are both able to reach significantly lower terminal KL divergence, due to the structure of the final iterates. In particular, with this low number of particles, unbiased methods such as MALA still have regions of mass that are not represented by particles, leading to a higher KL divergence. This is demonstrated in Figure 3, where the particle positions under the Langevin algorithms are less structured.

The acceleration in ARWP for this simple case is mild, manifesting as a slightly faster convergence with the same $T = 0.05$. This is due to the slightly larger allowed step-size in ARWP. In particular, ARWP is able to use a step-size of 0.3 instead of 0.2 for BRWP, which diverges for step-size 0.3. From (38), the optimal step-size for ARWP for optimally chosen damping is given by

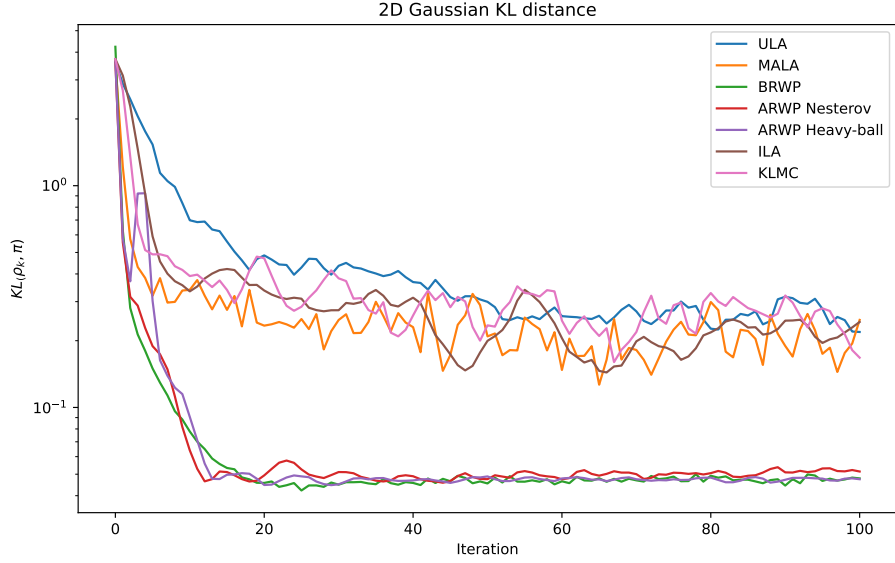$$h^*_{\text{ARWP}} = \frac{1}{\sqrt{2}}(0.1)^{1/2}\sqrt{\frac{0.1 + 0.05}{0.1 - 0.05}} \approx 0.387.$$

FIGURE 2. Convergence in KL divergence for the 2D Gaussian, run with 100 particles over 100 iterations. We observe that the deterministic methods ARWP and BRWP enjoy particle-wise convergence, indicated by the smaller oscillations between iterations. The accelerated Langevin methods ILA and KLMC continue to evolve due to the Brownian motion in the velocity. We observe that while BRWP has a faster initial convergence rate, both ARWP-Nesterov and ARWP-Heavy-ball reach their steady states faster. This is consistent with classical optimization results.
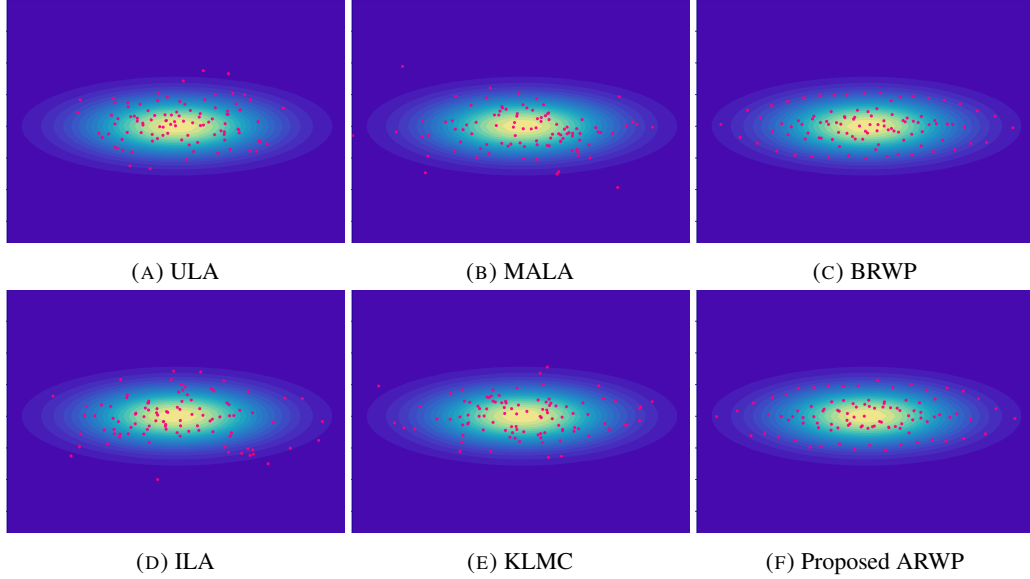

(A) ULA   (B) MALA   (C) BRWP

(D) ILA   (E) KLMC   (F) Proposed ARWP

FIGURE 3. Particle positions after 100 iterations for the 2D Gaussian with condition number $\kappa = 50$, run with 100 particles. We observe that both the accelerated and non-accelerated Langevin algorithms look more randomly sampled, as the particles do not interact. Moreover, the proposed ARWP method has a similarly structured but slightly messier terminal position compared to BRWP. Both ARWP and BRWP particle positions converge and do not move.

However, the theoretical optimal step-size for BRWP is given by [37, Cor. 1]

$$h_{\text{BRWP}}^* = \frac{1}{2}(0.1)\frac{0.1 + 0.05}{0.1 - 0.05} = 0.15,$$

which is smaller than that of ARWP.

Remark 5. *By taking a slightly larger step size, we sacrifice some speed in the fast directions for small covariance, while accelerating the slow directions for large covariance. Taking a slightly larger than optimal step-size usually results in acceleration.*

5.3. **2D Rosenbrock.** We consider the (scaled) Rosenbrock function [30]

$$V(x, y) = \frac{1}{20}\left[(1 - x)^2 + 100(y - x^2)^2\right],$$

with the gradient operator

$$\begin{bmatrix}\partial_x V \\ \partial_y V\end{bmatrix}(x, y) = \frac{1}{20}\begin{bmatrix}2(1 - x) - 400x(y - x^2) \\ 200(y - x^2)\end{bmatrix}.$$

This is a difficult non-log-concave distribution to sample. The high Lipschitz constants away from the valley require a small step size, leading to slow exploration along the valley. In particular, the mass of the distribution is distributed away from the valley near the minimizer[2]; only about 30% of the mass is distributed within the square region indicated within the figure [30, Fig. 1]. Therefore, it is desirable for a method to be able to sample from the tails of this distribution. We run the methods with 100 particles initialized with distribution $\mathcal{N}(0, I_2)$ up to 500 iterations.

Figure 4 plots the evolution of ARWP-Nesterov, ILA with a low damping parameter, KLMC, and ULA. We observe that ARWP-Nesterov is able to properly diffuse along the parabolic potential well, while keeping an appropriate number of particles near the origin. Section F contains some additional comparisons with other baselines ILA and BRWP, as well as a hyperparameter ablation for ARWP-Heavy-ball.

5.4. **Multi-Modal Gaussian Mixture.** We now consider a four-mode weighted Gaussian mixture in two dimensions. In this case, the potential is given by

$$V(x) = -\log\left[-\sum_{i=1}^4 w_i \exp\left(-2\pi\sigma_i^2\left\|\frac{x - c_i}{\sigma_i^2}\right\|^2\right)\right],$$

where the centers are given by $\{c_i\} = \{(0, 0), (3, 0), (-3, -1), (-3, 1)\}$, weights $\{w_i\} = \{1, 0.5, 0.5, 0.5\}$, and bandwidths $\{\sigma_i^2\} = \{0.5, 0.25, 0.25, 0.25\}$. This potential is a large well at the origin, with a smaller well on one side, and two smaller wells on the other side. This is run with 100 particles, with initial distribution $\mathcal{N}((3, 0), I_2)$ in a suboptimal well. The methods are run for 400 iterations to allow for sufficient mixing.

Figure 5 gives the evolution of the KL divergence of the various methods, run with optimal parameters as found using a grid search to minimize divergence at iteration 400. The KL divergence is approximated using a Gaussian KDE with bandwidth 0.1, and integrated over the grid $[-5, 5]^2$ with mesh size $\Delta x = 0.01$. We observe that the KL divergence of the ARWP methods are able to decrease faster than the compared Langevin methods as well as BRWP. Moreover, the combination of the acceleration and modified kernel allows for the particles to diffuse into all the potential wells in a structured manner, leading to a lower terminal divergence.

Figure 6 plots the particles at iterations 10, 50, and 200 for ARWP-Heavy-ball, ILA, KLMC and MALA. Each of the methods are able to diffuse to the opposite wells by iteration 200. We observe again a structured phenomenon for ARWP across all of the different wells.

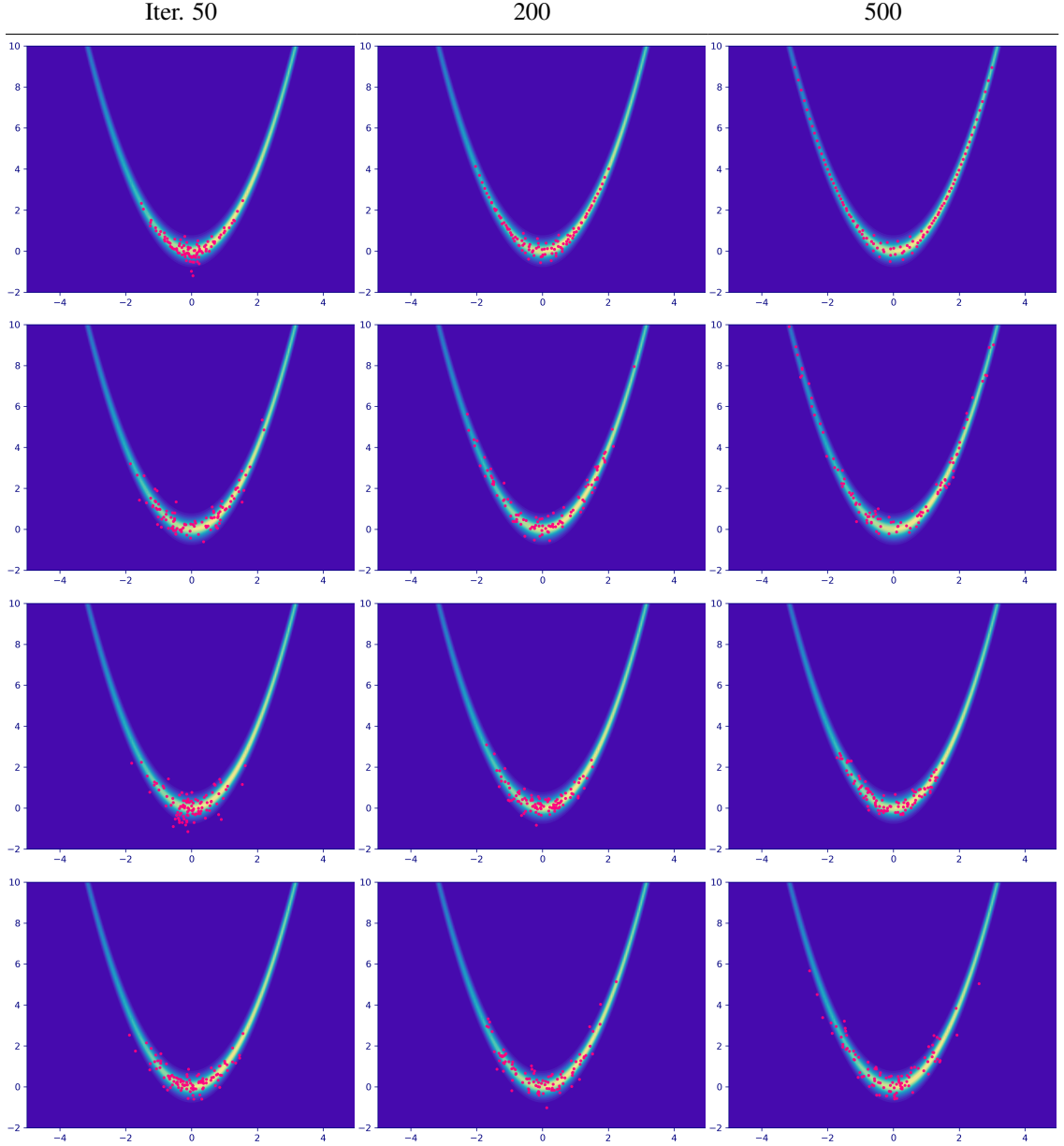[2]This also means that approximating the KL divergence using numerical integration is expensive.

FIGURE 4. Particle evolution for Rosenbrock distribution at iterations 50, 200, 500. Top to bottom: (1) ARWP-Nesterov ($T = \eta = 0.02$), (2) "underdamped ILA" ($\eta = 0.05$, damping parameter $= 2$), (3) KLMC ($\eta = 0.01$, damping parameter $= 5$), and (4) ULA ($\eta = 0.01$). We observe that ARWP and ILA are better at exploring the tails than KLMC. However, ILA and ULA both have particles straying away from the main parabola due to time-discretization bias.

