Preconditioned Regularized Wasserstein Proximal Sampling Methods

Hong Ye Tan*, Stanley Osher†, and Wuchen Li†

Abstract. We consider sampling from a Gibbs distribution by evolving finitely many particles. We propose a preconditioned version of a recently proposed noise-free sampling method, governed by approximating the score function with the numerically tractable score of a regularized Wasserstein proximal operator. This is derived by a Cole–Hopf transformation on coupled anisotropic heat equations, yielding a kernel formulation for the preconditioned regularized Wasserstein proximal. The diffusion component of the proposed method is also interpreted as a modified self-attention block, as in transformer architectures. For quadratic potentials, we provide a discrete-time non-asymptotic convergence analysis and explicitly characterize the bias, which is dependent on regularization and independent of step-size. Experiments demonstrate acceleration and particle-level stability on various log-concave and non-log-concave toy examples to Bayesian total-variation regularized image deconvolution, and competitive/better performance on non-convex Bayesian neural network training when utilizing variable preconditioning matrices.

Key words. Sampling, Cole-Hopf, regularized Wasserstein proximal, preconditioning, kernel formula, score approximation, MCMC

MSC codes. 65C05, 62G07

1. Introduction.

We are interested in sampling from a Gibbs distribution

$$\pi(x) \propto \exp(-\beta V(x)), \tag{1.1}$$

where $V : \mathbb{R}^d \to \mathbb{R}$ is some $\mathcal{C}^1$ potential, and $\beta > 0$ is a constant. Such problems arise frequently in data science, such as uncertainty quantification in imaging [28], Bayesian statistical inference [12], inverse problems [42], and computational physics [25]. Methods for sampling have also been connected to recent machine learning methods, such as simulating stochastic differential equations (SDEs) for diffusion models [40], unsupervised learning [44], and Bayesian neural networks [31].

Current popular sampling methods to solve this are Monte Carlo Markov chain (MCMC) methods, which bypass the usual difficulty arising from the intractability of the normalizing constant $\int_{\mathbb{R}^d} \exp(-\beta V(x))\,\mathrm{d}x$. There is a wide variety of Monte Carlo methods each with different convergence rates and conditions on the potential V. Notable examples include the unadjusted Langevin algorithm and the Metropolis-adjusted Langevin algorithm [36, 10], with recent algorithms considering proximal operators and Moreau–Yosida envelopes [11], subgradients [16], and annealed Moreau regularization [15]. The theoretical core of MCMC is to construct a Markov chain with some invariant distribution $\hat{\pi}$ that approximates the target distribution π.

To construct such a Markov chain, a crucial theoretical component is exponential or geometric ergodicity of the chain. In typical Langevin methods, this is given by the Wiener diffusion term, which adds random noise at every step of the chain [35]. These chains can

*Department of Mathematics, University of California, Los Angeles, 90095 (hyt35,sjo@math.ucla.edu).
†Department of Mathematics, University of South Carolina, Columbia, SC 29208 (wuchen@mailbox.sc.edu)

be interpreted as particular discretizations of SDEs, such as Euler–Maruyama in the case of the unadjusted Langevin algorithm. These SDEs correspond to Fokker–Planck equations [34], which are ODEs in density space, which in turn correspond to score-based particle evolutions based on Liouville equations [26]. Here, the score function is defined by the gradient of the log density.

Deviating from MCMC methods, score-based methods which crucially require knowledge of the score of the distribution at any given time-step. The difficulty manifests as at each time step, only an empirical measure is known, from which the score has to be approximated. Various methods of score approximation include kernel density estimation [6, 49], adaptive kernels [50, 47, 5], and approximation using neural networks [4, 8, 32]. However, many such methods suffer from mode collapse and parameter/initialization sensitivity [41, 30, 14].

Recently, [45] proposed a noise-free sampling based on the regularized Wasserstein proximal operator (RWPO), called the backwards regularized Wasserstein proximal (BRWP) method. In this case, the score is approximated using the score of the RWPO, which is shown to correspond to a semi-implicit discretization of a Fokker–Planck equation. Similarly to the unadjusted Langevin algorithm (ULA), and unlike Langevin algorithms involving Metropolis steps (such as the Metropolis-adjusted Langevin algorithm MALA), BRWP can be shown to preserve the Gaussianity of a distribution at each iteration for quadratic target potentials. In particular, in continuous time, the mixing time dependency on dimension d in the Gaussian case scales only as $\mathcal{O}(\log(d))$, whereas ULA scales as $\mathcal{O}(d^3)$ and MALA as $\mathcal{O}(d^2)$. For low dimensional distributions, the particles evolved through BRWP were found to be structured, with the outermost particles appearing to lie on level-set contours.

The BRWP method has also been extended with tensor-train techniques [19], as well as to the non-smooth Bayesian LASSO problem by using L^1-splittings based on the Laplace approximation [18]. On the theoretical side, [45] originally demonstrate linear convergence in total variation for quadratic potentials in the continuous case. [17] further extend this to strongly log-concave distributions and smooth initial distributions, showing that the BRWP iterations with equal regularization parameter and time-step indeed form a first-order method for the Fokker–Planck equation.

In this work, we aim to accelerate convergence of BRWP by utilizing preconditioners. The contributions of this work are as follows:

1. In Section 3, we derive the coupled PDE system that we wish to subsequently discretize. Utilizing similar techniques to [29], we demonstrate that the Laplacian regularizations in the Benamou–Brenier formulation of the Wasserstein proximal may be suitably replaced with second-order terms involving the preconditioning matrix. This is derived using a Cole–Hopf transform on a pair of forward-backward anisotropic heat equations.
2. In Section 4 we propose the preconditioned BRWP (PBRWP) method in Algorithm 1. This is given by taking a backwards Euler discretization on the Fokker–Planck component of the PDE derived in the previous section. This yields a semi-implicit discretization of the Liouville equation. We provide expressions for the score approximation in the case of finitely many particles as an interacting particle system. Subsection 4.1 interprets the interactions as a modified self-attention block, which implicitly allows for efficient parallel computation using common GPU libraries.
3. In the case of Gaussian initial and target measures (corresponding to the infinite-

particle limit), Section 5 provides a non-asymptotic analysis of the convergence rate in the discrete step-size regime. This characterizes the convergence rates in terms of the regularization parameter and the preconditioning matrix. We additionally provide various properties such as maximum allowable regularization parameters, Wasserstein contraction, a mean-variance contraction-diffusion inequality, and a maximum norm bound based on the number of particles.

4. In Section 6, the proposed PBRWP method is compared on a variety of toy examples, a Bayesian imaging problem, and Bayesian neural network training. We additionally introduce some heuristic modifications to improve the performance of PBRWP in high dimension scenarios, corresponding to scaling laws in attention-based models. We observe in low dimensions that the proposed method converges faster and stably in KL divergence and maintains the structured particle phenomenon of BRWP. We find rapid convergence to the MAP estimator in while maintaining deviations between particles for Bayesian image deconvolution, and convergence to better trained networks compared to other score-based MCMC methods.

2. Background and Notation. This section covers some basic definitions, including Wasserstein spaces, the Benamou–Brenier PDE formulation, and the non-preconditioned regularized Wasserstein proximal.

Throughout, $M \in \mathbb{R}^{d\times d}$ is a symmetric positive definite matrix $M \in \mathrm{Sym}_{++}(\mathbb{R}^d)$. We define the scaled norms and inner product on $\mathbb{R}^d$ as

$$\langle u, v\rangle_M = u^\top M^{-1} v, \quad \|\cdot\|_M^2 = \langle \cdot, \cdot\rangle_M. \tag{2.1}$$

Definition 2.1 ([39]). *Let $\mathcal{P}_2(\mathbb{R}^d)$ be the 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*

$$\mathcal{W}_2(\mu,\nu) = \inf_{\pi\in\Gamma(\mu,\nu)} \int \|x-y\|^2 \mathrm{d}\pi(x,y), \tag{2.2}$$

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

$$\int_{\mathbb{R}^d} \pi(x,y)\,\mathrm{d}y = \mu(x), \int_{\mathbb{R}^d} \pi(x,y)\,\mathrm{d}x = \nu(y). \tag{2.3}$$

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

$$\mathrm{WProx}_{T,V}(\rho_0) \coloneqq \operatorname*{arg\,min}_{q\in\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} V(x)q(x)\,\mathrm{d}x + \frac{\mathcal{W}(\rho_0,q)^2}{2T}. \tag{2.4}$$

The Wasserstein proximal can be equivalently reformulated as a coupled PDE system, sometimes known as the Benamou–Brenier or dynamical formulation [2]. Using this, [29]

defines the *regularized Wasserstein proximal* by adding a diffusive term to the continuity equation of the Benamou–Brenier formulation. The resulting mean-field control problem is

$$\inf_{\rho,v,q} \int_0^T \int_{\mathbb{R}^d} \frac{1}{2}\|v(t,x)\|^2 \rho(t,x)\,\mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}^d} V(x)q(x)\,\mathrm{d}x\,, \tag{2.5a}$$

$$\partial_t \rho(t,x) + \nabla\cdot(\rho(t,x)v(t,x)) = \beta^{-1}\Delta\rho(t,x), \quad \rho(0,x)=\rho_0(x),\ \rho(T,x)=q(x). \tag{2.5b}$$

From this problem, the regularized Wasserstein proximal operator (RWPO) $\mathrm{WProx}_{T,V}$ is defined as the solution density function ρ of (2.5). Taking the regularization parameter $\beta = 0$ recovers the Wasserstein proximal (2.4) through the Benamou–Brenier formulation of the Wasserstein distance. The RWPO admits the following coupled PDE formulation, consisting of a forward-time Fokker-Planck equation in ρ, and a backward-time viscous Hamilton-Jacobi equation in a dual variable Φ,

$$\begin{cases} \partial_t \rho(t,x) + \nabla_x\cdot(\rho(t,x)\nabla_x\Phi(t,x)) = \beta^{-1}\Delta_x\rho(t,x), & \text{(2.6a)}\\ \partial_t\Phi(t,x) + \dfrac{1}{2}\|\nabla_x\Phi(t,x)\|^2 = -\beta^{-1}\Delta_x\Phi(t,x), & \text{(2.6b)}\\ \rho(0,x)=\rho_0(x), \quad \Phi(T,x) = -V(x). & \text{(2.6c)} \end{cases}$$

Consider the Cole–Hopf transformation

$$\begin{cases} \eta(t,x) = e^{\beta\Phi(t,x)/2} \\ \hat\eta(t,x) = \rho(t,x)e^{-\beta\Phi(t,x)/2} \end{cases} \Leftrightarrow \begin{cases} \Phi(t,x) = 2\beta^{-1}\log\eta(t,x) \\ \rho(t,x) = \eta(t,x)\hat\eta(t,x). \end{cases}$$

This transforms the coupled PDEs (2.6) to the coupled forward-backward heat equations

$$\begin{cases} \partial_t\hat\eta(t,x) = \beta^{-1}\Delta\hat\eta(t,x), \\ \partial_t\eta(t,x) = -\beta^{-1}\Delta\eta(t,x), \\ \eta(0,x)\hat\eta(0,x) = \rho_0(x), \quad \eta(T,x) = e^{\beta\Phi(T,x)/2} = e^{-\beta V(x)/2}, \end{cases} \tag{2.7}$$

which may be solved using the heat kernel [29]

$$G_t(x,y) = \frac{1}{(4\pi\beta^{-1}t)^{d/2}} e^{-\beta\frac{\|x-y\|^2}{4t}}, \quad \eta(t,x) = G_{T-t} * \eta_T = G_{T-t} * e^{-\beta V/2}.$$

Translating back, the terminal solution $t = T$ to the coupled PDEs (2.6) is given by

$$\rho(T,x) = \int_{\mathbb{R}^d} K(x,y)\rho_0(y)\,\mathrm{d}y\,, \tag{2.8}$$

$$K(x,y) = \frac{\exp\left(-\frac{\beta}{2}(V(x) + \frac{\|x-y\|^2}{2T})\right)}{\int_{\mathbb{R}^d}\exp\left(-\frac{\beta}{2}(V(z) + \frac{\|z-y\|^2}{2T})\right)\mathrm{d}z}. \tag{2.9}$$

3. Preconditioned Regularized Wasserstein Proximal. We have seen that the regularized Wasserstein proximal is a particular Laplacian regularization of the Benamou–Brenier formulation, corresponding to coupled heat equations and computable using a kernel formula [29]. In

this section, we define a geometry-aware elliptic regularization to the Benamou–Brenier formulation, corresponding to coupled anisotropic heat equations under a Cole–Hopf transformation. This manifests as computation with a scaled heat kernel, allowing for more generality in the proximal operator.

From here onwards, let $M \in \mathbb{R}^{d\times d}$ be a symmetric positive definite matrix $M \in \mathrm{Sym}_{++}(\mathbb{R}^d)$. Define the anisotropic heat kernel

$$G_{t,M}(x,y) \coloneqq \frac{1}{(4\pi\beta^{-1}t)^{d/2}|M|^{1/2}} e^{-\beta\frac{(x-y)^\top M^{-1}(x-y)}{4t}}. \tag{3.1}$$

This is a scaled version of the standard heat kernel, and it is also a Green's function for an anisotropic heat equation.

Proposition 3.1. *The anisotropic kernel $G_{t,M}$ is a Green's function for the following PDE:*

$$\begin{cases} \partial_t u - \beta^{-1}\nabla\cdot(M\nabla u) = 0, \\ u(0,x) = \delta(y). \end{cases}$$

Proof. This is from a direct computation. The boundary condition arises from a change of variables with the standard heat kernel. The derivation can be found in A.1. ∎

We now define the preconditioned regularized Wasserstein proximal operator as the terminal time solution to a particular set of PDEs. Then, we show that it is equivalently a convolution involving another inhomogeneous kernel, with the kernel itself defined as a convolution with the anisotropic heat kernel.