5.5. **Bayesian Neural Networks.** For a high-dimensional non-log-concave target distribution, we consider the Bayesian neural network experiment as done in [41, 37, 38]. This consists of training some neural networks over five UCI[3] datasets. We compare against existing baselines given by various gradient flows.

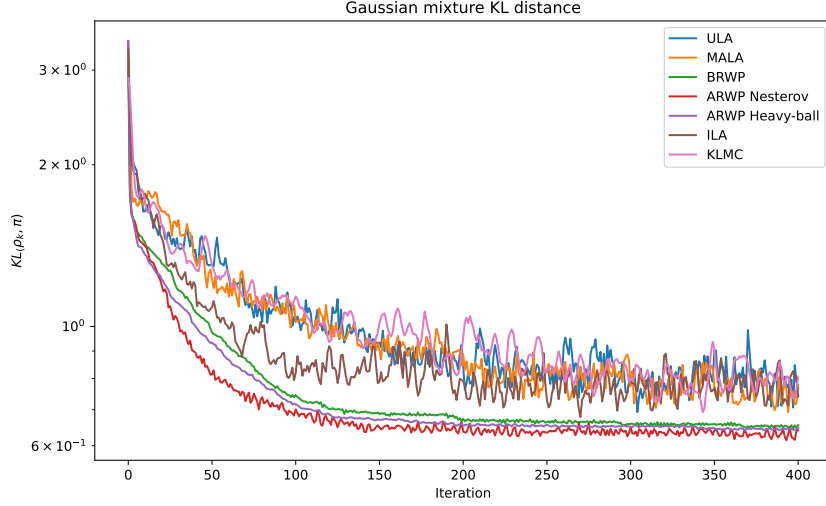[3]https://archive.ics.uci.edu/datasets

FIGURE 5. KL divergence between the particles and the underlying distribution for the Gaussian mixture. We observe that ARWP method converges faster than BRWP, and the particle cloud stabilizes with lower KL divergence than the corresponding Langevin methods. KLMC experiences mode collapse, which persists through hyperparameter changes.

Each particle is represented by a two-hidden-layer ReLU neural network, each with 50 neurons and default Gaussian initialization. The epoch and batch-size hyperparameters are taken as in [41], and the methods are run with $N = 10$ "particles". The step-size and T hyperparameters are chosen by a grid search ranging from $\eta \in [2 \times 10^{-2}, 3 \times 10^{-1}]$ and $T \in [10^{-3}, 10^{-2}]$. Values reported are averaged over 20 independent runs.

Table 1 reports the test root-mean-square error for training with BRWP, accelerated information gradient (AIG), Wasserstein gradient flow [41], and Stein variational gradient descent [40]. We observe that ARWP(-Nesterov) is able to consistently outperform BRWP on this task. Compared with Adam, the particle-based methods are also able to find networks with better generalization.

TABLE 1. Test root-mean-square-error (RMSE) on test datasets on various Bayesian neural network tasks, averaged over 20 runs. Bold indicates smallest in row, underlined denotes second smallest. ARWP(-Nesterov) consistently performs better than BRWP on this task, with a higher variance. This indicates that the particles are able to find better test-generalization, at the cost of also finding some poorer particles.

Dataset	Adam	ARWP	BRWP	AIG	WGF	SVGD
Boston	$3.350_{\pm 8.33e-1}$	$2.902_{\pm 7.25e-1}$	$3.309_{\pm 5.31e-1}$	$2.871_{\pm 3.41e-3}$	$3.077_{\pm 5.52e-3}$	$\mathbf{2.775_{\pm 3.78e-3}}$
Combined	$3.971_{\pm 1.79e-1}$	$\mathbf{3.939_{\pm 1.89e-1}}$	$3.975_{\pm 3.94e-2}$	$4.067_{\pm 9.27e-1}$	$4.077_{\pm 3.85e-4}$	$4.070_{\pm 2.02e-4}$
Concrete	$4.698_{\pm 4.85e-1}$	$\mathbf{4.257_{\pm 8.46e-1}}$	$4.478_{\pm 2.05e-1}$	$4.440_{\pm 1.34e-1}$	$4.883_{\pm 1.93e-1}$	$4.888_{\pm 1.39e-1}$
Kin8nm	$0.089_{\pm 2.72e-3}$	$\underline{0.089}_{\pm 2.47e-3}$	$\underline{0.089}_{\pm 6.06e-6}$	$0.094_{\pm 5.56e-6}$	$0.096_{\pm 3.36e-5}$	$0.095_{\pm 1.32e-5}$
Wine	$0.629_{\pm 4.01e-2}$	$0.608_{\pm 3.43e-2}$	$0.623_{\pm 1.35e-3}$	$0.606_{\pm 1.40e-5}$	$0.614_{\pm 3.48e-4}$	$\mathbf{0.604_{\pm 9.89e-5}}$

6. DISCUSSION

This work introduces the accelerated regularized Wasserstein proximal (ARWP) method for sampling from a target distribution. There are several accelerated schemes in probability density space. One is from overdamped Langevin to kinetic Langevin dynamics. The other is to add a
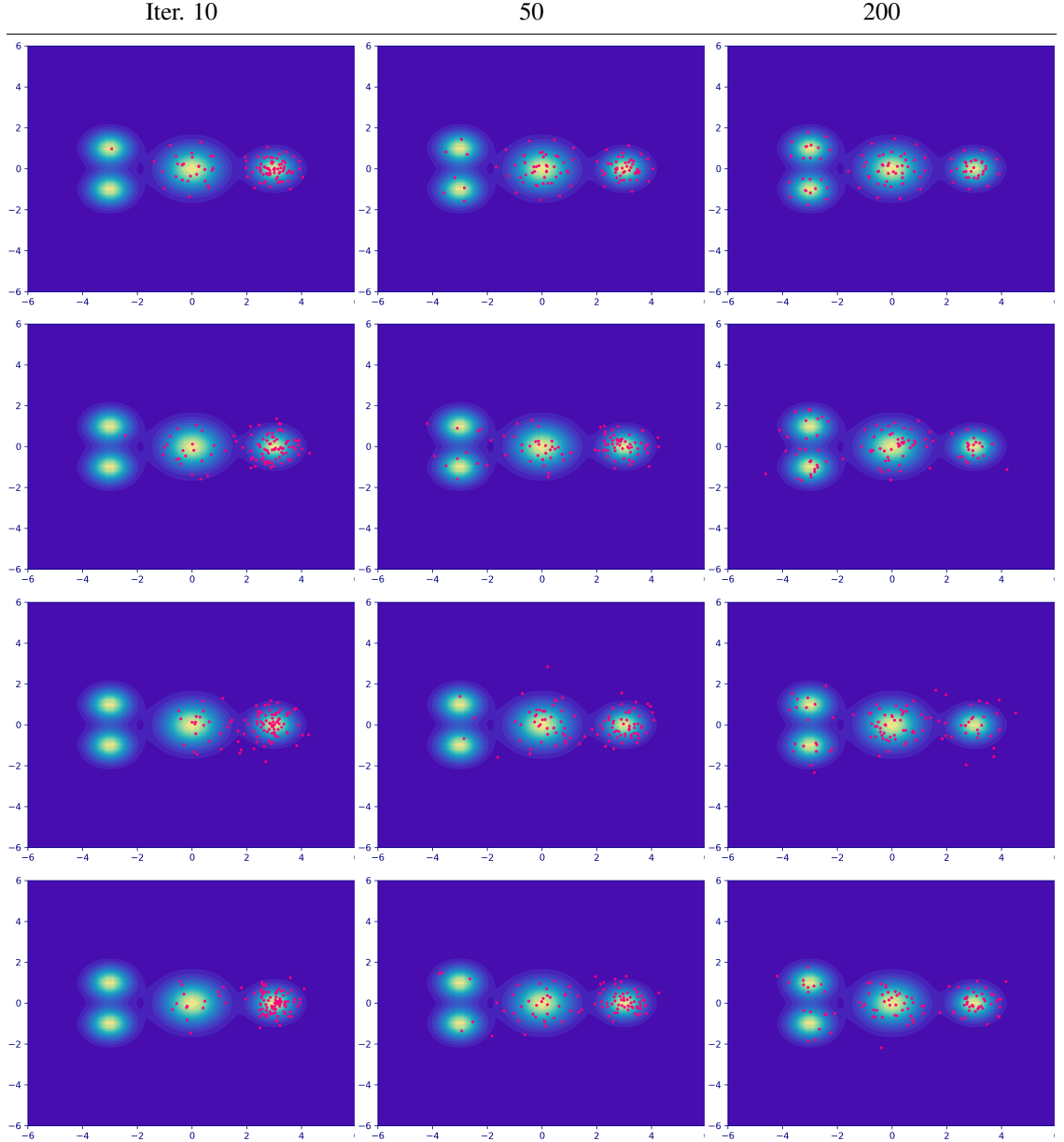
FIGURE 6. Top to bottom: particles for (1) ARWP-Heavy-ball ($\eta = 0.6$, $T = 0.1$, $a = 1$), (2) ILA ($\eta = 0.3$, damping $= 1$), (3) KLMC ($\eta = 1$, $a = 1$), and (4) MALA ($\eta = 0.3$), observed at iterations 10, 50, and 200. We observe reasonable mixing aside from KLMC, and a typical structured phenomenon at iteration 200 for ARWP. KLMC appears to exhibit larger bias around the rightmost well. Note that the particles have not stabilized at this point, and continue to flow from the right potential wells to the left wells.

momentum variable to the score-based ODE. The ARWP method then arises by replacing the score in the latter accelerated information gradient flow with a computationally tractable kernel approximation, given by the regularized Wasserstein proximal operator. For quadratic target potentials, we provide a detailed Lyapunov analysis in terms of the damping parameter in continuous time and an asymptotic discrete-time mixing rate via linearization. Moreover, we achieve a faster asymptotic contraction

rate than that of kinetic Langevin dynamics. Experiments demonstrate better tail exploration than accelerated Langevin methods and the characteristic structured-particle phenomenon.

In a similar vein to the fast iterative shrinkage thresholding algorithm [3], one may ponder whether or not a similar acceleration can hold using the (unregularized) Wasserstein proximal in order to accelerate the Wasserstein proximal gradient method [33]. While a Wasserstein proximal gradient method can be written down using some appropriate exponential maps, and acceleration for geodesically convex functions on manifolds can exist [26], acceleration on Wasserstein manifolds has not been explored in the literature. One possible direction would be to consider applying RWPO within the Wasserstein proximal gradient algorithm [33], relating the RWPO-based methods with classical proximal descent algorithms. The relationship between FISTA with an added score term, with the corresponding dynamics in density space, is also an open question.

References

[1] Kazunori Akiyama, Antxon Alberdi, Walter Alef, Keiichi Asada, Rebecca Azulay, Anne-Kathrin Baczko, David Ball, Mislav Baloković, John Barrett, Dan Bintley, et al. First m87 event horizon telescope results. iv. imaging the central supermassive black hole. *The Astrophysical Journal Letters*, 875(1):L4, 2019.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[4] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[5] Yu Cao, Jianfeng Lu, and Lihan Wang. On explicit l^2-convergence rate estimate for under-damped Langevin dynamics. *Archive for Rational Mechanics and Analysis*, 247(5):90, 2023.

[6] José A Carrillo, Young-Pil Choi, and Oliver Tse. Convergence to equilibrium in Wasserstein distance for damped Euler equations with interaction forces. *Communications in Mathematical Physics*, 365(1):329–361, 2019.

[7] José Antonio Carrillo, Katy Craig, and Francesco S Patacchini. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58:1–53, 2019.

[8] Shi Chen, Qin Li, Oliver Tse, and Stephen J Wright. Accelerating optimization over the space of probability measures. *Journal of machine learning research*, 26(31):1–40, 2025.

[9] Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *Journal of Machine Learning Research*, 21(92):1–72, 2020.

[10] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on learning theory*, pages 300–323. PMLR, 2018.

[11] Katy Craig and Andrea Bertozzi. A blob method for the aggregation equation. *Mathematics of computation*, 85(300):1681–1717, 2016.

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

[13] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019.

[14] Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[15] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Conference on learning theory*, pages 793–797. PMLR, 2018.

[16] Alexander Falk, Andreas Habring, Christoph Griesbacher, and Thomas Pock. An inertial Langevin algorithm. *arXiv preprint arXiv:2510.06723*, 2025.

[17] Fuqun Han, Stanley Osher, and Wuchen Li. Splitting regularized Wasserstein proximal algorithms for nonsmooth sampling problems. *arXiv preprint arXiv:2502.16773*, 2025.

[18] Fuqun Han, Stanley Osher, and Wuchen Li. Tensor train based sampling algorithms for approximating regularized Wasserstein proximal operators. *SIAM/ASA Journal on Uncertainty Quantification*, 13(2):775–804, 2025.

[19] Arieh Iserles. *A first course in the numerical analysis of differential equations*. Number 44. Cambridge university press, 2009.

[20] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[21] Benedict Leimkuhler and Charles Matthews. Rational construction of stochastic numerical methods for molecular sampling. *Applied Mathematics Research eXpress*, 2013(1):34–56, 2013.