Definition 3.2. *For a symmetric positive definite matrix M, define the* preconditioned regularized Wasserstein proximal operator *(PRWPO)* $\mathrm{WProx}^M_{T,V} : \rho_0 \mapsto \rho_T$ *to be the terminal density of the following PDE system*

$$\begin{cases} \partial_t\rho(t,x) + \nabla\cdot(\rho(t,x)\nabla\Phi(t,M^{-1}x)) = \beta^{-1}\nabla\cdot(M\nabla\rho)(t,x), \\ \partial_t\Phi(t,M^{-1}x) + \frac{1}{2}\|\nabla\Phi(t,M^{-1}x)\|^2_M = -\beta^{-1}\operatorname{Tr}\big(M^{-1}(\nabla^2\Phi)(t,M^{-1}x)\big), \\ \rho(0,x) = \rho_0(x), \quad \Phi(T,M^{-1}x) = -V(x). \end{cases} \tag{3.2}$$

Equivalently, the density of $\mathrm{WProx}^M_{T,V}\,\rho_0$ *is given by*

$$\rho(T,x) = \int_{\mathbb{R}^d} K(x,y,\beta,M,T,V)\rho(0,y)\,\mathrm{d}y\,, \tag{3.3}$$

where the normalized kernel is given by

$$K(x,y,\beta,M,T,V) \coloneqq \frac{e^{-\frac{\beta}{2}(V(x)+\frac{\|x-y\|^2_M}{2T})}}{\int_{\mathbb{R}^d} e^{-\frac{\beta}{2}(V(z)+\frac{\|z-y\|^2_M}{2T})}\,\mathrm{d}z}.$$

Compared to (2.9), the preconditioned kernel replaces the Euclidean norm $\|x-y\|^2$ with the scaled norm $\|x-y\|^2_M$, which is by construction. We show that the terminal density and the kernel formula are equivalent in the following section.

3.1. Proof of equivalence. Like in [29], the kernel formula (3.3) arises from a Cole–Hopf transformation to the underlying coupled PDE system. We consider the following coupled forward-backward anisotropic heat equations, which admit the kernel $G_{t,M}$,

$$\begin{cases} \partial_t \hat{\eta}(t,x) = \beta^{-1} \nabla \cdot (M \nabla \hat{\eta}(t,x)), \\ \partial_t \eta(t,x) = -\beta^{-1} \nabla \cdot (M \nabla \eta(t,x)), \\ \eta(0,x)\hat{\eta}(0,x) = \rho_0(x), \quad \eta(T,x) = e^{\beta \Phi(T,x)/2} = e^{-\beta V(x)/2}. \end{cases} \tag{3.4}$$

We now transform this back into a set of coupled PDEs with the following Cole–Hopf transform,

$$\begin{cases} \eta(t,x) = e^{\beta \Phi(t, M^{-1}x)/2} \\ \hat{\eta}(t,x) = \rho(t,x) e^{-\beta \Phi(t, M^{-1}x)/2} \end{cases} \Leftrightarrow \begin{cases} \Phi(t,x) = 2\beta^{-1} \log \eta(t, Mx) \\ \rho(t,x) = \eta(t,x)\hat{\eta}(t,x) \end{cases}. \tag{3.5}$$

The boundary conditions transform to yield the desired $\rho(0,x) = \rho_0(x)$, $\Phi(T, M^{-1}x) = -V(x)$. We first compute the second (backwards) heat equation, dealing solely in the dual variable Φ:

$$\begin{aligned} \partial_t \eta(t,x) &= \frac{\beta}{2} \partial_t \Phi(t, M^{-1}x) e^{\beta \Phi(t, M^{-1}x)/2}, \\ \nabla \eta(t,x) &= \frac{\beta}{2} M^{-1} \nabla \Phi(t, M^{-1}x) e^{\beta \Phi(t, M^{-1}x)/2}. \end{aligned}$$

The RHS of the second heat equation yields

$$\begin{aligned} &\nabla \cdot (M \nabla \eta(t,x)) \\ &= \nabla \cdot \left(\frac{\beta}{2} \nabla \Phi(t, M^{-1}x) e^{\beta \Phi(t, M^{-1}x)/2} \right) \\ &= \left(\frac{\beta}{2} \operatorname{Tr}\big(M^{-1} \nabla^2 \Phi(t, M^{-1}x)\big) + \frac{\beta^2}{4} \nabla \Phi(t, M^{-1}x)^\top M^{-1} \nabla \Phi(t, M^{-1}x) \right) e^{\beta \Phi(t, M^{-1}x)/2}. \end{aligned}$$

Therefore the backwards heat equation evolving from time $t = T$ to $t = 0$

$$\partial_t \eta(t,x) = -\beta^{-1} \nabla \cdot (M \nabla \eta(t,x))$$

becomes a PDE in the dual variable

$$\partial_t \Phi(t, M^{-1}x) + \frac{1}{2} \|\nabla \Phi(t, M^{-1}x)\|_M^2 = -\beta^{-1} \operatorname{Tr}\big(M^{-1} (\nabla^2 \Phi)(t, M^{-1}x)\big). \tag{3.6}$$

This is the desired Hamilton–Jacobi equation in (3.2). Now computing the transformation for the forward heat equation $\hat{\eta}$:

$$\begin{aligned} \partial_t \hat{\eta}(t,x) &= \left(\partial_t \rho - \frac{\beta}{2} \rho \partial_t \Phi(t, M^{-1}x) \right) e^{-\beta \Phi(t, M^{-1}x)/2}, \\ \nabla \hat{\eta}(t,x) &= \left(\nabla \rho - \frac{\beta}{2} \rho M^{-1} \nabla \Phi(t, M^{-1}x) \right) e^{-\beta \Phi(t, M^{-1}x)/2}. \end{aligned}$$

The RHS can be computed as

$$\begin{aligned}\nabla\cdot(M\nabla\hat{\eta}(t,x)) = \nabla\cdot(M\nabla\rho)e^{-\beta\Phi(t,M^{-1}x)/2} - \beta\nabla\rho^\top\nabla\Phi(t,M^{-1}x)e^{-\beta\Phi(t,M^{-1}x)/2}\\ - \frac{\beta}{2}\rho\operatorname{Tr}\big(M^{-1}\nabla^2\Phi(t,M^{-1}x)\big)e^{-\beta\Phi(t,M^{-1}x)/2}\\ + \frac{\beta^2}{4}\rho\nabla\Phi(t,M^{-1}x)^\top M^{-1}\nabla\Phi(t,M^{-1}x)e^{-\beta\Phi(t,M^{-1}x)/2}.\end{aligned}$$

Applying the transformation to the forward heat equation in $\hat{\eta}$ yields

$$\begin{aligned}\partial_t\rho - \frac{\beta}{2}\rho\partial_t\Phi(t,M^{-1}x) = \beta^{-1}\nabla\cdot(M\nabla\rho) - \langle\nabla\rho,\nabla\Phi(t,M^{-1}x)\rangle\\ - \frac{1}{2}\rho\operatorname{Tr}\big(M^{-1}\nabla^2\Phi(t,M^{-1}x)\big) + \frac{\beta}{4}\rho\|\nabla\Phi(t,M^{-1}x)\|_M.\end{aligned}$$

Simplifying with the Hamilton–Jacobi equation for Φ (3.6), we obtain

$$\begin{aligned}&\partial_t\rho + \langle\nabla\rho,\nabla\Phi(t,M^{-1}x)\rangle + \rho\operatorname{Tr}\big(M^{-1}\nabla^2\Phi(t,M^{-1}x)\big) = \beta^{-1}\nabla\cdot(M\nabla\rho)\\ &\quad\Rightarrow \partial_t\rho + \nabla\cdot(\rho(t,x)\nabla\Phi(t,M^{-1}x)) = \beta^{-1}\nabla\cdot(M\nabla\rho).\end{aligned}$$

This is the desired (modified) Fokker–Planck equation in (3.2). We have demonstrated that the Cole–Hopf transformation (3.5) transforms the coupled anisotropic heat equations (3.4) into a diffusive coupled PDE system: a forward-time Fokker–Planck equation, and a backward-time Hamilton–Jacobi equation. Going backwards, we may infer that the coupled system admits a kernel solution corresponding to the anisotropic heat equation, and showing equivalence in Definition 3.2. We summarize this in the following proposition.

Proposition 3.3. *The coupled PDE system*

$$\text{(3.7)}\qquad \begin{cases}\partial_t\rho(t,x) + \nabla\cdot(\rho(t,x)\nabla\Phi(t,M^{-1}x)) = \beta^{-1}\nabla\cdot(M\nabla\rho)(t,x),\\ \partial_t\Phi(t,M^{-1}x) + \frac{1}{2}\|\nabla\Phi(t,M^{-1}x)\|_M^2 = -\beta^{-1}\operatorname{Tr}\big(M^{-1}(\nabla^2\Phi)(t,M^{-1}x)\big),\\ \rho(0,x) = \rho_0(x), \quad \Phi(T,M^{-1}x) = -V(x),\end{cases}$$

admits the following kernel formulation, where $G_{t,M}(x,y)$ is given in (3.1):

$$\text{(3.8)}\qquad \begin{cases}\rho(t,x) = \left(G_{T-t,M} * e^{-\beta V/2}\right)(x)\cdot\left(G_{t,M} * \frac{\rho_0}{(G_{T,M}*e^{-\beta V/2})}\right)(x),\\ \Phi(t,x) = 2\beta^{-1}\log\left(G_{T-t,M} * e^{-\beta V/2}\right)(Mx).\end{cases}$$

Proof. The PDEs under the change of variables (3.5) becomes the decoupled anisotropic heat equations (3.4). Using the Green's function, the solutions are given by

$$\eta(t,x) = G_{T-t,M} * \eta_T = (G_{T-t,M} * e^{-\beta V/2})(x).$$

From the initial boundary conditions,

$$\hat{\eta}(0,x) = \frac{\rho_0(x)}{(G_{T,M} * e^{-\beta V/2})(x)}.$$

From the Green's function again,

$$\hat{\eta}(t,x) = \left(G_{t,M} * \frac{\rho_0}{(G_{T,M} * e^{-\beta V/2})} \right)(x).$$

Since $\rho = \eta\hat{\eta}$ from (3.5),

$$\rho(t,x) = \left(G_{T-t,M} * e^{-\beta V/2} \right)(x) \cdot \left(G_{t,M} * \frac{\rho_0}{(G_{T,M} * e^{-\beta V/2})} \right)(x). \tag{3.9}$$

■

In particular, plugging in $t = T$, we have the following kernel formula for the PRWPO which is defined as the terminal solution.

Corollary 3.4. *The preconditioned regularized Wasserstein proximal admits the kernel formula:*

$$\begin{aligned} \mathrm{WProx}^{M}_{T,V}\, \rho_0(x) &= e^{-\beta V(x)/2} \int_{\mathbb{R}^d} \frac{G_{T,M}(y,x)\rho_0(y)}{(G_{T,M} * e^{-\beta V/2})(y)} \mathrm{d}y \\ &= \int_{\mathbb{R}^d} K(x,y,\beta,M,T,V)\rho(0,y)\,\mathrm{d}y\,, \end{aligned} \tag{3.10}$$

where

$$K(x,y,\beta,M,T,V) \coloneqq \frac{e^{-\frac{\beta}{2}(V(x)+\frac{\|x-y\|_M^2}{2T})}}{\int_{\mathbb{R}^d} e^{-\frac{\beta}{2}(V(z)+\frac{\|z-y\|_M^2}{2T})}\,\mathrm{d}z}. \tag{3.11}$$

Note that the PRWPO operator is well defined. The parabolic heat equations given by the Cole–Hopf transform are well defined for all time. Therefore, any probability density will map to a unique output, and moreover the mapping is continuous according to the regularity of the heat equations.

For notation purposes, we shorten the kernel (3.11) to $K_M(x,y)$. We may interpret K_M as a Markov kernel: for any y, $\int_x K_M(x,y)\,\mathrm{d}x = 1$, and moreover K_M is measurable. This is useful to determine how each particle affects the score, given by applying K_M to a Dirac mass.

4. Preconditioned Backwards Regularized Wasserstein Proximal Method. Equipped with a preconditioned version of the regularized Wasserstein proximal operator, we can proceed like in [45] and devise a sampling algorithm based on the Fokker–Planck component of the coupled PDEs. The Fokker–Planck component can be written as

$$\partial_t \rho(t,x) + \nabla \cdot \left(\rho(t,x)\nabla\Phi(t,M^{-1}x) - \beta^{-1}M\rho\nabla(\log\rho)(t,x) \right) = 0. \tag{4.1}$$

Using Liouville's equation with the Hamiltonian $\nabla\Phi(t,M^{-1}x) - \beta^{-1}M\nabla(\log\rho)(x)$, the corresponding particle formulation with density satisfying (4.1) is

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \nabla\Phi(t,M^{-1}X) - \beta^{-1}M\nabla\log\rho(t,X). \tag{4.2}$$

This can be seen to approximate the original Liouville equation since $\nabla\Phi(t, M^{-1}x) \approx -MV(x)$. Performing a backwards Euler discretization on this ODE, we define the following *preconditioned backward regularized Wasserstein proximal* (PBRWP) method:

$$X_{k+1} = X_k + \eta\left(-M\nabla V(X_k) - \beta^{-1}M\nabla \log \mathrm{WProx}^M_{T,V}\,\rho_k(X_k)\right), \tag{4.3}$$

where $\mathrm{WProx}^M_{T,V}\,\rho_k$ is the PRWPO of the distribution of X_k as given in (3.7). As in [45], this can be evaluated using the kernel formula (3.11).

Note that the formal limit as $T \to 0$ and $\eta \to 0$ is the score-based ODE with potential V,

$$\frac{\mathrm{d}X}{\mathrm{d}t} = -M\nabla V(X) - \beta^{-1}M\nabla \log \rho(X), \tag{4.4}$$

which has a stationary solution $\pi \propto \exp(-\beta V)$.

Based on (4.3) and the kernel formulation (3.11), we proceed as in [45] and consider the case where the input ρ_k to the regularized Fokker–Planck equation is given by an empirical measure. This arises when updating a collection of points that together approximate the underlying distribution.