[22] Benedict J Leimkuhler, Daniel Paulin, and Peter A Whalley. Contraction and convergence rates for discretized kinetic Langevin dynamics. *SIAM Journal on Numerical Analysis*, 62(3):1226–1258, 2024.

[23] Ruilin Li, Hongyuan Zha, and Molei Tao. Hessian-free high-resolution Nesterov acceleration for sampling. In *International Conference on Machine Learning*, pages 13125–13162. PMLR, 2022.

[24] Wuchen Li, Siting Liu, and Stanley Osher. A kernel formula for regularized Wasserstein proximal operators. *Research in the Mathematical Sciences*, 10(4):43, 2023.

[25] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

[26] Yuanyuan Liu, Fanhua Shang, James Cheng, Hong Cheng, and Licheng Jiao. Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. *Advances in Neural Information Processing Systems*, 30, 2017.

[27] Yi-An Ma, Niladri S Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L Bartlett, and Michael I Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3):1942–1992, 2021.

[28] Edward Nelson. *Dynamical theories of Brownian motion*, volume 3. Princeton university press, 1967.

[29] Yurii Evgen'evich Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.

[30] Filippo Pagani, Martin Wiegand, and Saralees Nadarajah. An n-dimensional Rosenbrock distribution for Markov chain Monte Carlo testing. *Scandinavian Journal of Statistics*, 49(2):657–680, 2022.

[31] Hannes Risken. Fokker-Planck equation. In *The Fokker-Planck equation: methods of solution and applications*, pages 63–95. Springer, 1989.

[32] Peter J Rossky, Jimmie D Doll, and Harold L Friedman. Brownian dynamics as smart Monte Carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633, 1978.

[33] Adil Salim, Anna Korba, and Giulia Luise. The Wasserstein proximal gradient algorithm. *Advances in Neural Information Processing Systems*, 33:12356–12366, 2020.

[34] Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 87. Springer, 2015.

[35] Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[36] Amirhossein Taghvaei and Prashant Mehta. Accelerated flow for probability distributions. In *International conference on machine learning*, pages 6076–6085. PMLR, 2019.

[37] Hong Ye Tan, Stanley Osher, and Wuchen Li. Noise-free sampling algorithms via regularized Wasserstein proximals. *Research in the Mathematical Sciences*, 11(4):65, 2024.

[38] Hong Ye Tan, Stanley Osher, and Wuchen Li. Preconditioned regularized Wasserstein proximal sampling. *arXiv preprint arXiv:2509.01685*, 2025.

[39] Ryan J Tibshirani, Samy Wu Fung, Howard Heaton, and Stanley Osher. Laplace meets Moreau: Smooth approximation to infimal convolutions using Laplace's method. *Journal of Machine Learning Research*, 26(72):1–36, 2025.

[40] Dilin Wang, Ziyang Tang, Chandrajit Bajaj, and Qiang Liu. Stein variational gradient descent with matrix-valued kernels. *Advances in neural information processing systems*, 32, 2019.

[41] Yifei Wang and Wuchen Li. Accelerated information gradient flow. *Journal of Scientific Computing*, 90:1–47, 2022.

[42] Xinzhe Zuo, Stanley Osher, and Wuchen Li. Gradient-adjusted underdamped Langevin dynamics for sampling. *SIAM/ASA Journal on Uncertainty Quantification*, 13(4):1735–1765, 2025.

APPENDIX A. UNDERDAMPED LANGEVIN EQUATION

One possible accelerated counterpart to the Langevin equation is the so-called *kinetic equation* [12]. This corresponds to the *underdamped Langevin diffusion* [10]. This is the equivalent of Nesterov acceleration in the gradient space [41]. We can write down the the standard (overdamped) Langevin diffusion, given by

$$\mathrm{d}X_t = -\nabla V(X)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W\,.$$

The underdamped Langevin dynamics is then given by the following, where X and P are spatial and momentum parameters respectively,

$$\mathrm{d}\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} P \\ -(aP + u\nabla V(X)) \end{bmatrix}\mathrm{d}t + \sqrt{2au}\begin{bmatrix} 0 \\ I \end{bmatrix}\mathrm{d}W\,, \tag{53}$$

where $a > 0$ is a friction coefficient, and $u > 0$ is an inverse mass. In this case, the Brownian motion can be seen to only act on the momentum variable P. In the case $u = 1$, if we scale a to infinity, the limit of the kinetic Langevin dynamics yields the standard overdamped Langevin diffusion [28]. The distribution of this diffusion converges to its invariant distribution in $\mathbb{R}^{2d}$,

$$f_*(x,p) \propto \exp\left(-V(x) - \frac{1}{2u}\|p\|^2\right).$$

The corresponding accelerated Fokker–Planck equation is known as the *Klein–Kramers equation*, which is an evolution of the joint density in phase space $f(x,p)$. The update is given by the second-order update [31]

$$\partial_t f + p \cdot \nabla_x f - u^{-1}\nabla_p f \cdot \nabla V(x) = a\nabla_p \cdot (pf) + u^{-1}a\Delta_p^2 f. \tag{54}$$

Various convergence results can be found in [22]. This can also be seen as Hamilton's equations corresponding to the Hamiltonian

$$H(x,p) = V(x) + \frac{1}{2u}\|p\|^2.$$

An alternative is given by considering a different accelerated Fokker–Planck equation, with the same stationary distribution. In the continuous-time optimization setting, an accelerated gradient flow is given by

$$\dot{x} = p, \quad \dot{p} = -ap - \nabla V(x).$$

The analogous dynamics in the probability space are given by [8], referred to by the authors as *heavy-ball flow*,

$$\partial_t f + p \cdot \nabla_x f - \nabla_p \cdot \left(\left(ap + \nabla_x \frac{\delta E}{\delta \rho}\left[\int_{\mathbb{R}^d} f(x,p)\,\mathrm{d}p\right]\right)f\right) = 0. \tag{55}$$

Here, E is some divergence or metric to the stationary distribution, such as the relative entropy/KL divergence, and $\delta E/\delta\rho$ represents the first variation. The main difference with the Klein–Kramers equation is the second order term: instead of having a Laplacian in momentum space $\Delta_p f_t(x,p)$ over the joint density $f(x,p)$, one has a mixed gradient $\nabla_p \cdot (f\nabla_x)$ over the marginal $\int f(x,p)\,\mathrm{d}p$. In the case $E(\rho) = \mathrm{D}_{\mathrm{KL}}(\rho\|\pi)$, writing ρ_t for the marginal over p, the heavy-ball flow is specialized as

$$\partial_t f + p \cdot \nabla_x f - \nabla_p \cdot [(ap + \nabla V(x) + \nabla \log \rho_t(x))f] = 0. \tag{56}$$

This equation describes the phase-space measure corresponding the following particle evolution [8, Eq. 94]

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} P \\ -aP - \nabla V(X) - \nabla \log \rho_t(X) \end{bmatrix}. \tag{57}$$

Other than considering this as an analogue of Nesterov acceleration in measure space, another method of deriving this equation Eq. (55) is by damping an appropriate Hamiltonian flow in the Wasserstein-2 space. This is interpreted by [41] as arising from a measure-valued analog of Nesterov acceleration using the Wasserstein metric. See more details in [41].

APPENDIX B. PROOFS

B.1. **Convergence rate of linearized discrete time update.** This section shows the convergence rate given in Section 4.3. Recall the update matrix (in each dimension) is given by

$$\begin{pmatrix} \delta_{n+1} \\ g_{n+1} \end{pmatrix} = [I + \eta A] \begin{pmatrix} \delta_n \\ g_n \end{pmatrix}, \quad A = \begin{bmatrix} 0 & 2\lambda - 4Tk_+^{-1} \\ -\lambda^{-2} & -a \end{bmatrix}. \tag{58}$$

The eigenvalues of A are

$$\chi_{\pm} = \frac{1}{2}[\mathrm{Tr} \pm \sqrt{\mathrm{Tr} - 4\det}]$$

$$= \frac{1}{2}\left[-a \pm \sqrt{a^2 - 4\lambda^{-2}(2\lambda - 4TK_+^{-1})}\right].$$

and the step-size has to be chosen such that for every eigenvalue, $|1 + \eta\chi_{\pm}| < 1$. The contraction/convergence rate is the largest of the values $|1 + \eta\chi_{\pm}|$ over all eigenvalues. We recall a technical assumption that $T \leq (1 + \sqrt{2})^{-1}\lambda_{\min}$. This ensures that the function $\lambda \mapsto \lambda^{-1}\frac{\lambda - T}{\lambda + T}$ is (strictly) decreasing over $[\lambda_{\min}, \lambda_{\max}]$.

B.1.1. *All complex eigenvalues, low critical damping.* This happens if for each eigenvalue λ of Λ,

$$a^2 \leq 8\lambda^{-1}(1 - 2Tk_+^{-1}\lambda^{-1}). \tag{59}$$

The step-size condition on η is

$$\left|1 + \frac{\eta}{2}\left[-a \pm \sqrt{a^2 - 4\lambda^{-2}(2\lambda - 4TK_+^{-1})}\right]\right| < 1.$$

The absolute value is less than 1 if and only if

$$(1 - \frac{a\eta}{2})^2 + \frac{\eta^2}{4}(-a^2 + 4\lambda^{-2}(2\lambda - 4Tk_+^{-1})) < 1$$

$$\Leftrightarrow 1 - a\eta + \eta^2\lambda^{-2}(2\lambda - 4Tk_+^{-1}) < 1.$$

Rearranging, the step-size condition that yields convergence is

$$\eta < \frac{a}{2\lambda^{-1}(1 - 2Tk_+^{-1}\lambda^{-1})}.$$

The rate is fastest when the norm of $1 + \eta\chi_{\pm}$ is minimized, which occurs at

$$\eta = \frac{a}{4\lambda^{-1}(1 - 2Tk_+^{-1}\lambda^{-1})}. \tag{60}$$

Since this has to be true for every eigenvalue, it is sufficient (and necessary) for (59) to be true for $\lambda_{\max}$. The maximal step-size is also given by the smallest value of (60), i.e. for $\lambda_{\min}$. This yields the parameters

$$a = 2\sqrt{2}\lambda_{\max}^{-1/2}\sqrt{1 - 2Tk_{+,\max}^{-1}\lambda_{\max}^{-1}}, \quad \eta = \frac{1}{\sqrt{2}}\lambda_{\max}^{-1/2}\lambda_{\min}\frac{\sqrt{1 - 2Tk_{+,\max}^{-1}\lambda_{\max}^{-1}}}{1 - 2Tk_{+,\min}^{-1}\lambda_{\min}^{-1}}.$$

To obtain the form in Proposition 2, one may observe that

$$1 - 2Tk_+^{-1}\lambda^{-1} = \frac{\lambda - T}{\lambda + T}.$$

Then, the damping and optimal step-size are given by

$$a = 2\sqrt{2}\lambda_{\max}^{-1/2}\sqrt{\frac{\lambda_{\max} - T}{\lambda_{\max} + T}}, \quad \eta = \frac{1}{\sqrt{2}}\lambda_{\max}^{-1/2}\lambda_{\min}\sqrt{\frac{\lambda_{\max} - T}{\lambda_{\max} + T}\frac{\lambda_{\min} + T}{\lambda_{\min} - T}}.$$

The rate is given when the complex part is largest, i.e. when λ is minimized. It is given by

$$\max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \left| 1 + \frac{\eta}{2}\left[-a \pm \sqrt{a^2 - 4\lambda^{-2}(2\lambda - 4TK_+^{-1})} \right] \right|$$

$$= \sqrt{1 - a\eta + 2\eta^2\lambda_{\min}^{-1}(1 - 2Tk_{+,\min}^{-1}\lambda_{\min}^{-1})}$$

$$= \sqrt{1 - \kappa^{-1}\frac{\lambda_{\max} - T}{\lambda_{\max} + T}\frac{\lambda_{\min} + T}{\lambda_{\min} - T}}.$$

B.1.2. *All real eigenvalues, high critical damping.* This happens if for each eigenvalue λ,

$$a^2 > 8\lambda^{-1}(1 - 2Tk_+^{-1}\lambda^{-1}). \tag{61}$$

Since both eigenvalues are less than 0 and $\chi_- < \chi_+ < 0$, we need only check that $1 + \eta\chi_- > -1$. The step-size condition on η such that the iteration is stable/convergent, thus becomes

$$1 + \frac{\eta}{2}[-a - \sqrt{a^2 - 4\lambda^{-2}(2\lambda - 4TK_+^{-1})}] > -1$$

The condition for stability satisfies

$$\eta \le \frac{4}{a + \sqrt{a^2 - 4\lambda^{-2}(2\lambda - 4TK_+^{-1})}}$$

We can now consider the largest possible optimal a, which corresponds to $\lambda_{\min}$,

$$a = 2\sqrt{2}\lambda_{\min}^{-1/2}\sqrt{\frac{\lambda_{\min} - T}{\lambda_{\min} + T}}. \tag{62}$$

Using this damping parameter, we have that

$$a \ge 2\sqrt{2}\lambda^{-1/2}\sqrt{\frac{\lambda - T}{\lambda + T}}, \tag{63}$$

for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, and all the eigenvalues of the update matrix $I + \eta A$ are real. Moreover, if the step-size satisfies

$$\eta \le \frac{2}{a},$$

then we have that $|1 + \eta\chi_+| \ge |1 + \eta\chi_-|$. This can be seen by solving the equality $1 + \eta\chi_+ = -(1 + \eta\chi_-)$. Therefore, the rate for a given eigenvalue is given by

$$1 + \eta\chi_+ = 1 - \frac{\eta}{2}\left[a - \sqrt{a^2 - 8\lambda^{-1}\frac{\lambda - T}{\lambda + T}} \right]. \tag{64}$$

Up to the first order, this rate is controlled by the following lemma.

Lemma 2. *For a constant $c > 0$, the function*

$$q : [\sqrt{c}, \infty) \to \mathbb{R}, \quad q(x) = x - \sqrt{x^2 - c}, \tag{65}$$

can be rewritten as

$$q(x) = \frac{c}{x + \sqrt{x^2 - c}} > \frac{c}{2x}. \tag{66}$$

If we take $\eta = 2/a$, the eigenvalues for convergence are all real, and the rate corresponding to an eigenvalue $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ is given by,

$$1 - \frac{\eta}{2}\left[a - \sqrt{a^2 - 8\lambda^{-1}\frac{\lambda - T}{\lambda + T}} \right].$$

This rate is slowest (i.e. largest) when λ is maximized. To summarize, the choice of damping a and step-size is

$$a = 2\sqrt{2}\lambda_{\min}^{-1/2}\sqrt{\frac{\lambda_{\min} - T}{\lambda_{\min} + T}}, \quad \eta = \frac{2}{a}.$$

This choice of parameters yields the rate

$$1 - \frac{\eta}{2}\left[a - \sqrt{a^2 - 8\lambda_{\max}^{-1}\frac{\lambda_{\max} - T}{\lambda_{\max} + T}}\right] = 1 - a^{-1}\left[a - \sqrt{a^2 - 8\lambda_{\max}^{-1}\frac{\lambda_{\max} - T}{\lambda_{\max} + T}}\right]$$

$$= \sqrt{1 - \kappa^{-1}\frac{\lambda_{\max} - T}{\lambda_{\max} + T}\frac{\lambda_{\min} + T}{\lambda_{\min} - T}}.$$

APPENDIX C. CONVERGENCE OF ACCELERATED REGULARIZED CONTINUOUS-TIME UPDATE

Recall the continuous-time covariance update, given by the coupled ODEs Equation (42),

$$\begin{cases} \dot{\tilde{\sigma}}_t = 2g_t\tilde{\sigma}_t - 4g_t T k_+^{-1}, \\ \dot{g}_t = -ag_t - g_t^2 - \lambda^{-1} + \tilde{\sigma}_t^{-1}. \end{cases} \tag{67}$$

This differs from the linearized case due to the introduction of the g_t^2 term, which changes the dynamics away from $g_t \approx 0$. Moreover, we have a forcing term within $\dot{\tilde{\sigma}}_t$.

Let us additionally define the time-dependent variable

$$b_\pm = \lambda^{-1/2} \pm \tilde{\sigma}_t^{-1/2}. \tag{68}$$

For ease of notation, we denote the KL divergence between two zero-mean Gaussians directly using their covariances, $\mathrm{D_{KL}}(\Sigma_1, \Sigma_2) = \mathrm{D_{KL}}(\mathcal{N}(0, \Sigma_1), \mathcal{N}(0, \Sigma_2))$. We define a Lyapunov function by

$$\mathcal{E}_t = (\tilde{\sigma}_t - 2T k_+^{-1})[b_- + g_t]^2 + 2\mathrm{D_{KL}}(\tilde{\sigma}_t, \lambda). \tag{69}$$

From [41, Prop. 8], we have a specialized bound, stronger than the traditional log-Sobolev inequality using the Bakry–Emery criterion. In the case where E is the relative entropy, so that the first variation satisfies $\delta E(\rho_t, \rho_*)/\delta\rho_t = \log(\rho_t/\rho_*) + 1$. One has the stronger bound

$$\mathrm{D_{KL}}(\tilde{\sigma}, \lambda) \leq \tilde{\sigma}\sqrt{\lambda}b_-^2 b_+ - \frac{1}{2}\tilde{\sigma}b_-^2. \tag{70}$$

The proof by specializing this result to Gaussians is delayed until Section C.1, given in Corollary 2.

It remains to compute the time derivative of $\mathcal{E}_t$, which we wish to show is negative. We have the following expressions for the time derivative of $\mathcal{E}_t$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}2\mathrm{D_{KL}}(\tilde{\sigma}_t, \lambda) &= \frac{\mathrm{d}}{\mathrm{d}t}(\tilde{\sigma}_t\lambda^{-1} - \log(\tilde{\sigma}_t\lambda^{-1}) - 1) \\ &= \dot{\tilde{\sigma}}_t(\lambda^{-1} - \tilde{\sigma}_t^{-1}) \\ &= (2g_t\tilde{\sigma}_t - 4g_t T k_+^{-1})(\lambda^{-1} - \tilde{\sigma}_t^{-1}), \end{aligned}$$

In addition,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(b_- + g_t) &= \frac{\mathrm{d}}{\mathrm{d}t}(\lambda^{-1/2} - \tilde{\sigma}_t^{-1/2} + g_t) \\ &= -ag_t - g_t^2 - \lambda^{-1} + \tilde{\sigma}_t^{-1} + \frac{1}{2}\dot{\tilde{\sigma}}_t\tilde{\sigma}_t^{-3/2} \\ &= -ag_t - g_t^2 - \lambda^{-1} + \tilde{\sigma}_t^{-1} + g_t\tilde{\sigma}_t^{-1/2} - 2Tk_+^{-1}g_t\tilde{\sigma}_t^{-3/2}. \end{aligned}$$

We can therefore compute the time derivative $\dot{\mathcal{E}}_t$ as follows:

$$
\begin{aligned}
\dot{\mathcal{E}}_t ={}& 2g(\tilde{\sigma} - 2Tk_+^{-1})[b_- + g]^2 \\
&+ 2(\tilde{\sigma} - 2Tk_+^{-1})(b_- + g)(-ag - g^2 - \lambda^{-1} + \tilde{\sigma}^{-1} + g\tilde{\sigma}^{-1/2} - 2Tk_+^{-1}g\tilde{\sigma}^{-3/2}) \\
&+ 2g(\tilde{\sigma} - 2Tk_+^{-1})(\lambda^{-1} - \tilde{\sigma}^{-1}) \\
={}& 2(\tilde{\sigma} - 2Tk_+^{-1})(b_- + g)(-ag - g^2 + g(b_- + g) + g\tilde{\sigma}^{-1/2} - 2Tk_+^{-1}g\tilde{\sigma}^{-3/2}) \\
&+ 2(\tilde{\sigma} - 2Tk_+^{-1})(b_-)(-\lambda^{-1} + \tilde{\sigma}^{-1}) \\
={}& -2(\tilde{\sigma} - 2Tk_+^{-1})b_-^2 b_+ \\
&+ 2g(\tilde{\sigma} - 2Tk_+^{-1})(b_- + g)(-a + \lambda^{-1/2} - 2Tk_+^{-1}\tilde{\sigma}^{-3/2}).
\end{aligned}
\tag{71}
$$

From the bound (70) on $\mathrm{D_{KL}}(\tilde{\sigma}_t, \lambda)$,

$$
\mathcal{E}_t \leq (\tilde{\sigma} - 2Tk_+^{-1})[b_- + g]^2 + 2\tilde{\sigma}\sqrt{\lambda}b_-^2 b_+ - \tilde{\sigma}b_-^2.
$$

Rearranging,

$$
-2\tilde{\sigma}\sqrt{\lambda}b_-^2 b_+ \leq -\mathcal{E}_t + (\tilde{\sigma} - 2Tk_+^{-1})[b_- + g]^2 - \tilde{\sigma}b_-^2.
\tag{72}
$$

Substituting into (71), and noting that $\tilde{\sigma} - 2Tk_+^{-1}$ and b_+ are positive,

$$
\begin{aligned}
\dot{\mathcal{E}}_t ={}& -2(\tilde{\sigma} - 2Tk_+^{-1})b_-^2 b_+ \\
&+ 2g(\tilde{\sigma} - 2Tk_+^{-1})(b_- + g)[-a + \lambda^{-1/2} - 2Tk_+^{-1}\tilde{\sigma}^{-3/2}] \\
\leq{}& \left(\frac{2\tilde{\sigma} - 4Tk_+^{-1}}{2\tilde{\sigma}\sqrt{\lambda}}\right)\left[-\mathcal{E}_t + (\tilde{\sigma} - 2Tk_+^{-1})[B_- + g]^2 - \tilde{\sigma}b_-^2\right] \\
&+ 2g(\tilde{\sigma} - 2Tk_+^{-1})(b_- + g)[-a + \lambda^{-1/2} - 2Tk_+^{-1}\tilde{\sigma}^{-3/2}] \\
={}& (\lambda^{-1/2} - 2\tilde{\sigma}^{-1}Tk_+^{-1}\lambda^{-1/2})\left[-\mathcal{E}_t + (\tilde{\sigma} - 2Tk_+^{-1})[b_- + g]^2 - \tilde{\sigma}b_-^2\right] \\
&+ 2g(\tilde{\sigma} - 2Tk_+^{-1})(b_- + g)[-a + \lambda^{-1/2} - 2Tk_+^{-1}\tilde{\sigma}^{-3/2}] \\
={}& -(\lambda^{-1/2} - 2\tilde{\sigma}^{-1}Tk_+^{-1}\lambda^{-1/2})\mathcal{E}_t \tag{73a} \\
&+ \lambda^{-1/2}\tilde{\sigma}^{-1}(\tilde{\sigma} - 2Tk_+^{-1})[(\tilde{\sigma} - 2Tk_+^{-1})[b_- + g]^2 - \tilde{\sigma}b_-^2] \tag{73b} \\
&+ 2g(\tilde{\sigma} - 2Tk_+^{-1})(b_- + g)[-a + \lambda^{-1/2} - 2Tk_+^{-1}\tilde{\sigma}^{-3/2}]. \tag{73c}
\end{aligned}
$$

C.0.1. *One-dimensional case: critical momentum.* We now control the latter two terms (73b) and (73c), by showing their sum is negative. Then, we can use Grönwall's inequality to conclude.

$$
\begin{aligned}
&\lambda^{-1/2}\tilde{\sigma}^{-1}(\tilde{\sigma} - 2Tk_+^{-1})[(\tilde{\sigma} - 2Tk_+^{-1})[b_- + g]^2 - \tilde{\sigma}b_-^2] \\
&+ 2g(\tilde{\sigma} - 2Tk_+^{-1})(b_- + g)[-a + \lambda^{-1/2} - 2Tk_+^{-1}\tilde{\sigma}^{-3/2}] \\
={}& (\tilde{\sigma} - 2Tk_+^{-1})\begin{bmatrix} \lambda^{-1/2}\tilde{\sigma}^{-1}[2\tilde{\sigma}b_- g + \tilde{\sigma}g^2 - 2Tk_+^{-1}(b_- + g)^2] \\ +2g(b_- + g)[-a + \lambda^{-1/2} - 2Tk_+^{-1}\tilde{\sigma}^{-3/2}] \end{bmatrix} \\
={}& (\tilde{\sigma} - 2Tk_+^{-1})\begin{bmatrix} (2\lambda^{-1/2} - 2a + 2\lambda^{-1/2})b_- g + (\lambda^{-1/2} - 2a + 2\lambda^{-1/2})g^2 \\ -2Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1}(b_- + g)^2 - 4Tk_+^{-1}g(b_- + g)\tilde{\sigma}^{-3/2} \end{bmatrix} \\
={}& (\tilde{\sigma} - 2Tk_+^{-1})\begin{bmatrix} -\lambda^{-1/2}g^2 \\ -2Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1}(b_- + g)^2 - 4Tk_+^{-1}g(b_- + g)\tilde{\sigma}^{-3/2} \end{bmatrix},
\end{aligned}
$$

where the last equality holds if we take the momentum parameter

$$
a = 2\lambda^{-1/2}.
\tag{74}
$$

It remains to use the control on g^2 and $(b_- + g)^2$ to bound the final $g(b_- + g)$ term. The component inside the bracket is a quadratic in g:

$$-\lambda^{-1/2}g^2 - 2Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1}(b_- + g)^2 - 4Tk_+^{-1}g(b_- + g)\tilde{\sigma}^{-3/2}$$

$$= g^2(-\lambda^{-1/2} - 2Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1} - 4Tk_+^{-1}\tilde{\sigma}^{-3/2})$$

$$+ gb_-(-4Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1} - 4Tk_+^{-1}\tilde{\sigma}^{-3/2}) - 2Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1}b_-^2.$$

The coefficient of g^2 is negative. Maximizing over all possible g, the above expression is upper bounded by

$$-2Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1}b_-^2 - \frac{b_-^2(-4Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1} - 4Tk_+^{-1}\tilde{\sigma}^{-3/2})^2}{4(-\lambda^{-1/2} - 2Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1} - 4Tk_+^{-1}\tilde{\sigma}^{-3/2})}$$

$$= \frac{1}{4(-\lambda^{-1/2} - 2Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1} - 4Tk_+^{-1}\tilde{\sigma}^{-3/2})}$$

$$\cdot \left[-8Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1}b_-^2(-\lambda^{-1/2} - 2Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1} - 4Tk_+^{-1}\tilde{\sigma}^{-3/2}) \right.$$

$$\left. - b_-^2(-4Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1} - 4Tk_+^{-1}\tilde{\sigma}^{-3/2})^2 \right]$$

$$= \frac{b_-^2}{4c_t} \cdot \left[8Tk_+^{-1}\lambda^{-1}\tilde{\sigma}^{-1} + 16T^2k_+^{-2}\lambda^{-1}\tilde{\sigma}^{-2} + 32T^2k_+^{-2}\lambda^{-1/2}\tilde{\sigma}^{-5/2}) \right.$$

$$\left. - 16\left(T^2k_+^{-2}\lambda^{-1}\tilde{\sigma}^{-2} + 2T^2k_+^{-2}\lambda^{-1/2}\tilde{\sigma}^{-5/2} + T^2k_+^{-2}\tilde{\sigma}^{-3} \right) \right]$$

$$= \frac{b_-^2}{4c_t}\left[8Tk_+^{-1}\lambda^{-1}\tilde{\sigma}^{-1} - 16T^2k_+^{-2}\tilde{\sigma}^{-3} \right]. \tag{75}$$

where c_t indicates the negative denominator $c_t = (-\lambda^{-1/2} - 2Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1} - 4Tk_+^{-1}\tilde{\sigma}^{-3/2})$. This quantity is negative if the term in (75) is positive. This is the case if the following relationship on $\tilde{\sigma}$ and T holds:

$$8TK_+^{-1}\lambda^{-1}\tilde{\sigma}^{-1} - 16T^2K_+^{-2}\tilde{\sigma}^{-3} \geq 0$$

$$\Leftrightarrow \lambda^{-1} - 2TK_+^{-1}\tilde{\sigma}^{-2} \geq 0$$

$$\Leftrightarrow \tilde{\sigma}^2 \geq 2TK_+^{-1}\lambda. \tag{76}$$

This states that the variance of the regularized Wasserstein proximal can not be too small, or that the regularization has to be chosen to be sufficiently small.