Suppose that ρ_k is a uniform mixture of Dirac masses at points $\{\mathbf{x}_1^{(k)}, ..., \mathbf{x}_N^{(k)}\}$. From the kernel formulation (3.11), we have the following expressions to compute the score (dropping the k superscripts):

$$\mathrm{WProx}^M_{T,V}\,\rho_k(\mathbf{x}_i) = \frac{1}{N}\sum_{j=1}^N K_M(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{N}\sum_{j=1}^N \frac{\exp\left[-\frac{\beta}{2}\left(V(\mathbf{x}_i) + \frac{\|\mathbf{x}_i - \mathbf{x}_j\|_M^2}{2T}\right)\right]}{\mathcal{Z}(\mathbf{x}_j)}, \tag{4.5a}$$

$$\nabla\,\mathrm{WProx}^M_{T,V}\,\rho_k(\mathbf{x}_i) = \frac{1}{N}\sum_{j=1}^N \frac{\left(-\frac{\beta}{2}\left(\nabla V(\mathbf{x}_i) + M^{-1}\frac{\mathbf{x}_i - \mathbf{x}_j}{T}\right)\right)\exp\left[-\frac{\beta}{2}\left(V(\mathbf{x}_i) + \frac{\|\mathbf{x}_i - \mathbf{x}_j\|_M^2}{2T}\right)\right]}{\mathcal{Z}(\mathbf{x}_j)}, \tag{4.5b}$$

$$\mathcal{Z}(\mathbf{x}_j) \coloneqq \int_{\mathbb{R}^d} e^{-\frac{\beta}{2}(V(z) + \frac{\|z - \mathbf{x}_j\|_M^2}{2T})}\,\mathrm{d}z\,. \tag{4.5c}$$

We may observe the score $\nabla \log \rho = \frac{\nabla\rho}{\rho}$ as a particular weighted sum. Recall the definition of the softmax function: for a vector $v \in \mathbb{R}^d$, the softmax is the vector given by

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

satisfying $\sum_j \mathrm{softmax}(v)_j = 1$. We may rewrite the score in terms of the softmax function,

which makes it clear that it acts as an inter-particle diffusive term:

$$\begin{aligned}\nabla \log \mathrm{WProx}^M_{T,V}\,\rho_k(\mathbf{x}_i) &= \frac{\nabla\,\mathrm{WProx}^M_{T,V}\,\rho_k}{\mathrm{WProx}^M_{T,V}\,\rho_k}(\mathbf{x}_i)\\ &= \frac{\sum_{j=1}^N \frac{\left(-\frac{\beta}{2}\left(\nabla V(\mathbf{x}_i)+M^{-1}\frac{\mathbf{x}_i-\mathbf{x}_j}{T}\right)\right)\exp\left[-\frac{\beta}{2}\left(V(\mathbf{x}_i)+\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2_M}{2T}\right)\right]}{\mathcal{Z}(\mathbf{x}_j)}}{\sum_{j=1}^N \frac{\exp\left[-\frac{\beta}{2}\left(V(\mathbf{x}_i)+\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2_M}{2T}\right)\right]}{\mathcal{Z}(\mathbf{x}_j)}}\\ &= -\frac{\beta\nabla V(\mathbf{x}_i)}{2} - \frac{\beta}{2T}M^{-1}\mathbf{x}_i + \frac{\beta M^{-1}}{2T}\sum_{j=1}^N \mathrm{softmax}(U_{i,\cdot})_j\mathbf{x}_j,\end{aligned}$$

where we define the interaction kernel $U_{i,j} = U_{i,j}(\mathbf{x}_1, ..., \mathbf{x}_N)$ as the matrix

$$U_{i,j} = -\frac{\beta}{2}\left(V(\mathbf{x}_i) + \frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2_M}{2T}\right) - \log \mathcal{Z}(\mathbf{x}_j). \tag{4.6}$$

From (4.3) we can write the particle-wise PBRWP iterations

$$\begin{aligned}\mathbf{x}_i^{(k+1)} &= \mathbf{x}_i^{(k)} + \eta\left(-M\nabla V(\mathbf{x}_i^{(k)}) - \beta^{-1}M\nabla \log \mathrm{WProx}^M_{T,V}\,\rho_k(\mathbf{x}_i^{(k)})\right)\\ &= \mathbf{x}_i^{(k)} - \frac{\eta}{2}M\nabla V(\mathbf{x}_i^{(k)}) + \frac{\eta}{2T}\left(\mathbf{x}_i^{(k)} - \sum_{j=1}^N \mathrm{softmax}(U^{(k)}_{i,\cdot})_j\mathbf{x}_j^{(k)}\right)\\ &= \mathbf{x}_i^{(k)} - \frac{\eta}{2}M\nabla V(\mathbf{x}_i^{(k)}) + \frac{\eta}{2T}\left(\sum_{j=1}^N \mathrm{softmax}(U^{(k)}_{i,\cdot})_j(\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)})\right).\end{aligned}$$

We note that since $U_{i,j}$ is only used in the softmax function, it is possible to add a constant for each column without changing the dynamics. In particular, we define a simpler matrix $W_{i,j}$ such that

$$\mathrm{softmax}(U_{i,\cdot})_j = \mathrm{softmax}(W_{i,\cdot})_j,$$

defined by adding constants depending only on the row i:

$$\begin{aligned}W_{i,j} &:= U_{i,j} + \frac{\beta V(\mathbf{x}_i)}{2}\\ &= -\beta\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2_M}{4\beta T} - \log \mathcal{Z}(\mathbf{x}_j).\end{aligned} \tag{4.7}$$

This has the advantage of avoiding one computation of V, which is already implicitly used in the normalizing constants $\mathcal{Z}$. To further simplify, let $\mathbf{X}$ be the matrix of particles

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

The above iteration can thus be written as

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} - \frac{\eta}{2} M \nabla V(\mathbf{X}^{(k)}) + \frac{\eta}{2T}\left(\mathbf{X}^{(k)} - \mathbf{X}^{(k)}\mathrm{softmax}(W^{(k)})^\top\right), \tag{4.8}$$

where $\nabla V(\mathbf{X}) = \begin{bmatrix}\nabla V(\mathbf{x}_1) & \dots & \nabla V(\mathbf{x}_N)\end{bmatrix}$, W is defined in (4.7), and $\mathrm{softmax}(W)_{i,j} = \mathrm{softmax}(W_{i,\cdot})_j$ is a row-stochastic matrix (and therefore, $\mathrm{softmax}(W)^\top$ is a column-stochastic matrix). This formulation has the advantage of being easily parallelizable (or serially computed, depending on memory restrictions), assuming that the normalizing constant can be easily computed. Summarizing, the PBRWP update is given in Algorithm 1, which details the steps used in practice.

Algorithm 1: PBRWP: Preconditioned Backwards Regularized Wasserstein Proximal

Data: Initial points $\mathbf{x}_1^{(1)}, \dots, \mathbf{x}_N^{(1)} \in \mathbb{R}^d$, potential $V : \mathbb{R}^d \to \mathbb{R}$, preconditioner $M \in \mathrm{Sym}_{++}(\mathbb{R}^d)$, regularization parameter $T > 0$, diffusion $\beta > 0$, step-size $\eta > 0$, iteration count K.

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

Tensorize $\mathbf{X}^{(1)} = \begin{bmatrix}\mathbf{x}_1^{(1)} & \dots & \mathbf{x}_N^{(1)}\end{bmatrix} \in \mathbb{R}^{d\times N}$;

for $k = 1, \dots, K$ **do**

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

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

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

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

end

Remark 4.1. The corresponding limit of (4.3) as $\eta \to 0$ satisfies the modified Fokker–Planck equation

$$\partial_t \rho = \beta^{-1} \nabla \cdot \left(\rho M \nabla \log \frac{\mathrm{WProx}_{T,V}^M(\rho)}{\exp(-\beta V)}\right). \tag{4.9}$$

A stationary solution is $\hat{\pi}$ satisfies $\mathrm{WProx}_{T,V}^M(\hat{\pi}) \propto \exp(-\beta V)$. This is a semi-implicit discretization of (4.1), where $\rho \nabla \log \rho(t,x)$ is replaced with $\rho \nabla \log \mathrm{WProx}_{T,V}^M(\rho)$, and the dual variable Φ is also replaced with the terminal condition $\Phi(T, M^{-1}x) = -V(x)$.

4.1. Transformer reformulation. We may connect (4.8) to self-attention mechanisms, which are commonly employed in transformers [48], and have recently attracted interpretations as particle evolutions [38, 18]. Recall that an attention block takes the following form: for

queries $Q \in \mathbb{R}^{d\times N}$, keys $K \in \mathbb{R}^{d\times N}$ and values $V \in \mathbb{R}^{d_V\times N}$ of the prescribed dimensions[1],

$$\mathrm{Attn}(Q;K,V) = V\mathrm{softmax}\left(\frac{Q^\top K}{\sqrt{d}}\right)^\top \in \mathbb{R}^{d_V\times N}. \tag{4.10}$$

We may connect this with the softmax term in (4.8). The argument of the softmax (4.7) is

$$\begin{aligned} W_{i,j} &= -\beta\frac{\|\mathbf{x}_i-\mathbf{x}_j\|_M^2}{4T} - \log \mathcal{Z}(\mathbf{x}_j) \\ &= \beta\frac{\mathbf{x}_i^\top M^{-1}\mathbf{x}_j}{2T} - \log\mathcal{Z}(\mathbf{x}_j) - \beta\frac{\|\mathbf{x}_i\|_M^2+\|\mathbf{x}_j\|_M^2}{4T}. \end{aligned}$$

Discarding the $\sqrt{d}$ scaling[2] of (4.10) and the constant-in-j term $\|\mathbf{x}_i\|_M^2$, the diffusion term $\mathbf{X}^{(K)}\mathrm{softmax}(W^{(k)})^\top$ for N particles takes the form of masked attention

$$V\mathrm{softmax}\left(Q^\top K - \mathbf{1}\mathbf{z}^\top\right)^\top, \tag{4.11}$$

where $Q^\top K = \frac{\beta}{2T}\mathbf{X}^\top M^{-1}\mathbf{X}$, $\mathbf{z}\in\mathbb{R}^N$ is the vector of normalizing constants $\mathbf{z}_j = \log\mathcal{Z}(\mathbf{x}_j) + \beta\frac{\|\mathbf{x}_j\|_M^2}{4T}$, and $V=\mathbf{X}$.

For an input $H \in \mathbb{R}^{d\times N}$, the self-attention framework imposes the particular forms $V = W_V H$, $K = W_K H$, $V = W_V H$, where $W_V, W_K \in \mathbb{R}^{d\times d}$, $W_V \in \mathbb{R}^{d_V\times d}$ are some admissible matrices. By taking $H = \mathbf{X}$, the preconditioning arises as $W_V^\top W_K = M^{-1}$, and $W_V = I$, making the diffusion in PBRWP interpretable as a modified self-attention.

Remark 4.2. In the other direction, modeling transformers as interacting particle systems, [38] consider self-attention transformers with symmetric interaction and an additional condition $W_Q^\top W_K = W_K^\top W_Q = -W_V$, demonstrating that transformers can be interpreted as a particular Wasserstein gradient flow, converging to a nonlinear PDE. [13] consider transformers acting on the sphere $\mathbb{S}^{d-1}$ in the special case $W_Q = W_K = I$, presenting various conditions on the dimension, number of particles, and temperature such that particles will cluster. [7] demonstrate well-posedness of the induced PDE for various forms of self-attention, as well as a continuous-time analysis of the Gaussian-preserving properties. These works consider the continuous limit of the transformers to obtain these links.

5. Analysis in Gaussian distributions.

In the particular case where $V(x) = \frac{x^\top\Sigma^{-1}x}{2}$, i.e. the target distribution is $\mathcal{N}(0,\beta^{-1}\Sigma)$, the PRWPO and PBRWP iterations can be computed in closed form. Suppose M is now such that $cM \preceq \Sigma \preceq CM$, where $0 < c \le C < \infty$, and $T < c$. For Gaussian distributions, the PRWPO admits a closed form as follows.

Proposition 5.1. *For a Gaussian density $\rho_k = \mathcal{N}(\mu_k,\Sigma_k)$ and the quadratic potential $V(x) = x^\top\Sigma^{-1}x/2$, the PRWPO operator satisfies*

$$\mathrm{WProx}_{T,V}^M(\rho_k) = \mathcal{N}(\tilde{\mu}_k,\tilde{\Sigma}_k), \tag{5.1}$$

[1]The matrices are transposed compared to the literature in order to be consistent with the convention in this work. The usual form is $\mathrm{Attn}(Q;K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V$, where the softmax is taken row-wise as in this work. Both forms are equivalent.

[2]This will be reintroduced in Subsection 6.4 as a heuristic modification in high dimensions.

where $\tilde{\mu}_k, \tilde{\Sigma}_k$ are given by

$$\tilde{\mu}_k = (I + TM\Sigma^{-1})^{-1}\mu_k, \tag{5.2}$$

$$\tilde{\Sigma}_k = 2\beta^{-1}T\left(T\Sigma^{-1} + M^{-1}\right)^{-1} + \left(T\Sigma^{-1} + M^{-1}\right)^{-1} M^{-1}\Sigma_k M^{-1}\left(T\Sigma^{-1} + M^{-1}\right)^{-1}. \tag{5.3}$$

Moreover, the kernel $K_M(x, y)$ in (3.11) takes the form of a Gaussian density

$$K_M(\cdot, y) \sim \mathcal{N}\left(\left(\frac{\Sigma^{-1}}{2} + \frac{M^{-1}}{2T}\right)^{-1}\frac{M^{-1}}{2T}y, \left(\frac{\beta\Sigma^{-1}}{2\beta} + \frac{\beta M^{-1}}{2T}\right)^{-1}\right). \tag{5.4}$$

Proof Sketch. The kernel formula can be computed in closed form for Gaussians, and the product of two Gaussian densities is again proportional to a Gaussian density. The full proof is given in B.2. ∎

Since the PRWPO maps Gaussian distributions to Gaussian distributions, the score function is linear, and the PBRWP iterations also preserve Gaussianity.