As a sanity check, we may verify that for the choice $T < \lambda$, this inequality holds near the terminal variance $\tilde{\sigma} \approx \lambda$. The necessary condition becomes

$$\lambda^2 \geq 2TK_+^{-1}\lambda$$

$$\Leftrightarrow \lambda \geq 2TK_+^{-1} = \frac{2T}{1 + T\lambda^{-1}} = \lambda + \frac{T - \lambda}{1 + T\lambda^{-1}},$$

which is equivalent to $T \leq \lambda$, as initially assumed.

Returning to Equation (73a), we have shown that under the assumption (76) on the variance and regularization, and using the momentum parameter choice $a = 2\lambda^{-1/2}$ in (74), the Grönwall-type inequality holds:

$$\dot{\mathcal{E}}_t \leq -(\lambda^{-1/2} - 2Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1})\mathcal{E}_t$$

$$+ (\tilde{\sigma} - 2Tk_+^{-1})b_-^2 \frac{8Tk_+^{-1}\lambda^{-1}\tilde{\sigma}^{-1} - 16T^2k_+^{-2}\tilde{\sigma}^{-3}}{4(-\lambda^{-1/2} - 2Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1} - 4Tk_+^{-1}\tilde{\sigma}^{-3/2})}$$

$$\leq -(\lambda^{-1/2} - 2Tk_+^{-1}\lambda^{-1/2}\tilde{\sigma}^{-1})\mathcal{E}_t.$$

In particular, close to the terminal distribution $\tilde{\sigma} \approx \lambda$, one has the asymptotic rate

$$\mathcal{E}_t = \mathcal{O}(e^{-rt}), \tag{77}$$

where the rate is

$$\begin{aligned}
r &= \lambda^{-1/2} - 2Tk_+^{-1}\lambda^{-1/2}\lambda^{-1} \\
&= \left(1 - 2Tk_+^{-1}\lambda^{-1}\right)\lambda^{-1/2} \\
&= \left(\frac{\lambda - T}{\lambda + T}\right)\lambda^{-1/2}.
\end{aligned}$$

C.0.2. *Subcritical damping, removing the step-size condition.* Consider $a \in (\lambda^{-1/2}, 2\lambda^{-1/2}]$. We wish to drop the condition that $\tilde{\sigma}$ is bounded below when deriving our rate, in order to get a global convergence.

$$\mathcal{F}_t = (\tilde{\sigma}_t - 2Tk_+^{-1})[b_- + g_t]^2 + 2\chi D_{\mathrm{KL}}(\tilde{\sigma}_t, \lambda),$$

and $\chi = \chi(a, \lambda, T) > 0$ is to be determined. Taking the time derivative of $\mathcal{F}_t$, one gets

$$\begin{aligned}
\frac{\dot{\mathcal{F}}_t}{2(\tilde{\sigma} - 2Tk_+^{-1})} &= g[b_- + g]^2 \\
&\quad + (b_- + g)(-ag - g^2 - \lambda^{-1} + \tilde{\sigma}^{-1} + g\tilde{\sigma}^{-1/2} - 2Tk_+^{-1}g\tilde{\sigma}^{-3/2}) \\
&\quad + \chi g(\lambda^{-1} - \tilde{\sigma}^{-1}) \\
&= (b_- + g)\left(gb_- + g^2 - ag - g^2 - b_- b_+ + g\tilde{\sigma}^{-1/2}(1 - 2Tk_+^{-1}\tilde{\sigma}^{-1})\right) \\
&\quad + \chi g b_- b_+ \\
&= (b_- + g)(g(b_- - a) - b_- b_+ + g\tilde{\sigma}^{-1/2}(1 - 2Tk_+^{-1}\tilde{\sigma}^{-1})) + \chi g b_- b_+ \\
&= (b_- + g)(g(b_- - a) + g\tilde{\sigma}^{-1/2}(1 - 2Tk_+^{-1}\tilde{\sigma}^{-1})) + (\chi - 1)g b_- b_+ \\
&\quad - b_-^2 b_+.
\end{aligned}$$

Note since $a \geq \lambda^{-1/2}$, that $b_- - a < 0$. Moreover, since

$$D_{\mathrm{KL}}(\tilde{\sigma}, \lambda) \leq \tilde{\sigma}\sqrt{\lambda}b_-^2 b_+ - \frac{1}{2}\tilde{\sigma}b_-^2, \tag{78}$$

we have

$$\begin{aligned}
-2\chi\tilde{\sigma}\sqrt{\lambda}b_-^2 b_+ &\leq -\mathcal{F}_t + (\tilde{\sigma} - 2Tk_+^{-1})[b_- + g]^2 - \tilde{\sigma}\chi b_-^2 \\
\Leftrightarrow \tilde{\sigma}\chi(1 - 2\sqrt{\lambda}b_+)b_-^2 &\leq -\mathcal{F}_t + (\tilde{\sigma} - 2Tk_+^{-1})[b_- + g]^2 \\
\Leftrightarrow -b_-^2 &\leq \frac{-\mathcal{F}_t + (\tilde{\sigma} - 2Tk_+^{-1})[b_- + g]^2}{\chi\tilde{\sigma}(2\sqrt{\lambda}b_+ - 1)},
\end{aligned}$$

where the equivalence is since $1 - 2\sqrt{\lambda}b_+ = -1 - 2\tilde{\sigma}^{-1/2}\lambda^{1/2} < 0$. It remains to use the term $b_-^2 b_+$ to determine the rate. Let $r > 0$ be some rate parameter. Then,

$$\begin{aligned}
\frac{\dot{\mathcal{F}}_t}{2(\tilde{\sigma} - 2Tk_+^{-1})} &\leq (b_- + g)(g(b_- - a) + g\tilde{\sigma}^{-1/2}(1 - 2Tk_+^{-1}\tilde{\sigma}^{-1})) + (\chi - 1)g b_- b_+ \\
&\quad + rb_+\frac{-\mathcal{F}_t + (\tilde{\sigma} - 2Tk_+^{-1})[b_- + g]^2}{\chi\tilde{\sigma}(2\sqrt{\lambda}b_+ - 1)} - (1 - r)b_-^2 b_+ \\
&= -\frac{rb_+}{\chi\tilde{\sigma}(2\sqrt{\lambda}b_+ - 1)}\mathcal{F}_t \\
&\quad + g(b_- + g)(b_- - a + \tilde{\sigma}^{-1/2}(1 - 2Tk_+^{-1}\tilde{\sigma}^{-1})) + (\chi - 1)g b_- b_+ \\
&\quad + rb_+\frac{(\tilde{\sigma} - 2Tk_+^{-1})[b_- + g]^2}{\chi\tilde{\sigma}(2\sqrt{\lambda}b_+ - 1)} - (1 - r)b_-^2 b_+. \tag{79}
\end{aligned}$$

It remains to find a maximal $r = r_t > 0$, such that the sum of the last two terms is always negative. Considering it as a quadratic in $b_- + g$, one has

$$g(b_- + g)(b_- - a + \tilde{\sigma}^{-1/2}(1 - 2Tk_+^{-1}\tilde{\sigma}^{-1})) + (\chi - 1)gb_-b_+$$

$$+ rb_+ \frac{(\tilde{\sigma} - 2Tk_+^{-1})[b_- + g]^2}{\chi\tilde{\sigma}(2\sqrt{\lambda}b_+ - 1)} - (1 - r)b_-^2 b_+$$

$$= (b_- + g)^2(b_- - a + \tilde{\sigma}^{-1/2}(1 - 2Tk_+^{-1}\tilde{\sigma}^{-1})) - b_-(b_- + g)(b_- - a + \tilde{\sigma}^{-1/2}(1 - 2Tk_+^{-1}\tilde{\sigma}^{-1}))$$

$$+ (\chi - 1)(b_- + g)b_-b_+ - (\chi - 1)b_-^2 b_+$$

$$+ rb_+ \frac{(\tilde{\sigma} - 2Tk_+^{-1})}{\chi\tilde{\sigma}(2\sqrt{\lambda}b_+ - 1)}[b_- + g]^2 - (1 - r)b_-^2 b_+.$$

This is a quadratic $c_2(b_- + g)^2 + c_1(b_- + g) + c_0$, where

$$c_2 = (b_- - a + \tilde{\sigma}^{-1/2}(1 - 2Tk_+^{-1}\tilde{\sigma}^{-1})) + rb_+ \frac{(\tilde{\sigma} - 2Tk_+^{-1})}{\chi\tilde{\sigma}(2\sqrt{\lambda}b_+ - 1)},$$

$$c_1 = (\chi - 1)b_-b_+ - b_-(b_- - a + \tilde{\sigma}^{-1/2}(1 - 2Tk_+^{-1}\tilde{\sigma}^{-1})),$$

$$c_0 = -(\chi - r)b_-^2 b_+.$$

We wish to show that $c_2 < 0$, and furthermore that the maximum $c_0 - \frac{c_1^2}{4c_2}$ is negative. Equivalently, $c_1^2 - 4c_0c_2 < 0$.

Condition 1: quadratic coefficient is negative. The equivalent condition for this to hold is that

$$(b_- - a + \tilde{\sigma}^{-1/2}(1 - 2Tk_+^{-1}\tilde{\sigma}^{-1})) + rb_+ \frac{(\tilde{\sigma} - 2Tk_+^{-1})}{\chi\tilde{\sigma}(2\sqrt{\lambda}b_+ - 1)} < 0$$

$$\Leftrightarrow r\chi^{-1} \leq \frac{\tilde{\sigma}(a - b_- - \tilde{\sigma}^{-1/2}(1 - 2Tk_+^{-1}\tilde{\sigma}^{-1}))(2\sqrt{\lambda}b_+ - 1)}{b_+(\tilde{\sigma} - 2Tk_+^{-1})}. \tag{80}$$

The RHS is positive for all $\tilde{\sigma}$ if $a > \lambda^{-1/2}$.

Condition 2: quadratic is upper bounded by 0. Let $p = p_t := a - b_- - \tilde{\sigma}^{-1/2}(1 - 2Tk_+^{-1}\tilde{\sigma}^{-1}) = a - \lambda^{-1/2} + 2Tk_+^{-1}\tilde{\sigma}^{-3/2} > 0$. Rewriting the coefficients, we obtain

$$c_2 = -p + rb_+ \frac{\tilde{\sigma} - 2Tk_+^{-1}}{\chi\tilde{\sigma}(2\sqrt{\lambda}b_+ - 1)}, \tag{81}$$

$$c_1 = (\chi - 1)b_-b_+ + b_-p, \tag{82}$$

$$c_0 = -(\chi - r)b_-^2 b_+. \tag{83}$$

The equivalent condition is

$$0 \geq c_1^2 - 4c_0c_2 \tag{84}$$

$$\Leftrightarrow 0 \geq ((\chi - 1)b_+ + p)^2 - 4(-p + rb_+ \frac{\tilde{\sigma} - 2Tk_+^{-1}}{\chi\tilde{\sigma}(2\sqrt{\lambda}b_+ - 1)})(-(\chi - r)b_+). \tag{85}$$

This inequality has to be strict at $r = 0$ for a feasible rate to exist. One obtains the simplified quadratic inequality:

$$0 \geq ((\chi - 1)b_+ + p)^2 - 4(-p)(-\chi b_+).$$

One immediately observes now that if $a \in (\lambda^{-1/2}, 2\lambda^{-1/2}]$, then

$$p = a - \lambda^{-1/2} + 2Tk_+^{-1}\tilde{\sigma}^{-3/2} \leq \lambda^{-1/2} + 2Tk_+^{-1}\tilde{\sigma}^{-3/2} < b_+. \tag{86}$$

Therefore, taking $\chi = 1$ yields that the quadratic inequality holds strictly,

$$p^2 - 4pb_+ < 0.$$

One may now solve for r in Equation (85). Substituting $\chi = 1$, the quadratic inequality becomes

$$(p)^2 - 4(-p + rb_+ \frac{\tilde{\sigma} - 2Tk_+^{-1}}{\tilde{\sigma}(2\sqrt{\lambda}b_+ - 1)})(-(1-r)b_+)$$

This is a quadratic with positive coefficient in r^2 and is negative at $r = 0$. Since

$$\frac{p\tilde{\sigma}(2\sqrt{\lambda}b_+ - 1)}{b_+(\tilde{\sigma} - 2Tk_+^{-1})}, \quad 1 \tag{87}$$

are positive, the inflection point is also positive. Therefore, the rate (that also satisfies Equation (80)) is given by the smallest (positive) root. Since the quadratic is negative at $r = 0$, by the intermediate value theorem (IVT), the positive root must be smaller than the (positive) quantity in (80).

C.0.3. *Multi-dimensional case: overdamping.* Let $\zeta > 0$ be a constant to be chosen later, and define the Lyapunov function

$$\mathcal{F}_t = \zeta^{-1}(\tilde{\sigma}_t - 2Tk_+^{-1})[b_- + \zeta g_t]^2 + 2\zeta \mathrm{D}_{\mathrm{KL}}(\tilde{\sigma}_t \lambda).$$

Differentiating,

$$\begin{aligned}
\frac{\dot{\mathcal{F}}_t}{2(\tilde{\sigma} - 2Tk_+^{-1})} &= \zeta^{-1}(g)(b_- + \zeta g)^2 \\
&\quad + \zeta^{-1}(b_- + \zeta g)(\zeta(-ag - g^2 - \lambda^{-1} + \tilde{\sigma}^{-1}) + g\tilde{\sigma}^{-3/2}(\tilde{\sigma} - 2Tk_+^{-1})) \\
&\quad + \zeta g(\lambda^{-1} - \tilde{\sigma}^{-1}) \\
&= \zeta^{-1}(b_- + \zeta g) \\
&\quad \cdot \left[gb_- + \zeta g^2 + \zeta(-ag - g^2 - \lambda^{-1} + \tilde{\sigma}^{-1}) + g\tilde{\sigma}^{-3/2}(\tilde{\sigma} - 2Tk_+^{-1}) \right] \\
&\quad + \zeta gb_- b_+ \\
&= \zeta^{-1}(b_- + \zeta g) \\
&\quad \cdot \left[gb_- - a\zeta g - \zeta b_- b_+ + g\tilde{\sigma}^{-3/2}(\tilde{\sigma} - 2Tk_+^{-1}) \right] \\
&\quad + \zeta gb_- b_+ \\
&= \zeta^{-1}g(b_- + \zeta g) \cdot \left[b_- - a\zeta + \tilde{\sigma}^{-3/2}(\tilde{\sigma} - 2Tk_+^{-1}) \right] \\
&\quad - b_-^2 b_+.
\end{aligned}$$

Since

$$\mathrm{D}_{\mathrm{KL}}(\tilde{\sigma}, \lambda) \leq \tilde{\sigma}\sqrt{\lambda}b_-^2 b_+ - \frac{1}{2}\tilde{\sigma}b_-^2 = b_-^2(\sqrt{\lambda}b_+ - \frac{1}{2})\tilde{\sigma}, \tag{88}$$

we have that

$$\mathcal{F} \leq \zeta^{-1}(\tilde{\sigma}_t - 2Tk_+^{-1})[b_- + \zeta g_t]^2 + \zeta b_-^2(2\sqrt{\lambda}b_+ - 1)\tilde{\sigma},$$

and therefore we have the bound

$$-b_-^2 \leq \frac{-\mathcal{F} + \zeta^{-1}(\tilde{\sigma} - 2Tk_+^{-1})(b_- + \zeta g)^2}{\zeta(2\sqrt{\lambda}b_+ - 1)\tilde{\sigma}}. \tag{89}$$

Introducing a rate parameter $r > 0$, we again split the control term $-b_-^2 b_+ = -rb_-^2 b_+ - (1-r)b_-^2 b_+$, giving the derivative control

$$
\begin{aligned}
\frac{\dot{\mathcal{F}}_t}{2(\tilde{\sigma} - 2Tk_+^{-1})} &= \zeta^{-1}g(b_- + \zeta g) \cdot \left[b_- - a\zeta + \tilde{\sigma}^{-3/2}(\tilde{\sigma} - 2Tk_+^{-1}) \right] \\
&\quad - rb_-^2 b_+ - (1-r)b_-^2 b_+ \\
&\leq \zeta^{-1}g(b_- + \zeta g) \cdot \left[b_- - a\zeta + \tilde{\sigma}^{-3/2}(\tilde{\sigma} - 2Tk_+^{-1}) \right] \\
&\quad + rb_+ \frac{-\mathcal{F} + \zeta^{-1}(\tilde{\sigma} - 2Tk_+^{-1})(b_- + \zeta g)^2}{\zeta(2\sqrt{\lambda}b_+ - 1)\tilde{\sigma}} \\
&\quad - (1-r)b_-^2 b_+ \\
&= -\frac{rb_+}{\zeta(2\sqrt{\lambda}b_+ - 1)\tilde{\sigma}}\mathcal{F} + rb_+ \frac{\zeta^{-1}(\tilde{\sigma} - 2Tk_+^{-1})(b_- + \zeta g)^2}{\zeta(2\sqrt{\lambda}b_+ - 1)\tilde{\sigma}} \\
&\quad + \zeta^{-2}(b_- + \zeta g)^2[\lambda^{-1/2} - a\zeta - 2Tk_+^{-1}\tilde{\sigma}^{-3/2}] \\
&\quad - \zeta^{-2}b_-(b_- + \zeta g)[\lambda^{-1/2} - a\zeta - 2Tk_+^{-1}\tilde{\sigma}^{-3/2}] \\
&\quad - (1-r)b_-^2 b_+.
\end{aligned}
$$

In the last step, we expand b_-, and write $g(b_- + \zeta g) = \zeta^{-1}(b_- + \zeta g)^2 - \zeta^{-1}b_-(b_- + \zeta g)$. Define the auxiliary variable similarly to the previous section, as

$$
p = a\zeta - \lambda^{-1/2} + 2Tk_+^{-1}\tilde{\sigma}^{-3/2}.
$$

In order to have $p > 0$, we have the **necessary and sufficient condition:** $a\zeta \geq \lambda^{-1/2}$.

Consider the quadratic that is added to the $\mathcal{F}$ term. It can be written as $c_2(b_- + \zeta g)^2 + c_1(b_- + \zeta g) + c_0$, where

$$
c_2 = -\zeta^{-2}p + rb_+\zeta^{-2}\frac{\tilde{\sigma} - 2Tk_+^{-1}}{(2\sqrt{\lambda}b_+ - 1)\tilde{\sigma}}, \tag{90}
$$

$$
c_1 = \zeta^{-2}b_- p, \tag{91}
$$

$$
c_0 = -(1-r)b_-^2 b_+. \tag{92}
$$

We wish to show that the quadratic is always negative for all possible g.
Condition 1: $c_2 < 0$. This is equivalent to the inequality

$$
r \leq \frac{\tilde{\sigma}p(2\sqrt{\lambda}b_+ - 1)}{b_+(\tilde{\sigma} - 2Tk_+^{-1})}.
$$

Condition 2: $c_1^2 - 4c_2c_0 < 0$ when $r = 0$. This implies that the quadratic when $r = 0$ is strictly less than 0, which guarantees the existence of a positive rate by continuity. This condition can be written as

$$
\begin{aligned}
0 &> \zeta^{-4}p^2 - 4(-\zeta^{-2}p)(-b_+) \\
\Leftrightarrow 0 &> \zeta^{-2}p - 4b_+ \\
\Leftrightarrow 4\zeta^2 b_+ &> p.
\end{aligned}
$$

This holds for sufficiently large ζ since $p \sim a\zeta$ as $\zeta \to \infty$.

Let us take

$$
\zeta = a\lambda^{1/2}/2. \tag{93}
$$

Since $a \geq 2\lambda^{-1/2}$, the necessary condition $a\zeta \geq \lambda^{-1/2}$ holds and we have that $p > 0$. Therefore, the above equivalences hold. Moreover, we have that

$$
4\zeta^2 b_+ - p
$$
$$
\geq a^2 \lambda b_+ - \frac{1}{2}a^2\lambda^{1/2} + \lambda^{-1/2} - 2Tk_+^{-1}\tilde{\sigma}^{-3/2}
$$
$$
\geq \frac{1}{2}a^2\lambda^{1/2} + \lambda^{-1/2} + a^2\lambda\tilde{\sigma}^{-1/2} - \tilde{\sigma}^{-1/2} > 0.
$$

Therefore, the quadratic is always negative for $r = 0$. One obtains that the optimal rate is the smallest positive root of $c_1^2 - 4c_0c_2$, when written as a quadratic in r,

$$
\zeta^{-4}p^2 - 4(-\zeta^{-2}p + rb_+\zeta^{-2}\frac{\tilde{\sigma} - 2Tk_+^{-1}}{(2\sqrt{\lambda}b_+ - 1)\tilde{\sigma}})(-(1-r)b_+),
$$

and the rate becomes

$$
\dot{\mathcal{F}}_t \leq -\frac{2rb_+(\tilde{\sigma} - 2Tk_+^{-1})}{\zeta(2\sqrt{\lambda}b_+ - 1)\tilde{\sigma}}\mathcal{F}_t.
$$

C.1. **Proof of strengthened KL bound.** It remains to show the strengthened inequality (70). It states:

$$
\mathrm{D_{KL}}(\tilde{\sigma}, \lambda) \leq \tilde{\sigma}\sqrt{\lambda}b_-^2 b_+ - \frac{1}{2}\tilde{\sigma}b_-^2.
$$

where $b_\pm = \lambda^{-1/2} \pm \tilde{\sigma}^{-1/2}$. To use this, we use [41, Prop. 8], stated as follows. We first need the concept of convexity over a probability space.

Definition 3. *Let $\Omega \subset \mathbb{R}^n$ be some domain, and $\mathcal{P}(\Omega)$ the space of probability densities over Ω. For a density $\rho \in \mathcal{P}(\Omega)$, let $T_\rho\mathcal{P}(\Omega)$ and $T_\rho^*\mathcal{P}(\Omega)$ be the tangent and cotangent spaces at ρ respectively. A metric tensor is a (pointwise) invertible mapping $G(\rho) : T_\rho\mathcal{P}(\Omega) \to T_\rho^*\mathcal{P}(\Omega)$, which induces a metric inner product g_ρ on $T_\rho\mathcal{P}(\Omega)$ by*

$$
g_\rho(\sigma_1, \sigma_2) = \int \sigma_1 G(\rho)\sigma_2 \, \mathrm{d}x = \int \Phi_1 G(\rho)^{-1}\Phi_2 \, \mathrm{d}x, \quad \sigma_1, \sigma_2 \in T_\rho\mathcal{P}(\Omega),
$$

where Φ_i satisfies $\sigma_i = G(\rho)^{-1}\Phi_i$, $i = 1, 2$.

Let $E(\rho)$ be defined some functional over a probability space, such as the KL divergence. We say that $E(\rho)$ is α-strongly convex (with respect to a metric g) if for any $\rho \in \mathcal{P}(\Omega)$, for any $\sigma \in T_\rho\mathcal{P}(\Omega)$, we have

$$
g_\rho(\mathrm{Hess}\, E(\rho)\sigma, \sigma) \geq \beta g_\rho(\sigma, \sigma),
$$

where $\mathrm{Hess}\, E(\rho)$ is the Hessian operator w.r.t. the metric g_ρ.

The special case we consider is when the metric is given by the Wasserstein metric, with metric tensor

$$
G(\rho)^{-1}(\Phi) = -\nabla \cdot (\rho\nabla\Phi),
$$

and (tangent space) inner product (for $G(\rho)^{-1}\Phi_i \in T_\rho\mathcal{P}(\Omega)$,

$$
g_\rho(G(\rho)\Phi_1, G(\rho)\Phi_2) = \int \Phi_1 G(\rho)^{-1}\Phi_2 \, \mathrm{d}x.
$$

The enhanced KL property is given as follows.