Corollary 5.2. *For a Gaussian density $\rho_k = \mathcal{N}(0, \Sigma_k)$, the PBRWP iteration satisfies the update $\rho_{k+1} = \mathcal{N}(0, \Sigma_{k+1})$, where*

$$\Sigma_{k+1} = (I - \eta M\Sigma^{-1} + \eta\beta^{-1}M\tilde{\Sigma}_k^{-1})\Sigma_k(I - \eta M\Sigma^{-1} + \eta\beta^{-1}M\tilde{\Sigma}_k^{-1})^\top. \tag{5.5}$$

Proof. Follows from using Proposition 5.1 in the PBRWP update (4.3). ∎

Now that we have shown that the PBRWP iteration preserves Gaussianity, we wish to consider the convergence for discrete time-steps. In the case where the initial and target distributions are both Gaussian, we can show that the distributions generated by PBRWP with discrete step-sizes converge to a Gaussian. Moreover, the terminal distributions are independent of step-size, only depending on T, the terminal Gaussian and the preconditioner. The following result extends the result of [45] from continuous time to discrete time, as well as adding generality by working in the preconditioned setting. We consider the zero-mean initialization case for brevity, as the covariances do not depend on the mean; moreover, it can be shown that the means of the iterates converge linearly to the target mean.

Theorem 5.3. *Suppose $V(x) = x^\top\Sigma^{-1}x/2$, and fix a $\beta > 0$. For the target distribution $\pi = \mathcal{N}(0, \beta^{-1}\Sigma)$ and positive definite preconditioner M, let $c, C > 0$ be such that $cM \preceq \Sigma \preceq CM$. Let $T \in (0, c)$ so that the inverse PRWPO is well-defined. Then there exists a stationary distribution to the discrete-time PBRWP iterations, and it is Gaussian $\hat{\pi} = \mathcal{N}(0, \hat{\Sigma})$, satisfying $\mathrm{WProx}_{T,V}^M\hat{\pi} = \pi$. Moreover, it is unique within distributions with sufficiently decaying Fourier transforms.*

Suppose further that the initial Gaussian distribution $\rho_0 = \mathcal{N}(0, \Sigma_0)$ has its covariance commuting with Σ. Let ρ_k be the iterations generated by PBRWP, and let $\tilde{\rho}_k = \mathrm{WProx}_{T,V}^M(\rho_k)$. Further let λ be such that $\sqrt{M}^{-1}\Sigma_k\sqrt{M}^{-1}$ are uniformly bounded below by some λ (which holds as $\Sigma_k \to \Sigma$). Let the step-size η be bounded by

$$\eta\beta^{-1} \leq \min((\max_i |\lambda_i(\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})|)^{-1}/2, 3\lambda_{\min}(\tilde{\Xi}_k)/32)$$

Then the exact discrete-time PBRWP method has the following decay in KL divergence:

$$\mathrm{D}_{\mathrm{KL}}(\tilde{\rho}_{k+1}\|\pi) - \mathrm{D}_{\mathrm{KL}}(\tilde{\rho}_k\|\pi) \leq -\frac{\eta}{2C[\beta + 2\ T(1+TC^{-1})^{-1}(1+Tc^{-1})^2\lambda^{-1}]}\mathrm{D}_{\mathrm{KL}}(\tilde{\rho}_k\|\pi). \tag{5.6}$$

Sketch Proof. From the condition, there exists a (Gaussian) distribution $\mathcal{N}(0,\hat{\Sigma})$ such that $\mathrm{WProx}^M_{T,V}(\mathcal{N}(0,\hat{\Sigma})) = \mathcal{N}(0,\beta^{-1}\Sigma)$. Uniqueness comes from the closed-form kernel (5.4) and the invertibility of Gaussian convolution on the range of $\mathrm{WProx}^M_{T,V}$. Moreover, the PBRWP iterations generate Gaussian distributions, with closed-form covariance updates. The KL divergence between two Gaussians has a closed form in terms of the means and covariances, from which the decay comes using a Taylor expansion. The full proof is given in B.3. ∎

5.1. Additional properties. We briefly list some properties of the PRWPO and PBRWP, and refer to appendix C for more detailed statements and proofs.

1. *Conditions for PRWPO invertibility.* For a density ρ, there exists a unique distribution ρ_0 such that $\rho = \mathrm{WProx}^M_{T,V}\,\rho_0$ if and only if $\rho(x)e^{\beta V(x)/2}$ has Fourier transform decaying at least as fast than $e^{-T\beta^{-1}\xi^\top M\xi}$. In the Gaussian case $V = x^\top\Sigma^{-1}x/2$, this is equivalent to $\Sigma \succeq TM$.
2. *The regularized Wasserstein proximal is a $\mathcal{W}_2$ contraction.* For quadratic V, let $\zeta := (\frac{C^{-1}}{2} + \frac{1}{2T})^{-1}(\frac{1}{2T}) < 1$, where C is as in Theorem 5.3. For any μ,ν in Wasserstein-2 space,
$$\mathcal{W}_2(\mathrm{WProx}^M_{T,V}\,\mu, \mathrm{WProx}^M_{T,V}\,\nu) \leq \zeta\mathcal{W}_2(\mu,\nu). \tag{5.7}$$
3. *Mean-variance tradeoff.* For strongly convex V and any $\beta > 0$, the particles move together towards the minimizer, at a rate balanced between the inter-particle variance and the distances to the minimizer. In particular, let $\hat{x}$ be the minimizer of V, and let $\hat{\mathbf{X}} = \begin{bmatrix}\hat{x} & \dots & \hat{x}\end{bmatrix} = \hat{x}\mathbf{1}_N^\top \in \mathbb{R}^{d\times N}$. In particular, suppose V is μ-relatively strongly convex with respect to M. Then for any $\mathbf{X}\in\mathbb{R}^{d\times N}$,
$$\langle\Delta, M^{-1}(\mathbf{X}-\hat{\mathbf{X}})\rangle_{\mathrm{Frob}} \leq \|\mathbf{X}-\hat{\mathbf{X}}\|_{2,M}\left(-\frac{\mu}{2}\|\mathbf{X}-\hat{\mathbf{X}}\|_{2,M} + \frac{(1+\sqrt{N})}{2T}\|\mathbf{X}-\frac{1}{N}\mathbf{X}\mathbf{1}_N\mathbf{1}_N^\top\|_{2,M}\right).$$
4. *Diffusion is bounded based on number of particles.* For quadratic $V = x^\top\Sigma^{-1}x/2$, suppose $M = \Sigma$ (and $T < 1$). Suppose that $\mathbf{x}_1$ is an exterior point of the convex hull of $\{\mathbf{x}_i\}$, and further that $\delta > 0$ is such that for all $j \neq 1$,
$$\mathbf{x}_1^\top M^{-1}\mathbf{x}_j \leq \mathbf{x}_1^\top M^{-1}\mathbf{x}_1 - \delta\|\mathbf{x}_1\|_M. \tag{5.8}$$
Then, if $\mathbf{x}_1$ has sufficiently large norm, in particular assuming,
$$\delta\|\mathbf{x}_1\|_M \geq 2\beta^{-1}T\log\left(\frac{2(N-1)}{T}\right), \tag{5.9}$$
then for small step-size, the updated point satisfies $\|\mathbf{x}_1^{(k+1)}\|_M < \|\mathbf{x}_1^{(k)}\|_M$. In other words, the norm of the outermost particle scales at most as $\mathcal{O}(\log N)$.

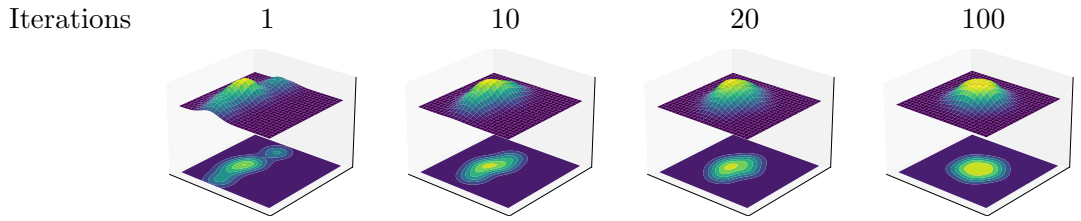


Figure 1. *Evolution of the preconditioned regularized Wasserstein proximal* $\mathrm{WProx}^{I}_{T=0.2,I}$ *for the 2-dimensional standard Gaussian, done with 5 particles and step-size of* $\eta = 0.1$*. The bandwidth is automatically determined by the regularization. As suggested by theory, the PRWPO of the particles approaches the standard Gaussian.*

6. Experiments. In this section, we present various two-dimensional toy examples to demonstrate the effects of adding a preconditioning matrix on the sampling behavior, followed by high-dimensional Bayesian total-variation regularized image deconvolution, and Bayesian neural network training. We compare against the MCMC-based baselines ULA, MALA and the mirror Langevin algorithm (MLA) [21, 23] for the low-dimensional examples, and various kernel-based methods to be detailed for the Bayesian neural network experiment. Throughout, we fix $\beta = 1$.

6.1. Gaussian low particle regime. We first consider the evolution of (P)BRWP for the two-dimensional standard Gaussian $\mathcal{N}(0, I_2)$, with the identity preconditioner $M = I$. In this case, for a fixed $T \in (0, 1)$, Theorem 5.3 gives that the RWPO (with parameter T) of the iterated distributions should approach $\mathcal{N}(0, I_2)$. Moreover, the RWPO for this quadratic distribution is given by the closed-form Gaussian kernel (5.4). In particular, each point gets mapped into a Gaussian with covariance $(\frac{1}{2} + \frac{1}{2T})^{-1} I$.

Figure 1 demonstrates the density evolution of the RWPO of the (P)BRWP iterations, evaluated with 5 particles, regularization $T = 0.2$ and step-size $\eta = 0.1$. We observe as expected from Theorem 5.3 that the RWPO converges to that of the standard Gaussian, even in the finite particle setting. Figure 2 demonstrates the effect of the number of particles from 3 to 6, evaluated at convergence at iteration 100 and with the same parameters above. In particular, we observe that the contours of the RWPO are initially triangular, then square and pentagonal, gradually becoming smoother and better approximating the standard Gaussian. We observe as well in the 6-particle case that while the contour is pentagonal, the density is more rounded at the origin, indicating that there is one particle there surrounded by 5 other particles.

6.2. Bimodal distribution. As a non-convex potential, we consider the two-moons bimodal distribution as given in [50, 45],

$$p(x) \propto \exp\big(-2(\|x\| - 3)^2\big) \left[\exp\big(-2(x_1 - 3)^2\big) + \exp\big(-2(x_1 + 3)^2\big)\right].$$

This is generated by the potential V with corresponding gradient ∇V as follows:

$$V(x) = 2(\|x\| - 3)^2 - 2\log\left[\exp\big(-2(x_1 - 3)^2\big) + \exp\big(-2(x_1 + 3)^2\big)\right], \tag{6.1a}$$

$$\nabla V(x) = 4\frac{(\|x\| - 3)x}{\|x\|} + \frac{4(x_1 - 3)\exp\big(-2(x_1 - 3)^2\big) + 4(x_1 + 3)\exp\big(-2(x_1 + 3)^2\big)}{\exp(-2(x_1 - 3)^2) + \exp(-2(x_1 + 3)^2)}\mathrm{e}_1, \tag{6.1b}$$

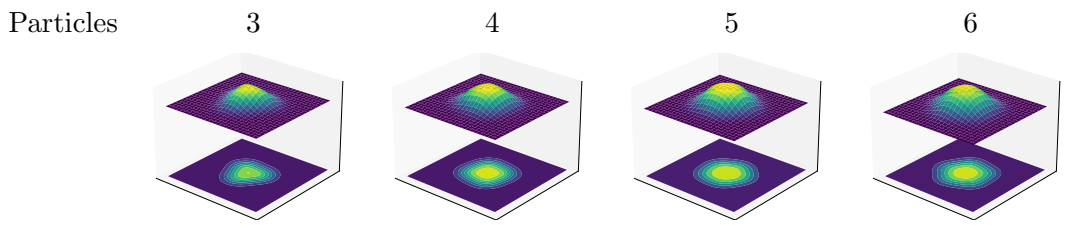


Figure 2. *Densities of the PRWPO* $\mathrm{WProx}^{I}_{0.2I}$ *for the 2-dimensional standard Gaussian at iteration 100, done with* $n \in \{3, 4, 5, 6\}$ *particles and a step-size of* $\eta = 0.1$*. We observe that the density of the Wasserstein proximal gradually becomes more spherical and Gaussian-like.*

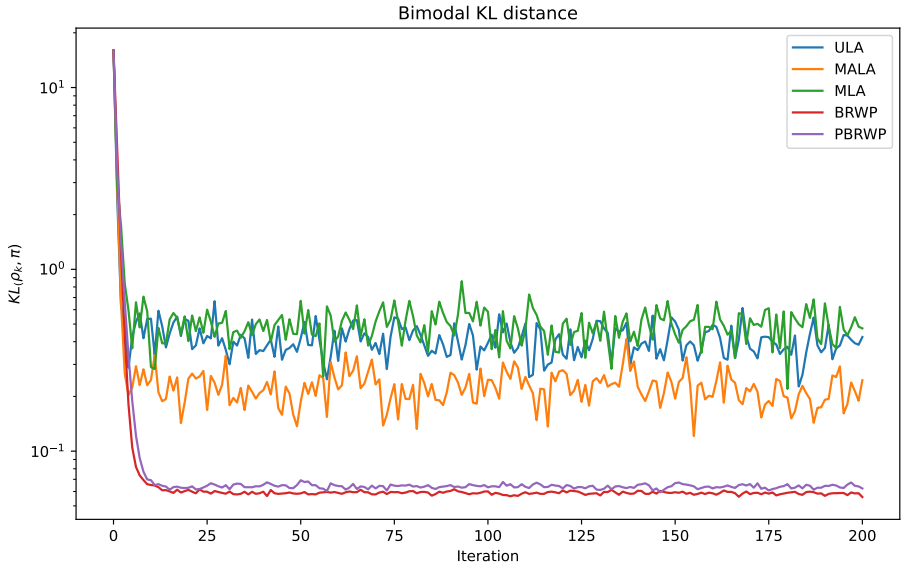


Figure 3. *Evolution of the KL divergence between baselines and BRWP-based methods for the bimodal distribution. Applied with 100 particles with initial distribution* $\mathcal{N}(0, I)$*, fixed step-size* $\eta = 0.1$ *and regularization parameter* $T = 0.05$*. We observe that the BRWP-based methods converge more smoothly even in later iterations, and the particles better approximate the target distribution.*

where e_1 is the first standard basis vector. We consider a low-variance initialization and a high-variance initialization. For the low-variance initialization, we consider the same setup as [45], where the particles are initially distributed as $\mathcal{N}(0, I)$; for the high-variance setup, the initial distribution is $\mathcal{N}(0, 6I)$. Figure 3 demonstrates that in this case, adding preconditioning has minimal effect on the convergence rate of BRWP and PBRWP, and indeed leads to a slightly higher bias. The bias can be seen qualitatively in Figure 4, which demonstrates that the particles retain the same structured behavior at convergence, but takes a different path at small iterations.