Proposition 5. *Let $E(\rho)$ be some potential energy functional in Wasserstein space, and suppose that E satisfies $\mathrm{Hess}(\alpha)$ for some $\alpha \geq 0$. Let ρ^* be the minimizer of E. Further let T_t be the optimal transport map from ρ_t to ρ^*. Then,*

$$
E(\rho^*) \geq E(\rho) + \int \left\langle T_t(x) - x, \nabla\frac{\delta E}{\delta\rho_t} \right\rangle \rho_t \, \mathrm{d}x + \frac{\alpha}{2}\int \|T_t(x) - x\|^2 \rho_t \, \mathrm{d}x. \tag{94}
$$

We have the following corollary for the special case where E is KL divergence and ρ_t, ρ^* are both Gaussians. Let $\rho^* = \mathcal{N}(0, \Lambda)$. We have the following representation of the KL divergence and transport map:

Lemma 3. *The KL distance between two zero-mean Gaussians in $\mathbb{R}^d$ is*

$$D_{\mathrm{KL}}(\Sigma_1, \Sigma_2) = \frac{1}{2}\left(\log \det \Sigma_2 \Sigma_1^{-1} - d + \mathrm{Tr}\left(\Sigma_2^{-1}\Sigma_1\right)\right). \tag{95}$$

Moreover, the optimal transport map T from $\mathcal{N}(0, \Sigma_1)$ to $\mathcal{N}(0, \Sigma_2)$ is linear,

$$T(x) = \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}x. \tag{96}$$

Substituting this into Proposition 5 yields the desired inequality. Let $\rho^* = \mathcal{N}(0, \Lambda)$ be the desired target distribution $\rho^* \propto \exp(-V)$, where $V(x) = x^\top \Lambda^{-1} x/2$.

Corollary 2. *Let $\rho^* = \mathcal{N}(0, \Lambda)$. Taking $E(\rho) = D_{\mathrm{KL}}(\rho\|\rho^*)$, the minimizer of E is ρ^*. Moreover, if $\rho = \mathcal{N}(0, \Sigma)$, then*

$$E(\rho) \le -\mathrm{Tr}\left((\Sigma^{1/2}\Lambda\Sigma^{1/2})^{1/2}(\Sigma^{-1/2}\Lambda^{-1}\Sigma^{-1/2} - I)\Sigma\right)$$
$$-\frac{\alpha}{2}\mathrm{Tr}\left(\left(\Sigma^{-1/2}(\Sigma^{1/2}\Lambda\Sigma^{1/2})^{1/2}\Sigma^{-1/2} - I\right)^2 \Sigma\right). \tag{97}$$

In the case where the covariance matrices Σ and Λ commute, one has the strengthened bound, where $B_\pm = \Lambda^{-1/2} \pm \Sigma^{-1/2}$,

$$E(\rho) \le \mathrm{Tr}\left(\Sigma\Lambda^{1/2}B_-^2 B_+ - \Sigma B_-^2/2\right).$$

Proof. The following two statements hold.

- $E(\rho)$ satisfies $\mathrm{Hess}(\alpha)$ where $\alpha = \lambda_{\max}^{-1}$. This follows from the Bakry–Emery criterion since $V(x)$ is $\lambda_{\max}^{-1}$-strongly convex.
- The optimal transport map is linear by Lemma 3.

We can now substitute into (94). For a quadratic form $x^\top A x$, we have that

$$\int x^\top A x \, d\mathcal{N}(0, \Sigma)(x) = \mathrm{Tr}(A\Sigma). \tag{98}$$

Further note that for the KL divergence,

$$\frac{\delta E}{\delta \rho} = \frac{1}{2}x^\top(\Lambda^{-1} - \Sigma^{-1})x + 1, \quad \text{therefore} \quad \nabla\frac{\delta E}{\delta \rho}(x) = (\Lambda^{-1} - \Sigma^{-1})x.$$

Since $E(\rho^*) = 0$, we have that

$$E(\rho) \le -\int \left\langle T(x) - x, \nabla\frac{\delta E}{\delta \rho}\right\rangle \rho \, dx - \frac{\alpha}{2}\int \|T(x) - x\|^2 \rho \, dx$$

$$= -\int x^\top \left(\Sigma^{-1/2}(\Sigma^{1/2}\Lambda\Sigma^{1/2})^{1/2}\Sigma^{-1/2}(\Lambda^{-1} - \Sigma^{-1})\right) x \, d\rho(x)$$

$$-\frac{\alpha}{2}\int x^\top \left(\Sigma^{-1/2}(\Sigma^{1/2}\Lambda\Sigma^{1/2})^{1/2}\Sigma^{-1/2} - I\right)^2 x \, d\rho(x)$$

$$= -\mathrm{Tr}\left((\Sigma^{1/2}\Lambda\Sigma^{1/2})^{1/2}(\Sigma^{-1/2}\Lambda^{-1}\Sigma^{-1/2} - I)\Sigma\right)$$

$$-\frac{\alpha}{2}\mathrm{Tr}\left(\left(\Sigma^{-1/2}(\Sigma^{1/2}\Lambda\Sigma^{1/2})^{1/2}\Sigma^{-1/2} - I\right)^2 \Sigma\right).$$

In the case where Λ and Σ commute, one may work over an eigenbasis and sharpen the bound Equation (94) to each dimension. In particular, the transport map simplifies to $T(x) = \Lambda^{1/2}\Sigma^{-1/2}x$, and the bound then becomes

$$E(\rho) \le -\mathrm{Tr}\left((\Lambda^{1/2}\Sigma^{-1/2} - I)(\Lambda^{-1} - \Sigma^{-1})\Sigma\right)$$

$$-\frac{1}{2}\mathrm{Tr}\left(\Lambda^{-1}\left(\Lambda^{1/2}\Sigma^{-1/2} - I\right)^2 \Sigma\right)$$

$$= \mathrm{Tr}\left(\Lambda^{1/2}(\Lambda^{-1/2} - \Sigma^{-1/2})(\Lambda^{-1} - \Sigma^{-1})\right) - \frac{1}{2}\mathrm{Tr}\left(\Sigma(\Lambda^{-1/2} - \Sigma^{-1/2})^2\right)$$

$$= \mathrm{Tr}\left(\Sigma\Lambda^{1/2}B_-^2 B_+ - \Sigma B_-^2/2\right),$$

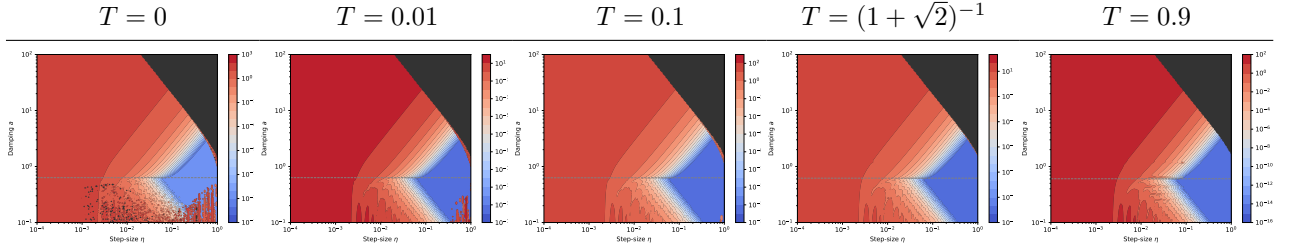| $T = 0$ | $T = 0.01$ | $T = 0.1$ | $T = (1 + \sqrt{2})^{-1}$ | $T = 0.9$ |

FIGURE 7. Contour plots of the covariance error of discrete-time ARWP (18), for target distribution $\Lambda = \mathcal{N}(0, \operatorname{diag}(1, 20))$, with initialization $\mathcal{N}(0, 40I)$, at 2000 iterations. We observe that for lower levels of T, numerical instability dominates in the highly underdamped setting $a < 20^{-1/2}$, due to the covariance getting close to zero.

as desired. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

APPENDIX D. DISCRETE TIME ARWP COVARIANCE UPDATE

Let $V(x) = x^\top \Lambda^{-1} x / 2$. Recall the original accelerated regularized kinetic system (18), updated as

$$\begin{cases} P_{k+1} = (1 - a\eta)P_k - \eta(\nabla V(X_k) + \nabla \log \operatorname{WProx} \rho_k(X_k)), \\ X_{k+1} = X_k + \eta P_{k+1}. \end{cases} \tag{99}$$

We recall the essential expressions relating the covariance of a Gaussian to its regularized Wasserstein proximal

$$K_\pm = 1 \pm T\lambda^{-1}, \quad B_\pm = \lambda^{-1/2} \pm \tilde{\sigma}^{-1/2}. \tag{100}$$

$$\tilde{\Sigma}_t = 2TK_+^{-1} + K_+^{-1}\Sigma_t K_+^{-1}, \quad \Sigma_t = K_+(\tilde{\Sigma}_t - 2TK_+^{-1})K_+. \tag{101}$$

Let us write $P_{k+1} = G_k X_k$, where G_k is some matrix. Then,

$$X_{k+1} = X_k + \eta P_{k+1} = (I + \eta G_k)X_k, \tag{102}$$

and the update of the momentum term

$$\begin{aligned} G_k X_k = P_{k+1} &= (1 - a\eta)P_k - \eta(\Lambda^{-1} - \tilde{\Sigma}_k^{-1})X_k \\ &= (1 - a\eta)G_{k-1}X_{k-1} - \eta(\Lambda^{-1} - \tilde{\Sigma}_k^{-1})X_k \\ &= (1 - a\eta)G_{k-1}(I + \eta G_{k-1})^{-1}X_k - \eta(\Lambda^{-1} - \tilde{\Sigma}_k)X_k. \end{aligned}$$

Therefore, the regularized WProx kinetic equation in covariance form becomes

$$X_{k+1} = (I + \eta G_k)X_k \Rightarrow \Sigma_{k+1} = (I + \eta G_k)\Sigma_k(I + \eta G_k)^\top, \tag{103}$$

$$G_k = (1 - a\eta)G_{k-1}(1 + \eta G_{k-1})^{-1} - \eta(\Lambda^{-1} - \tilde{\Sigma}_k^{-1}). \tag{104}$$

APPENDIX E. ABLATION ON REGULARIZATION PARAMETERS: GAUSSIAN CASE

We consider a 2D Gaussian target distribution $\Lambda = \operatorname{diag}(1, 20)$. Figure 7 ablates against the Wasserstein proximal regularization parameter T, which has to be fixed in $T \in [0, 1)$. We observe that the optimal damping is given by (36), which is consistent with the linearization analysis in Section 4.3. Moreover, as T decreases, the update becomes (numerically) unstable for small damping a. This is due to the update step (104): if T is small, then the regularized Wasserstein proximal variance matrix $\tilde{\Sigma}_k$ may also be very small, leading to a large update in the momentum matrix G. This makes the ODE system stiff, and can cause blowup if the step-size is not chosen to be sufficiently small.

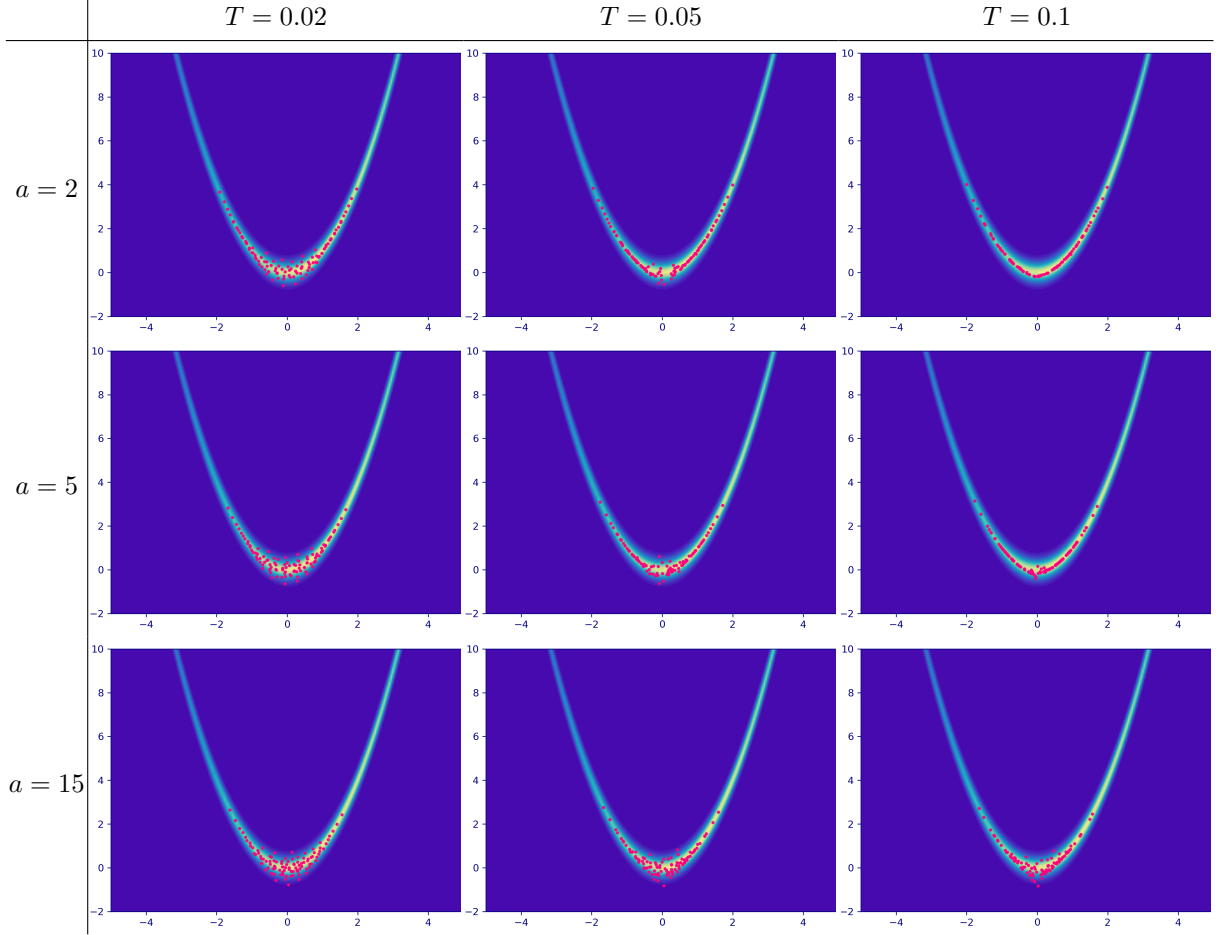FIGURE 8. Ablation of ARWP-Heavy-ball across some different choices of $T \in \{0.02, 0.05, 0.1\}$ and $a \in \{2, 5, 15\}$. All evaluated at iteration 200 and step-size $\eta = 0.02$. As T increases, the particles concentrate more around the parabola $y = x^2$.