We observe a different compression phenomenon with the high-variance initialization, when the samples are initially distributed around the modes instead of between the modes. Figure 5

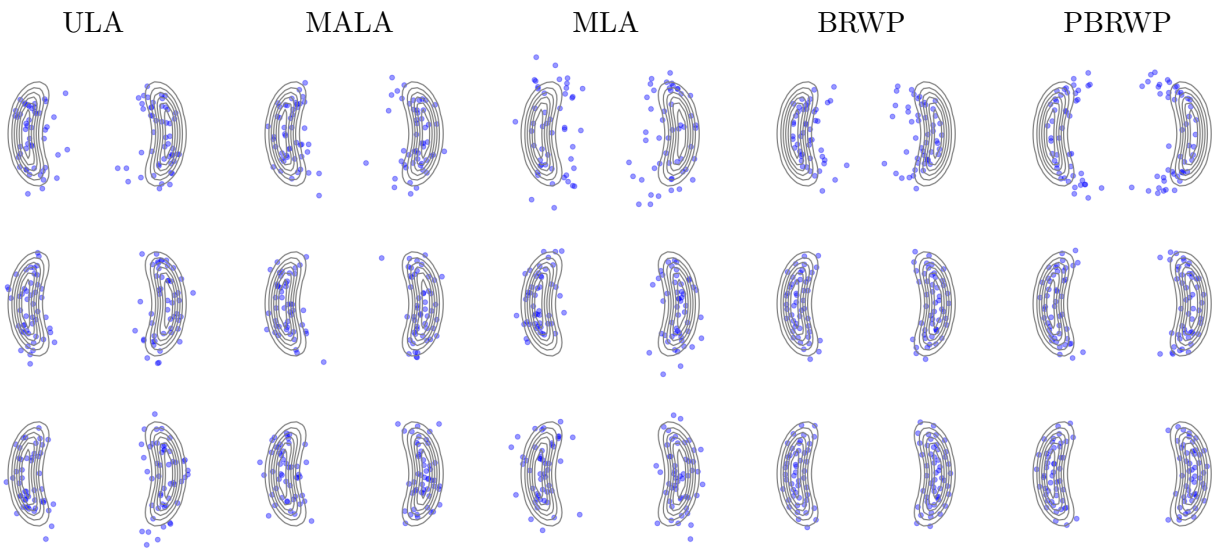


Figure 4. *Evolution of the various methods for the bimodal distribution at different iterations, with initial distribution $\mathcal{N}(0, I)$, which is contained between the moons. Evaluated with 100 particles, at iterations 2, 5 and 10 in the top, middle, and bottom rows respectively. Observe the stable behavior of BRWP and PBRWP, as opposed to the randomness of the Langevin methods.*

demonstrates that in this case, the preconditioning has a significant acceleration effect, where the KL divergence decreases faster than BRWP. Moreover, Figure 6 demonstrates that PBRWP particles are less noisy than BRWP particles at low iterations.

6.3. Annulus. Consider the two-dimensional annulus defined with the potential

$$V(x) = \left(\left\| \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} x \right\| - 3 \right)^2, \quad \nabla V(x) = 2 \frac{\left\| \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} x \right\| - 3}{\left\| \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} x \right\|} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}^2 x. \tag{6.2}$$

We consider $M = \mathrm{diag}([4, 1])$, which takes into account the scaling of the annulus in both directions. The initialization is taken to be the off-center Gaussian $\mathcal{N}((2, 2), I)$, and the particles should diffuse along the elliptical potential well. Figure 7 shows the evolution of various methods from this initialization, evaluated with 100 particles. We observe that the preconditioned methods MLA and PBRWP are able to diffuse faster than their non-preconditioned counterparts, namely covering the annulus at iteration 50. Figure 8 verifies this quantitatively, where the PBRWP particles converge significantly faster than BRWP, and to a lower minimum than the Langevin methods.

6.4. High dimensional examples and modifications. In high dimensions, approximating the normalizing constant with a Monte Carlo integral is highly inaccurate [45]. This manifests empirically as non-interaction between particles, or excessive interaction leading to diverging particles. Therefore, we require another approximation using Laplace's method [3], similarly to

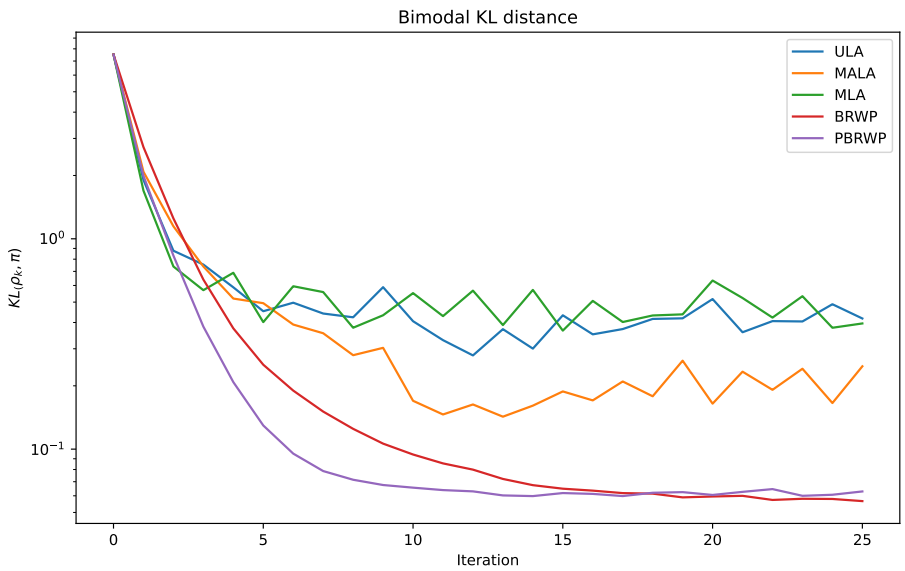


Figure 5. *Evolution of the KL divergence between baselines and BRWP-based methods for the bimodal distribution, initialized with large variance $\mathcal{N}(0, 6I)$. Applied with 100 particles, fixed step-size $\eta = 0.1$ and regularization parameter $T = 0.05$. In this case, the preconditioning accelerates the convergence.*

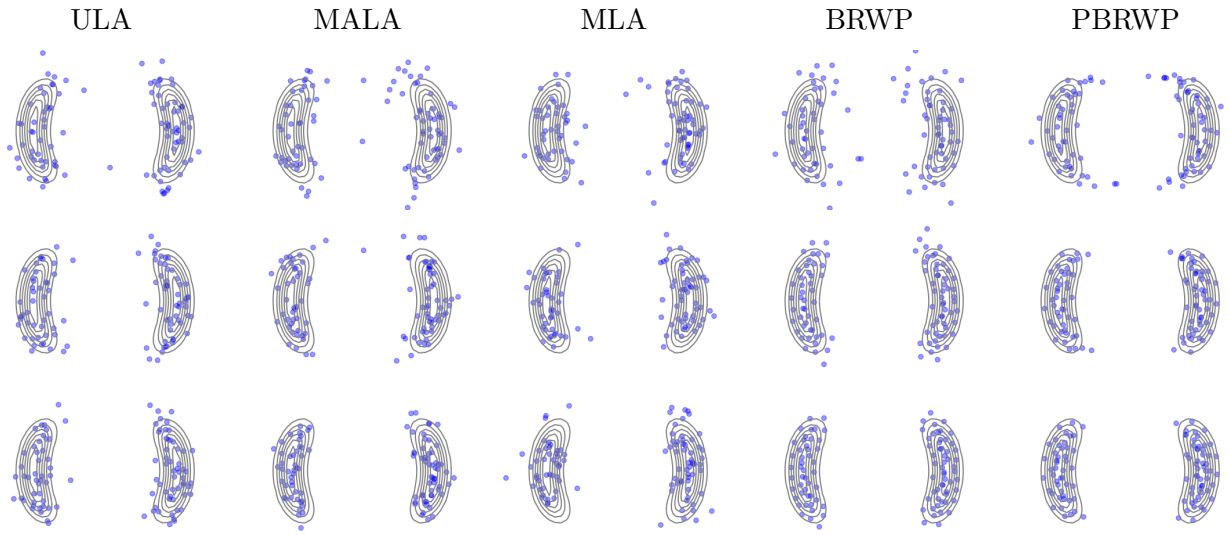


Figure 6. *Evolution of the various methods for the bimodal distribution, with large initial variance $\mathcal{N}(0, 6I)$, which surrounds the moons. Evaluated with 100 particles, at iterations 2, 5 and 10 in the top to bottom rows respectively. We observe again the structure behavior at convergence of the noise-free BRWP and PBRWP methods. Moreover, the preconditioning affects the empirical covariance of the particles in different directions, which can be seen in the semicircular artifacts between the modes at low iterations.*

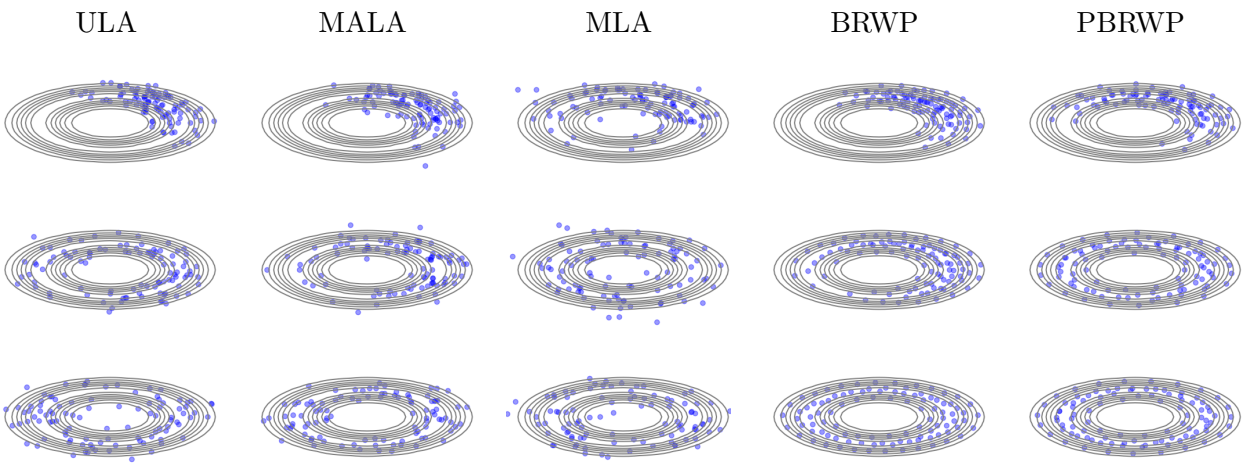


Figure 7. *Evolution of the various methods for the scaled annulus. Evaluated with 100 particles, at iterations 10, 50, and 200 from top to bottom respectively. We observe that PBRWP and MLA diffuse faster than their non-preconditioned counterparts. Moreover, PBRWP retains a similar level-set structure to BRWP.*

[18, 17]. This approximation has also been used in the other direction, e.g. [46] using integrals to approximate infimal convolutions.

Laplace's method consists of the following approximation: for a $\mathcal{C}^2$ function f and a continuous function g, assume that f has a unique global minimizer x^* and satisfies some coercivity condition. Then, we have the following approximation:

$$\sqrt{\det(\nabla^2 f(x^*))}\frac{\exp(f(x^*)/T)}{(2\pi T)^{d/2}}\int_{\mathbb{R}^d} g(x)\exp(-f(x)/T)\,\mathrm{d}x \to g(x^*) \quad \text{as } T \to 0^+. \tag{6.3}$$

Applying this with $f(x) = \beta\|x - \mathbf{x}_j\|_M^2/4$ and $g(x) = \exp(-\beta V(x)/2)$, we obtain the approximation

$$\mathcal{Z}(\mathbf{x}_j) = \int_{\mathbb{R}^d} \exp\left(-\frac{\beta}{2}\left(V(z) - \frac{\|z - \mathbf{x}_j\|_M^2}{2T}\right)\right)\mathrm{d}z \tag{6.4}$$

$$\approx \sqrt{\det M}\exp\left(-\frac{\beta}{2}V(\mathbf{x}_j)\right)C(\beta, T), \tag{6.5}$$

where $C = C(\beta, T)$ is a constant independent of $\mathbf{x}_j$. We note that actually only a Hölder type condition is required for a similar asymptotic to hold, albeit with a different normalizing factor [46, Thm. 2]. Using this, the $\log \mathcal{Z}(\mathbf{x}_j)$ term may be approximated by $-\beta V(\mathbf{x}_j)/2 + \log C(\beta, T) + \frac{1}{2}\log\det M$. For constant M, the latter two terms will disappear in the softmax. We refer to this as the Laplace approximation in the following section.

In high dimensions, the matching term $\|\mathbf{x}_i - \mathbf{x}_j\|_M$ can be large for $i \neq j$, leading to the interaction matrix $\mathrm{softmax}(U_{i,.})_j$ being close to identity. Similarly to the seminal work on transformers [48], we additionally propose to set the diffusion parameter as $\beta = d^{-1/2}$, which increases the diffusion. We refer to this as scaling in the following section.

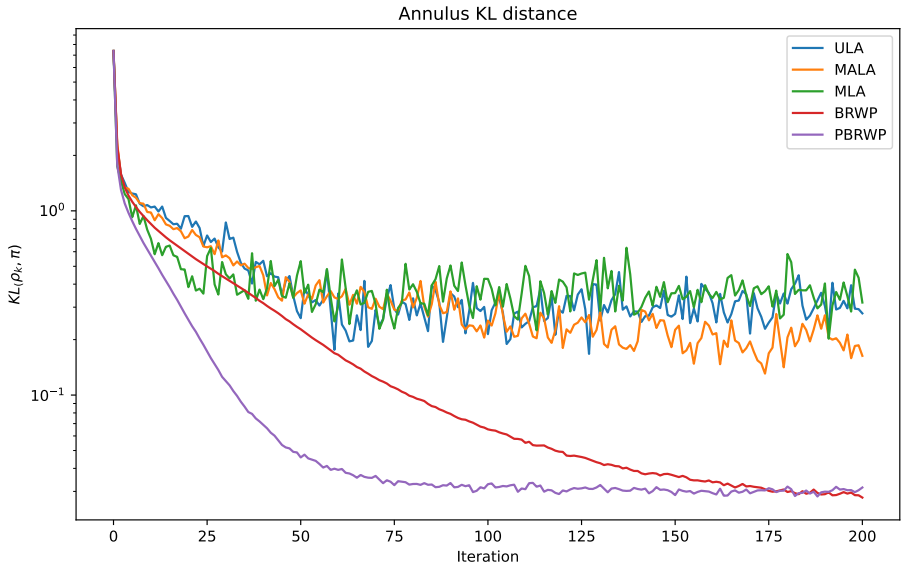


Figure 8. *Evolution of the KL divergence between baselines and BRWP-based methods for the scaled annulus. Applied with 100 particles, fixed step-size $\eta = 0.1$ and regularization parameter $T = 0.05$. We observe that the BRWP-based methods converge more smoothly.*

The resulting modified iteration using scaling and the Laplace approximation is as follows, where $\beta = d^{-1/2}$,

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} - \frac{\eta}{2} M \nabla V(\mathbf{X}^{(k)}) + \frac{\eta}{2T}\left(\mathbf{X}^{(k)} - \mathbf{X}^{(k)} \mathrm{softmax}(W^{(k)})^\top\right), \tag{6.6a}$$

$$W_{i,j}^{(k)} = -\beta \frac{\|\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)}\|_M^2}{4T} + \frac{\beta}{2} V(\mathbf{x}_j^{(k)}). \tag{6.6b}$$

6.5. High-dimensional Gaussian distributions. We demonstrate the effect of taking $\beta = d^{-1/2}$ instead of $\beta = 1$, as well as the Laplace approximation in Figure 9. We test on a 50-dimensional Gaussian with diagonal covariance $\Sigma = \mathrm{diag}(0.1, 0.2, ..., 5)$ and condition number 50. We let $M = \Sigma$ be the preconditioner as before, and plot the particles projected onto the first and last dimensions in the horizontal and vertical directions respectively. Recall that in the Gaussian case, the normalizing constant can be computed exactly in the finite particle setting. We compare PBRWP with step-size $\eta = 0.1$, $T = 0.02, 0.2, 0.9$, with and without both the scaling and Laplace approximation after convergence at 1000 iteration.

We observe that the Laplace approximation does not seem to qualitatively affect the diffusion behavior of the particles, with or without the $\beta = d^{-1/2}$ scaling modification. However, the scaling is able to counter the mode-collapse phenomenon that comes with high dimensions. By reducing the scale of the interaction matrix, the particles are able to repel each other from further away. Motivated by this, we use the scaling for high dimensions to avoid mode collapse.

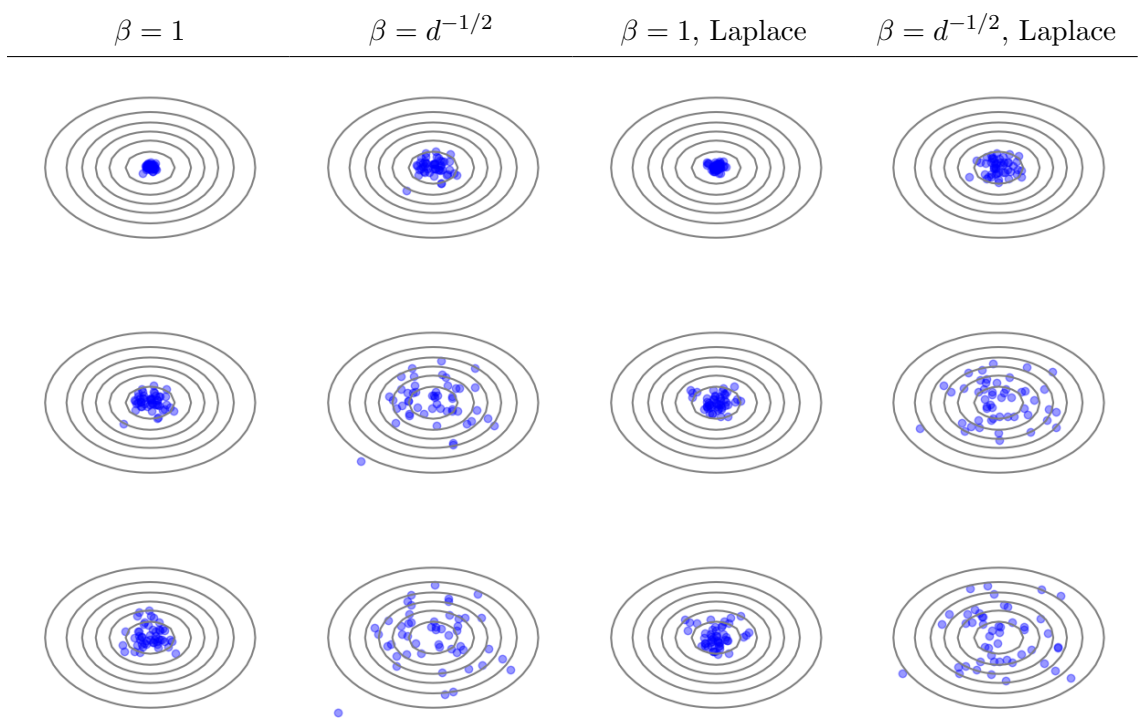


Figure 9. *Evolution of the high-dimensional modifications for the 50-dimensional Gaussian, at convergence in 1000 iterations. Evaluated with 50 particles, step-size $\eta = 0.1$ and regularizations $T = 0.02, 0.2, 0.9$ in the top, middle and bottom rows respectively. We observe little difference when using the Laplace approximation compared to the ground-truth, suggesting this is reasonable. The $\beta = d^{-1/2}$ scaling increases the diffusion, which is crucial for reducing mode-collapse in high dimensions.*

6.6. Deconvolution. We consider the Bayesian problem corresponding to the (convex) total-variation regularized (TV) objective [37],

$$V(x) = \frac{1}{2\sigma^2}\|Ax - y\|_2^2 + \lambda \mathrm{TV}(x), \tag{6.7}$$

where A is a convolution operator, y is a corrupted image, and $\mathrm{TV}(x) = \|Dx\|_1$ denotes the discrete total variation functional. For image deconvolution, the forward operator A takes the form $A = \mathcal{F}^*\Lambda\mathcal{F}$, where $\mathcal{F}$ is the (complex, unitary) matrix of the discrete Fourier transform, $\mathcal{F}^*$ is its inverse, and Λ is a diagonal matrix. In this case, the Hessian of the first term takes the simple form $\sigma^{-2}A^*A = \sigma^{-2}\mathcal{F}^*\Lambda^*\Lambda\mathcal{F}$. Similarly to SALSA [1], we may use a regularized version as a preconditioner $M = (A^*A + \tau I)^{-1}$, where $\tau > 0$ is some constant, taken to be $\tau = 0.5$. We note that sampling from the distribution $\mathcal{N}(0, M)$ is possible using the equation $\sqrt{M} = \mathcal{F}^*(\Lambda^*\Lambda + \tau I)^{-1/2}\mathcal{F}$, which allows for the use of MLA. We further note that using the regularized Hessian $(\sigma^2 A^*A + \tau I)^{-1}$ does not provide adequate preconditioning for PBRWP or MLA, likely due to the ill-conditioning of the TV term with respect to this metric.

We compare the mean and (norm of the channel-wise) standard deviation of the TV-regularized objective. In addition, we compute the PSNR of the finite-particle posterior mean

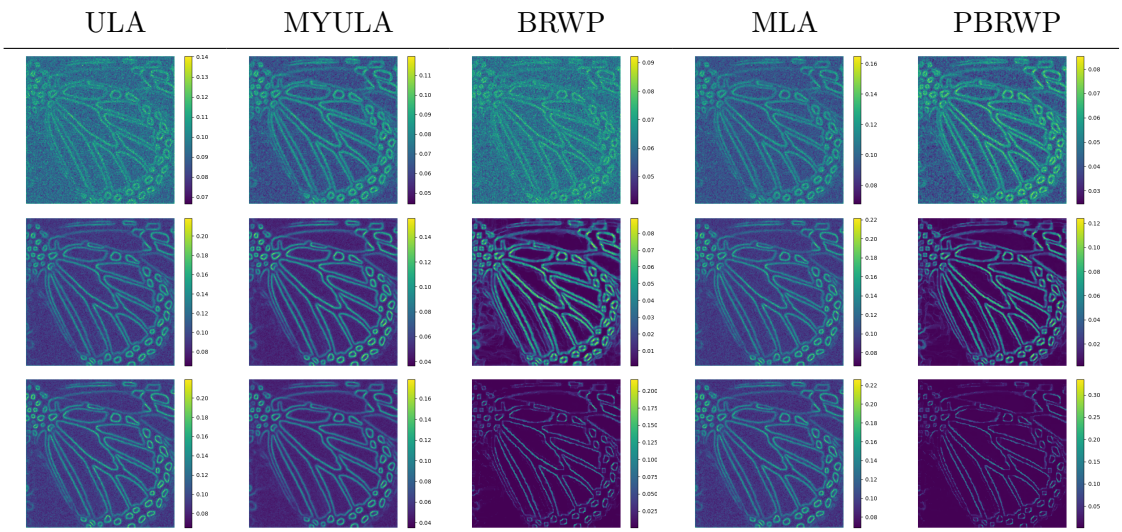


Figure 10. *Standard deviations for TV-regularized deconvolution. Run with 40 particles, and evaluated at iterations 20, 200 and 2000 in the top, middle and bottom rows respectively. We observe that the pixelwise variance of the noise-free methods BRWP and PBRWP are lower than their Langevin counterparts. In the 200 iteration regime, we see that the features of the noise-free BRWP and PBRWP methods have more constrast and lower variances in the piecewise constant areas.*

to the true TV-regularized solution, equivalent to running a fixed number of parallel chains for the Langevin methods. We do this to compare the small-particle behavior of the noise-free methods versus the Langevin methods. The minimizer of the objective (6.7) is computed using subgradient descent for 30,000 iterations, applied with decreasing step-sizes $1\mathrm{e}-\{4,5,6\}$ for 10,000 iterations each. The image considered is the butterfly image from the set3c dataset, which has a dimension of $3 \times 256 \times 256$.

We compare against ULA, MLA, and MYULA [11]. We note that the acceptance probability for MALA is almost degenerate for this high-dimensional problem and requires extremely small step-sizes, and is thus omitted due to slow convergence. MYULA takes the following form: for a composite potential of the form $\exp(-f(x)+g(x))$, a step-size $\tau > 0$ and a regularization parameter $\theta > 0$,

$$\text{(MYULA)} \qquad X_{k+1} = \left(1 - \frac{\tau}{\theta}\right) X_k - \tau \nabla f(X_k) + \frac{\tau}{\theta} \operatorname{prox}_{\theta g}(X_k) + \sqrt{2\tau} Z_{k+1},$$

where Z_{k+1} are independent standard Gaussians. This is the unadjusted Langevin algorithm, applied to the regularized potential where g is replaced by its Moreau–Yosida regularization with parameter θ. Applied to the deconvolution potential (6.7), we identify $f(x) = \frac{1}{2\sigma^2}\|Ax-y\|^2$ and $g(x) = \lambda\,\mathrm{TV}(x)$. The proximal of g is computed using the Chambolle–Pock algorithm [9] using the implementation in the DeepInverse Python library [43], and the parameter is fixed as $\theta = 1\mathrm{e}-4$.

The step-sizes for the baseline methods are chosen with a grid-search such that the PSNR is maximized after 2000 iterations, and can be found in the appendix. We compare with 40 particles due to GPU memory limitations from the $\mathcal{O}(N^2)$ scaling of fully parallelized

Table 1

PSNR (in dB) of the mean of 40 particles compared to the MAP estimate for the deconvolution problem. We observe that the preconditioned methods converge to the neighborhood of the MAP faster than their non-preconditioned counterparts, indicated by the values at 20 and 200 iterations. Moreover, the noise-free methods are much more tightly concentrated around the MAP estimate. However, the particle means of the preconditioned methods are slightly further away from the posterior mean at high iterations.

Iteration	ULA	MYULA	BRWP	MLA	PBRWP
20	23.28	23.53	23.52	23.86	24.27
200	25.43	25.79	27.81	25.78	29.90
2000	26.07	26.38	38.38	26.00	37.35

PBRWP[3].

The experimental setting differs from typical settings as follows, to highlight the difference between Langevin methods and the noise-free methods BRWP and PBRWP:

1. *Low sample count.* Langevin algorithms usually run many iterations (around 1e5) of the Markov chain to construct an estimate of the posterior mean. In our case, since the samples are approximately convergent for high iterations, taking means over iterations is not insightful, and we replace it with the mean over particles. The particles for Langevin methods can be computed using parallel chains.
2. *Short "burn-in" phase.* While typical Bayesian imaging methods run burn-in periods for up to 1e6 iterations, the noise-free methods again do not benefit from these long iterations. As our objective is rapid convergence, we consider only methods that are able to converge in the short iteration regime.
3. *Comparison with MAP estimate.* In high dimensions, computing the posterior mean is intractable, and is usually estimated. While it is possible to consider the convergence of the algorithm to the posterior mean by applying a Metropolis–Hastings correction step [16], this would compare the algorithms against an estimator with a-priori unknown uncertainty. We choose the MAP estimate instead as it is computable up to arbitrary accuracy for this convex task, and to compare the mode concentration of the algorithms.