APPENDIX F. ADDITIONAL PLOTS FOR ROSENBROCK DISTRIBUTION

Figure 8 plots the particle positions at iteration 200 for the Rosenbrock distribution. The particle count is $N = 100$, step-size $\eta = 0.02$, and the parameters are varied as $T \in [0.02, 0.05, 0.1]$ and $a \in [2, 5, 15]$. We observe that as T increases, the particles stay closer to the parabola $y = x^2$. This is consistent with the observation in [37] that the T parameter "shrinks" the local variance by T. The evolution is very similar for the different damping parameters in this case, possibly indicating overdamping.

To contrast this sensitivity, we compare with ILA, which has different behavior as the damping parameter changes as seen in Figure 9. For the less damped case, corresponding to a small Lipschitz constant estimate, the particles are able to explore the tails. However, in the more damped case, the particles do not explore the tails. KLMC does not explore the tails at all, as seen in Figure 10.

ARWP has an additional advantage over BRWP in its stability with respect to step-size. BRWP is only able to diffuse up to a certain height of the parabola before some instabilities occur, and the outermost particles begin to oscillate perpendicularly to the parabola.

APPENDIX G. HYPERPARAMETERS FOR EXPERIMENTS

We detail the hyperparameters used in the plots for the ill-conditioned Gaussian Section 5.2, Gaussian mixture Section 5.4, and Bayesian neural network examples Section 5.5. These are given
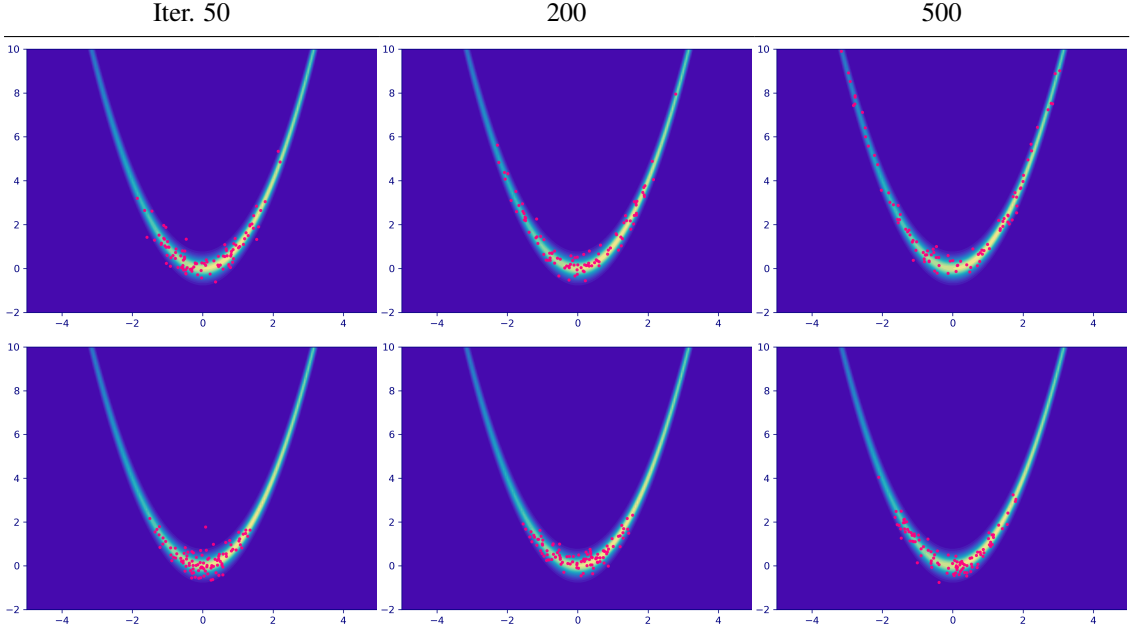
FIGURE 9. Top row: underdamped ILA ($\eta = 0.05$, $L = 2$; bottom row: overdamped ILA ($\eta = 0.05$, $L = 10$). Evaluated at iterations 50, 200 and 500. Overdamped ILA does not explore the tails.
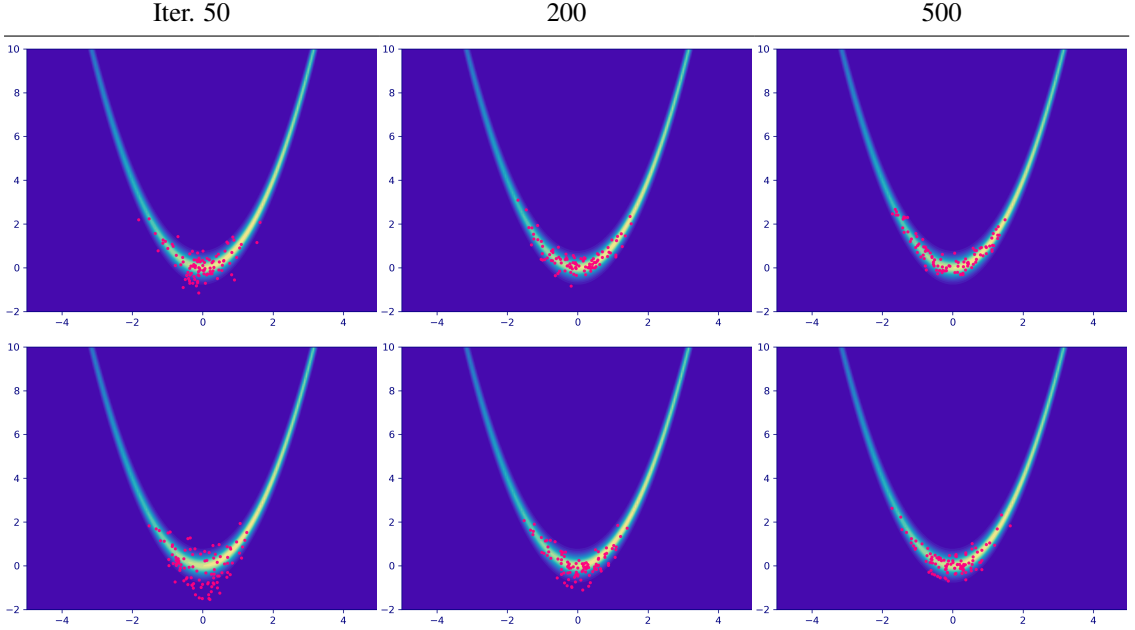

FIGURE 10. Top row: underdamped KLMC ($\eta = 0.01$, $a = 5$; bottom row: overdamped KLMC ($\eta = 0.01$, $a = 20$). Evaluated at iterations 50, 200 and 500. There is no exploration.

respectively in Tables 2 to 4. The parameters for Rosenbrock experiments are given in the previous section Section F.
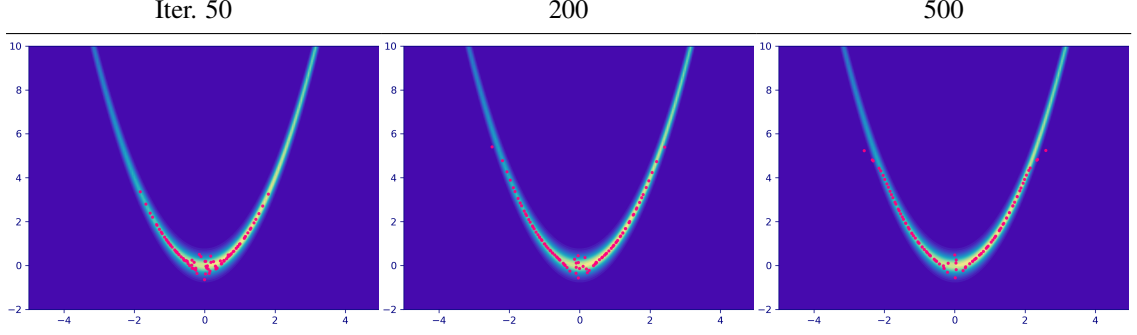
Iter. 50   200   500

FIGURE 11. Rosenbrock distribution with BRWP ($\eta = 0.02$, $T = 0.05$). Evaluated at iterations 50, 200 and 500. The convergence is fast but does not continue to explore the tails, as the outermost particles oscillate back and forth about the parabolic valley.

Method	Parameters
ULA	$\eta = 0.01$
MALA	$\eta = 0.05$
BRWP	$\eta = 0.2$, $T = 0.05$
ARWP-Nesterov	$\eta = 0.3$, $T = 0.05$
ARWP-Heavy-ball	$\eta = 0.3$, $T = 0.05$, $a = 1$
ILA	$\eta = 0.2$, $L = 5$
KLMC	$\eta = 0.2$, $a = 5$

TABLE 2. Ill-conditioned Gaussian

Method	Parameters
ULA	$\eta = 0.1$
MALA	$\eta = 0.3$
BRWP	$\eta = 0.5$, $T = 0.1$
ARWP-Nesterov	$\eta = 0.6$, $T = 0.1$
ARWP-Heavy-ball	$\eta = 0.6$, $T = 0.1$, $a = 1$
ILA	$\eta = 0.3$, $L = 1$
KLMC	$\eta = 0.2$, $a = 2$

TABLE 3. GMM

Dataset	Parameters
Boston	$\eta = 0.02$, $T = 0.01$
Combined	$\eta = 0.02$, $T = 0.01$
Concrete	$\eta = 0.15$, $T = 0.01$
Kin8nm	$\eta = 0.02$, $T = 0.01$
Wine	$\eta = 0.1$, $T = 0.005$

TABLE 4. BNN

APPENDIX H. DISCRETIZED KINETIC LANGEVIN ALGORITHMS

We detail the ILA and KLMC algorithms here, which were used as benchmarks for the low-dimensional examples.

H.1. **Inertial Langevin Algorithm.** The inertial Langevin algorithm (ILA) [16, Alg. 1.1] takes the following form. A particle is initialized with position X_0 and velocity $P_0 = 0$. The hyperparameters are a friction coefficient $\varepsilon \in [4/3, 7/4]$, a step-size Δt, and the Lipschitz constant of ∇V denoted as L. The ILA update then defines two new parameters:

$$\beta := 1 - \varepsilon \Delta t, \quad \tau := \Delta t^2 / L,$$

and updates the particle position as

$$X_{k+1} = X_k,$$
$$P_{k+1} = X_{k+1} - X_k.$$

For simplicity, we take the friction $\varepsilon = 1.5$, and use the Lipschitz constant as a tunable damping coefficient.

H.2. **Kinetic Langevin Monte Carlo.** The KLMC algorithm is given in [10, 12]. The algorithm is given by an exponential integrator, detailed as follows. For a damping parameter a and a step-size η, the KLMC update is given by [12, Sec. 3]:

$$\begin{bmatrix} X_{k+1} \\ P_{k+1} \end{bmatrix} = \begin{bmatrix} X_k + \psi_1(\eta) P_k - \psi_2(\eta) \nabla V(X_k) \\ \psi_0(\eta) P_k - \psi_1(\eta) \nabla V(X_k) \end{bmatrix} + \sqrt{2a} \begin{bmatrix} \xi_{k+1} \\ \xi'_{k+1} \end{bmatrix}, \quad (105)$$

where $\psi_0(t) = e^{-at}$, $\psi_{k+1}(t) = \int_0^t \psi_k(s)\,\mathrm{d}s$, and $\begin{bmatrix} \xi_{k+1} \\ \xi'_{k+1} \end{bmatrix}$ are $2d$-dimensional centered Gaussian vectors with covariance matrix given by

$$\begin{bmatrix} \xi_{k+1} \\ \xi'_{k+1} \end{bmatrix} \sim \mathcal{N}(0, \mathbf{C}), \quad \mathbf{C} = \int_0^\eta \begin{bmatrix} \psi_0(t) \\ \psi_1(t) \end{bmatrix} \begin{bmatrix} \psi_0(t) & \psi_1(t) \end{bmatrix} \mathrm{d}t.$$

The expressions are given in closed form as follows:

$$\psi_0(t) = e^{-at},$$
$$\psi_1(t) = a^{-1}(1 - e^{-at}),$$
$$\psi_2(t) = a^{-2}(at + e^{-at} - 1),$$
$$\mathbf{C} = \begin{bmatrix} (2a)^{-1}(1 - e^{-2a\eta}) & (2a^2)^{-1}(1 - e^{-a\eta})^2 \\ (2a^2)^{-1}(1 - e^{-a\eta})^2 & (2a^3)^{-1}(2a\eta - e^{-2a\eta} + 4e^{-a\eta} - 3) \end{bmatrix}.$$