Table 1 plots the peak signal-to-noise ratio (PSNR) of the particle means, compared to the MAP estimate (6.7) (equivalently, the minimizer of the potential). We observe that both BRWP and PBRWP converge to much higher values of PSNR, indicating better concentration around the MAP. Moreover, PBRWP has higher PSNR at lower iterations, indicating faster convergence. The lower PSNR of the preconditioned methods compared to the non-preconditioned methods can be explained by the differing noise and diffusion scales. Figure 10 plots the pixel-wise standard deviation of the particles at various iteration counts. We observe again that PBRWP has higher contrast compared to BRWP at low iterations, indicating faster convergence. Moreover, the noise-free methods have very low noise outside the edges. This low-rank phenomenon arises since diffusion comes from inter-particle interactions.

[3]It is possible to remove some parallelization by computing the interaction matrix column-wise, yielding $\mathcal{O}(N)$ memory constraint.

6.7. Bayesian neural networks. In this section, we empirically investigate the possibility of a variable preconditioning matrix M, applied to a high-dimensional non-convex potential. Recall the PBRWP iteration (4.8) for particles $\{\mathbf{x}_i\}_{i=1}^N$:

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} - \frac{\eta}{2} M \nabla V(\mathbf{X}^{(k)}) + \frac{\eta}{2T}\left(\mathbf{X}^{(k)} - \mathbf{X}^{(k)}\mathrm{softmax}(W^{(k)})^\top\right),$$
$$W_{i,j}^{(k)} = -\beta \frac{\|\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)}\|_M^2}{4T} - \log \mathcal{Z}(\mathbf{x}_j^{(k)}).$$

In the normalizing term, we empirically propose to replace the term $\|\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)}\|_M^2$ with a j- and k-dependent scaling $\|\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)}\|_{M_j^{(k)}}^2$. Using Laplace's approximation, we have

$$\mathcal{Z}(\mathbf{x}_j; M_j) = \int_{\mathbb{R}^d} \exp\left(-\frac{\beta}{2}(V(z) + \frac{\|z - \mathbf{x}_j\|_{M_j}^2}{2T})\right) \mathrm{d}z \tag{6.8}$$
$$\approx C(\beta, T)\sqrt{\det M_j}\exp\left(-\frac{\beta}{2}V(\mathbf{x}_j)\right). \tag{6.9}$$

Simplifying, the variable-preconditioner version of PBRWP takes the following form:

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} - \frac{\eta}{2}\mathbf{M}^{(k)}\nabla V(\mathbf{X}^{(k)}) + \frac{\eta}{2T}\left(\mathbf{X}^{(k)} - \mathbf{X}^{(k)}\mathrm{softmax}(W^{(k)})^\top\right), \tag{6.10}$$
$$W_{i,j}^{(k)} = -\beta\frac{\|\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)}\|_{M_j^{(k)}}^2}{4T} - \frac{1}{2}\log\det\left(M_j^{(k)}\right) + \frac{\beta}{2}V(\mathbf{x}_j), \tag{6.11}$$

where $\mathbf{M}^{(k)}\nabla V(\mathbf{X}^{(k)}) \coloneqq \left[M_1^{(k)}\nabla V(\mathbf{x}_1) \quad \ldots \quad M_N^{(k)}\nabla V(\mathbf{x}_N)\right]$.

For a neural network, the choice of preconditioner is a-priori unclear. For our experiments, we choose the diagonal preconditioner corresponding to the Adam optimizer [24], with exact implementation given in E. This has recently been linked to a diagonal approximation to the empirical Fisher information matrix [27, 22]. The diagonal form of the preconditioning matrix in this case makes it particularly amenable for the log-determinant in (6.11).

To test, we train neural networks on various regression datasets, with $N = 10$ particles as in [45, 50]. The potential V is given by the mean-squared error on the training dataset. Table 2 shows the root-mean-square error of the trained networks using PBRWP against various baselines. For the Adam baseline [24], we directly train from the initializations, where the step-size is chosen with a grid search to minimize the test RMSE. We observe that PBRWP is able to uniformly outperform BRWP on such a task, and is more competitive with kernel-based methods. This is analogous to Adam being empirically better than SGD for training neural networks.

7. Conclusion. This work proposes the Preconditioned Backwards Regularized Wasserstein Proximal (PBRWP) method Algorithm 1, a method of sampling without Langevin noise. Deriving the Cole–Hopf transform from a pair of anisotropic forward-backward heat equations, we obtain a coupled forward-time Fokker–Planck equation and backward-time Hamilton–Jacobi equation, both regularized with (elliptic) second-order differential operators. By applying a

Table 2

Test root-mean-square-error (RMSE) on test datasets on various Bayesian neural network tasks. Bold indicates smallest in row. We observe that the adaptive Fisher preconditioned BRWP uniformly outperforms BRWP on each of the BNN tasks. Adam and the noise-free methods both generally exhibit high variance in this setting, which may be due to the relatively small neural network architecture and sensitivity to initialization. We may further interpret the high variance of these methods as being able to find better trained models.

Dataset	Adam	PBRWP	BRWP	AIG	WGF	SVGD
Boston	$3.350_{\pm 8.33e-1}$	$2.866_{\pm 5.94e-1}$	$3.309_{\pm 5.31e-1}$	$2.871_{\pm 3.41e-3}$	$3.077_{\pm 5.52e-3}$	$\mathbf{2.775_{\pm 3.78e-3}}$
Combined	$3.971_{\pm 1.79e-1}$	$\mathbf{3.925_{\pm 1.52e-1}}$	$3.975_{\pm 3.94e-2}$	$4.067_{\pm 9.27e-1}$	$4.077_{\pm 3.85e-4}$	$4.070_{\pm 2.02e-4}$
Concrete	$4.698_{\pm 4.85e-1}$	$\mathbf{4.387_{\pm 4.88e-1}}$	$4.478_{\pm 2.05e-1}$	$4.440_{\pm 1.34e-1}$	$4.883_{\pm 1.93e-1}$	$4.888_{\pm 1.39e-1}$
Kin8nm	$0.089_{\pm 2.72e-3}$	$\mathbf{0.087_{\pm 2.67e-3}}$	$0.089_{\pm 6.06e-6}$	$0.094_{\pm 5.56e-6}$	$0.096_{\pm 3.36e-5}$	$0.095_{\pm 1.32e-5}$
Wine	$0.629_{\pm 4.01e-2}$	$0.612_{\pm 4.17e-2}$	$0.623_{\pm 1.35e-3}$	$0.606_{\pm 1.40e-5}$	$0.614_{\pm 3.48e-4}$	$\mathbf{0.604_{\pm 9.89e-5}}$

similar semi-implicit discretization to [45], we obtain particle swarm dynamics, interpretable as a self-attention mechanism. A discrete-time non-asymptotic convergence analysis is given in the case of quadratic potentials. Experiments demonstrate faster convergence with slightly more empirical bias, consistent with preconditioned Langevin methods.

While a fixed preconditioner M has be shown to provide acceleration for some simple variational problems, the experiments in Subsection 6.7 empirically demonstrate that it is also possible to have variable preconditioners, such as that approximating the empirical Fisher information matrix. Future work would consider rigorous justification of variable preconditioners $M = M(x)$, such as those arising from inhomogeneous heat equations, or from Bregman divergences, similarly to mirror descent methods. Given that the PBRWP may be considered as a self-attention framework, future work may consider the converse implication. Additional theoretical analysis is also important. One part is the continuous-time convergence of the modified Fokker–Planck type equation (4.9), which is not amenable to standard energy methods due to the variable kernel. The other is the discretization error from working with finitely many particles. While this arises from central limit and ergodicity results for standard Langevin methods, the final iteration approximation to the desired measure is less clear in the noiseless setting due to the per-particle convergence.

REFERENCES

[1] M. V. Afonso, J. M. Bioucas-Dias, and M. A. Figueiredo, *Fast image recovery using variable splitting and constrained optimization*, IEEE transactions on image processing, 19 (2010), pp. 2345–2356.

[2] J.-D. Benamou and Y. Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, Numerische Mathematik, 84 (2000), pp. 375–393.

[3] N. Bleistein and R. A. Handelsman, *Asymptotic expansions of integrals*, Ardent Media, 1975.

[4] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, *Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models*, IEEE transactions on pattern analysis and machine intelligence, (2021).

[5] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, *Kernel density estimation via diffusion*, Annals of Statistics, 38 (2010), pp. 2916–2957.

[6] J. A. Carrillo, K. Craig, and F. S. Patacchini, *A blob method for diffusion*, Calculus of Variations and Partial Differential Equations, 58 (2019), pp. 1–53.

[7] V. Castin, P. Ablin, J. A. Carrillo, and G. Peyré, *A unified perspective on the dynamics of deep*

transformers, arXiv preprint arXiv:2501.18322, (2025).
[8] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, *Neural ordinary differential equations*, Advances in neural information processing systems, 31 (2018).
[9] L. Condat, *A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms*, Journal of optimization theory and applications, 158 (2013), pp. 460–479.
[10] A. Durmus and É. Moulines, *High-dimensional Bayesian inference via the unadjusted Langevin algorithm*, Bernoulli, 25 (2019), pp. 2854–2882.
[11] A. Durmus, E. Moulines, and M. Pereyra, *Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau*, SIAM Journal on Imaging Sciences, 11 (2018), pp. 473–506.
[12] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, Chapman and Hall/CRC, 1995.
[13] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, *A mathematical perspective on transformers*, Bulletin of the American Mathematical Society, 62 (2025), pp. 427–479.
[14] A. Gramacki, *Nonparametric kernel density estimation and its computational aspects*, vol. 37, Springer, 2018.
[15] A. Habring, A. Falk, and T. Pock, *Diffusion at absolute zero: Langevin sampling using successive Moreau envelopes*, in 2025 IEEE Statistical Signal Processing Workshop (SSP), IEEE, 2025, pp. 61–65.
[16] A. Habring, M. Holler, and T. Pock, *Subgradient langevin methods for sampling from nonsmooth potentials*, SIAM Journal on Mathematics of Data Science, 6 (2024), pp. 897–925.
[17] F. Han, S. Osher, and W. Li, *Convergence of noise-free sampling algorithms with regularized Wasserstein proximals*, arXiv preprint arXiv:2409.01567, (2024).
[18] F. Han, S. Osher, and W. Li, *Splitting regularized Wasserstein proximal algorithms for nonsmooth sampling problems*, arXiv preprint arXiv:2502.16773, (2025).
[19] F. Han, S. Osher, and W. Li, *Tensor train based sampling algorithms for approximating regularized Wasserstein proximal operators*, SIAM/ASA Journal on Uncertainty Quantification, 13 (2025), pp. 775–804.
[20] J. R. Hershey and P. A. Olsen, *Approximating the Kullback Leibler divergence between Gaussian mixture models*, in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, vol. 4, IEEE, 2007, pp. IV–317.
[21] Y.-P. Hsieh, A. Kavis, P. Rolland, and V. Cevher, *Mirrored Langevin dynamics*, Advances in Neural Information Processing Systems, 31 (2018).
[22] D. Hwang, *Fadam: Adam is a natural gradient optimizer using diagonal empirical Fisher information*, arXiv preprint arXiv:2405.12807, (2024).
[23] Q. Jiang, *Mirror Langevin Monte Carlo: the case under isoperimetry*, Advances in Neural Information Processing Systems, 34 (2021), pp. 715–725.
[24] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980, (2014).
[25] W. Krauth, *Statistical mechanics: algorithms and computations*, vol. 13, OUP Oxford, 2006.
[26] R. Kubo, *Stochastic Liouville equations*, Journal of Mathematical Physics, 4 (1963), pp. 174–183.
[27] F. Kunstner, P. Hennig, and L. Balles, *Limitations of the empirical fisher approximation for natural gradient descent*, Advances in neural information processing systems, 32 (2019).
[28] R. Laumont, V. D. Bortoli, A. Almansa, J. Delon, A. Durmus, and M. Pereyra, *Bayesian imaging using plug & play priors: when Langevin meets Tweedie*, SIAM Journal on Imaging Sciences, 15 (2022), pp. 701–737.
[29] W. Li, S. Liu, and S. Osher, *A kernel formula for regularized Wasserstein proximal operators*, Research in the Mathematical Sciences, 10 (2023), p. 43.
[30] W. Li, W. Liu, J. Chen, L. Wu, P. D. Flynn, W. Ding, and P. Chen, *Reducing mode collapse with Monge–Kantorovich optimal transport for generative adversarial networks*, IEEE Transactions on Cybernetics, (2023).
[31] D. J. MacKay, *Bayesian neural networks and density networks*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 354 (1995), pp. 73–80.
[32] E. Nijkamp, R. Gao, P. Sountsov, S. Vasudevan, B. Pang, S.-C. Zhu, and Y. N. Wu, *MCMC*

should mix: learning energy-based model with neural transport latent space MCMC., in International Conference on Learning Representations (ICLR 2022)., 2022.

[33] K. B. Petersen, M. S. Pedersen, et al., *The matrix cookbook*, Technical University of Denmark, 7 (2008), p. 510.

[34] H. Risken, *Fokker-Planck equation*, in The Fokker-Planck equation: methods of solution and applications, Springer, 1989, pp. 63–95.

[35] G. O. Roberts and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, (1996), pp. 341–363.

[36] P. J. Rossky, J. D. Doll, and H. L. Friedman, *Brownian dynamics as smart Monte Carlo simulation*, The Journal of Chemical Physics, 69 (1978), pp. 4628–4633.

[37] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: nonlinear phenomena, 60 (1992), pp. 259–268.

[38] M. E. Sander, P. Ablin, M. Blondel, and G. Peyré, *Sinkformers: Transformers with doubly stochastic attention*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 3515–3530.

[39] F. Santambrogio, *Optimal transport for applied mathematicians*, vol. 87, Springer, 2015.

[40] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, *Score-based generative modeling through stochastic differential equations*, in International Conference on Learning Representations, 2021.

[41] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, *Veegan: Reducing mode collapse in GANs using implicit variational learning*, Advances in neural information processing systems, 30 (2017).

[42] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta numerica, 19 (2010), pp. 451–559.

[43] J. Tachella, M. Terris, S. Hurault, A. Wang, D. Chen, M.-H. Nguyen, M. Song, T. Davies, L. Davy, J. Dong, et al., *Deepinverse: A python package for solving imaging inverse problems with deep learning*, arXiv preprint arXiv:2505.20160, (2025).

[44] H. Y. Tan, Z. Cai, M. Pereyra, S. Mukherjee, J. Tang, and C.-B. Schönlieb, *Unsupervised training of convex regularizers using maximum likelihood estimation*, Transactions on Machine Learning Research, (2024).

[45] H. Y. Tan, S. Osher, and W. Li, *Noise-free sampling algorithms via regularized Wasserstein proximals*, Research in the Mathematical Sciences, 11 (2024), p. 65.

[46] R. J. Tibshirani, S. W. Fung, H. Heaton, and S. Osher, *Laplace meets Moreau: Smooth approximation to infimal convolutions using Laplace's method*, Journal of Machine Learning Research, 26 (2025), pp. 1–36.

[47] P. Van Kerm, *Adaptive kernel density estimation*, The Stata Journal, 3 (2003), pp. 148–156.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention is all you need*, Advances in neural information processing systems, 30 (2017).

[49] M. P. Wand and M. C. Jones, *Kernel smoothing*, CRC press, 1994.

[50] Y. Wang and W. Li, *Accelerated information gradient flow*, Journal of Scientific Computing, 90 (2022), pp. 1–47.

Appendix A. Proofs.

A.1. Anisotropic Green's function.

Recall (3.1):

$$G_{t,M}(x,y) \coloneqq \frac{1}{(4\pi\beta^{-1}t)^{d/2}|M|^{1/2}} e^{-\beta\frac{(x-y)^\top M^{-1}(x-y)}{4t}}. \tag{A.1}$$

Proposition A.1. *The anisotropic kernel $G_{t,M}(x,y)$ is a Green's function for the following PDE:*

$$\begin{cases} \partial_t u - \beta^{-1}\nabla\cdot(M\nabla u) = 0, \\ u(0,x) = \delta(y). \end{cases}$$

Proof.

$$\partial_t G_{t,M}(x,y) = \left(-\frac{d}{2t} + \beta\frac{(x-y)^\top M^{-1}(x-y)}{4t^2}\right) G_{t,M}(x,y)$$

$$\nabla_x G_{t,M}(x,y) = -\frac{\beta M^{-1}(x-y)}{2t} G_{t,M}(x,y)$$

$$\begin{aligned} \nabla_x\cdot(MG_{t,m}(x,y)) &= \nabla_x\cdot\left(-\frac{\beta(x-y)}{2t}G_{t,M}(x,y)\right) \\ &= \left(-\frac{d\beta}{2t} + \beta^2\frac{(x-y)^\top M^{-1}(x-y)}{4t^2}\right) G_{t,M}(x,y) \end{aligned}$$

Therefore $\partial_t G_{t,M} = \beta^{-1}\nabla\cdot(M\nabla G_{t,m})$. The boundary condition follows from a change of variables with the standard heat kernel. ∎

Appendix B. Proofs for Gaussian analysis. Recall that we wish to sample from some distribution $\exp(-\beta V(x))$, as in Section 5. Here we focus on the target distribution $\mathcal{N}(0,\beta^{-1}\Sigma)$, corresponding to potential $V(x) = x^\top\Sigma^{-1}x/2$. For a fixed preconditioning matrix $M \succ 0$, we will find it convenient to work in a rotated basis. For any covariance matrix Σ, we define a corresponding (positive definite) Ξ as $\Xi = \sqrt{M}^{-1}\Sigma\sqrt{M}^{-1}$.

Recall from (3.11) the normalized kernel

$$K_M(x,y) = \frac{\exp\left(-\frac{\beta}{2}(V(x) + \frac{\|x-y\|_M^2}{2T})\right)}{\int_{\mathbb{R}^d}\exp\left(-\frac{\beta}{2}(V(z) + \frac{\|z-y\|_M^2}{2T})\right)\mathrm{d}z}. \tag{B.1}$$

We may compute for a fixed y, noting $K_M(x,y)$ is a Gaussian density in x,

$$\begin{aligned} K_M(x,y) &\propto \exp\left(-\frac{\beta}{2}\left(\frac{x^\top \Sigma^{-1} x}{2} + \frac{(x-y)^\top M^{-1}(x-y)}{2T}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left[x^\top\left(\frac{\beta\Sigma^{-1}}{2} + \frac{\beta M^{-1}}{2T}\right)x - y^\top \frac{\beta M^{-1}}{2T}x - x^\top \frac{\beta M^{-1}}{2T}y + ...\right]\right) \\ &\propto \exp\Bigg(-\frac{1}{2}\Bigg[\left(x - \left(\frac{\Sigma^{-1}}{2} + \frac{M^{-1}}{2T}\right)^{-1}\frac{M^{-1}}{2T}y\right)^\top \left(\frac{\beta\Sigma^{-1}}{2} + \frac{\beta M^{-1}}{2T}\right) \\ &\qquad \left(x - \left(\frac{\Sigma^{-1}}{2} + \frac{M^{-1}}{2T}\right)^{-1}\frac{M^{-1}}{2T}y\right)\Bigg]\Bigg) \end{aligned}$$

We therefore have that

$$K_M(\cdot,y) \sim \mathcal{N}\left(\left(\frac{\Sigma^{-1}}{2} + \frac{M^{-1}}{2T}\right)^{-1}\frac{M^{-1}}{2T}y, \left(\frac{\beta\Sigma^{-1}}{2} + \frac{\beta M^{-1}}{2T}\right)^{-1}\right). \tag{B.2}$$

Utilizing that $\int K_M(x,y)\,\mathrm{d}x = 1$, the kernel can be written explicitly as

$$\begin{aligned} K_M(x,y) = \frac{1}{B}\exp\Bigg(-\frac{1}{2}\Bigg[&\left(x - \left(\frac{\Sigma^{-1}}{2} + \frac{M^{-1}}{2T}\right)^{-1}\frac{M^{-1}}{2T}y\right)^\top \left(\frac{\beta\Sigma^{-1}}{2} + \frac{\beta M^{-1}}{2T}\right) \\ &\left(x - \left(\frac{\Sigma^{-1}}{2} + \frac{M^{-1}}{2T}\right)^{-1}\frac{M^{-1}}{2T}y\right)\Bigg]\Bigg), \end{aligned} \tag{B.3}$$

$$B = (2\pi)^{d/2}\det\left(\left(\frac{\beta\Sigma^{-1}}{2} + \frac{\beta M^{-1}}{2T}\right)^{-1}\right)^{1/2}. \tag{B.4}$$

B.1. Additional definitions. Consider the scaled Fisher information

$$\mathcal{I}_\pi^M(\rho) = \int \langle \nabla \log\frac{\rho(x)}{\pi(x)}, M\nabla\log\frac{\rho(x)}{\pi(x)}\rangle \rho(x)\,\mathrm{d}x\,. \tag{B.5}$$

Supposing that the terminal distribution π satisfies the scaled log-Sobolev inequality, stating that for every locally Lipschitz function g, that

$$\frac{2}{c}\int \|\nabla g(x)\|^2_{M^{-1}}\mathrm{d}\pi \geq \int g^2\log g^2\mathrm{d}\pi - \left(\int g^2\mathrm{d}\pi\right)\log\left(\int g^2\mathrm{d}\pi\right), \tag{B.6}$$

applying with $g(x) = \sqrt{\mathrm{d}\rho/\mathrm{d}\pi}$ yields for any appropriate density

$$\mathrm{D}_{\mathrm{KL}}(\rho\|\pi) = \int \rho(x)\log\frac{\rho(x)}{\pi(x)}\mathrm{d}x \leq \frac{1}{2c}\mathcal{I}_\pi^M(\rho). \tag{B.7}$$

We note that by the Bakry–Emery criterion, a distribution $\exp(-U(x))$ satisfies the scaled log-Sobolev inequality with parameter κ if $\nabla^2 U \succeq \kappa M^{-1}$.

From [23, Prop. 1], if π satisfies the scaled log-Sobolev inequality, we have exponential convergence $\mathrm{D}_{\mathrm{KL}}(\rho_t||\pi) \leq \exp(-2\kappa t)\mathrm{D}_{\mathrm{KL}}(\rho_0\|\pi)$. Moreover, we have the following expression, which may be derived by direct computation,

$$\begin{aligned}\mathcal{I}^M_{\mathcal{N}(0,\Sigma_\infty)}(\mathcal{N}(0,\Sigma_t)) &= \mathrm{Tr}\big(\Sigma_t(\Sigma_\infty^{-1}-\Sigma_t^{-1})M(\Sigma_\infty^{-1}-\Sigma_t^{-1})\big)\\ &= \mathrm{Tr}\big(\Xi_t(\Xi_\infty^{-1}-\Xi_t^{-1})^2\big),\end{aligned}$$

where we denote $\Xi_t = \sqrt{M}^{-1}\Sigma_t\sqrt{M}^{-1}$, $\Xi_\infty = \sqrt{M}^{-1}\Sigma_\infty\sqrt{M}^{-1}$.

B.2. Regularized Wasserstein proximal of Gaussian.

Proposition B.1. *For a distribution $\rho_k \sim \mathcal{N}(\mu_k,\Sigma_k)$, the Wasserstein proximal satisfies* $\mathrm{WProx}^M_{T,V}\,\rho_k \sim \mathcal{N}(\tilde\mu_k,\tilde\Sigma_k)$, *where*

$$\tilde\mu_k = (I+TM\Sigma^{-1})^{-1}\mu_k, \tag{B.8}$$

$$\tilde\Sigma_k = 2\beta^{-1}T\left(T\Sigma^{-1}+M^{-1}\right)^{-1} + \left(T\Sigma^{-1}+M^{-1}\right)^{-1}M^{-1}\Sigma_k M^{-1}\left(T\Sigma^{-1}+M^{-1}\right)^{-1}. \tag{B.9}$$

Proof. We compute the closed-form update from a distribution $\rho_k \sim \mathcal{N}(\mu_k,\Sigma_k)$ from the update equation

$$\mathrm{WProx}^M_{T,V}\,\rho_k = \tilde\rho_k \propto \int_{\mathbb{R}^d} K_M(x,y)\rho_k(y)\,\mathrm{d}y\,, \tag{B.10a}$$

$$K_M(x,y) = \frac{\exp\Big(-\frac{\beta}{2}(V(x)+\frac{\|x-y\|_M^2}{2T})\Big)}{\int_{\mathbb{R}^d}\exp\Big(-\frac{\beta}{2}(V(z)+\frac{\|z-y\|_M^2}{2T})\Big)\,\mathrm{d}z}. \tag{B.10b}$$

We have from (B.3) that

$$K_M(x,\cdot) \propto \mathcal{N}((I+TM\Sigma^{-1})x, 2\beta^{-1}TM(T\Sigma^{-1}+M^{-1})M). \tag{B.11}$$

From [33], we have that the product of two Gaussian densities is again a Gaussian density, multiplied by a desired normalizing constant that is given by

$$\mathcal{N}_{(I+TM\Sigma^{-1})x}(\mu_k, \Sigma_k + 2\beta^{-1}TM(T\Sigma^{-1}+M^{-1})M). \tag{B.12}$$

We abuse the subscript notation to denote the Gaussian density evaluated at the subscript. The integral of the Gaussian density appears in both the numerator and denominator and is removed from the expression. We thus have that

$$\begin{aligned}\tilde\rho_k &\propto \int_{\mathbb{R}^d} K_M(x,y)\rho_k(y)\,\mathrm{d}y\\ &\propto \mathcal{N}_{(I+TM\Sigma^{-1})x}(\mu_k,\Sigma_k+2\beta^{-1}TM(T\Sigma^{-1}+M^{-1})M)\\ &\propto \mathcal{N}_x((I+TM\Sigma^{-1})^{-1}\mu_k,\\ &\qquad (I+TM\Sigma^{-1})^{-1}[\Sigma_k+2\beta^{-1}TM(T\Sigma^{-1}+M^{-1})M](I+TM\Sigma^{-1})^{-\top}).\end{aligned}$$

Manipulating yields the desired expressions for $\tilde\mu_k$ and $\tilde\Sigma_k$. ∎

B.3. Proof of discrete time convergence for Gaussians. Here we prove Theorem 5.3. We restate it here for convenience.

Theorem B.2. *Suppose $V(x) = x^\top \Sigma^{-1} x/2$, and fix a $\beta > 0$. For the target distribution $\pi = \mathcal{N}(0, \beta^{-1}\Sigma)$ and positive definite preconditioner M, let $c, C > 0$ be such that $cM \preceq \Sigma \preceq CM$. Let $T \in (0, c)$ so that the inverse PRWPO is well-defined. Then there exists a stationary distribution to the discrete-time PBRWP iterations, and it is Gaussian $\hat{\pi} = \mathcal{N}(0, \hat{\Sigma})$, satisfying $\mathrm{WProx}^M_{T,V}(\mathcal{N}(0, \hat{\Sigma})) = \mathcal{N}(0, \beta^{-1}\Sigma)$. Moreover, it is unique within distributions with sufficiently decaying Fourier transforms, specifically as given in Subsection C.1.*

Suppose further that the initial Gaussian distribution $\rho_0 = \mathcal{N}(0, \Sigma_0)$ has covariance commuting with Σ, let ρ_k be the iterations generated by PBRWP, and let $\tilde{\rho}_k = \mathrm{WProx}^M_{T,V}(\rho_k)$. Let λ be such that $\sqrt{M}^{-1}\Sigma_k\sqrt{M}^{-1}$ are uniformly bounded below by some λ (which holds as $\Sigma_k \to \Sigma$). Let the step-size η be bounded by

$$\eta\beta^{-1} \le \min((\max_i |\lambda_i(\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})|)^{-1}/2, 3\lambda_{\min}(\tilde{\Xi}_k)/32)$$

then the exact discrete-time PBRWP method has the following decay in KL divergence:

$$\mathrm{D_{KL}}(\tilde{\rho}_{k+1}\|\pi) - \mathrm{D_{KL}}(\tilde{\rho}_k\|\pi) \le -\frac{\eta}{2C[\beta + 2T(1+TC^{-1})^{-1}(1+Tc^{-1})^2\lambda^{-1}]}\mathrm{D_{KL}}(\tilde{\rho}_k\|\pi). \tag{B.13}$$

Proof. We show convergence rates and existence first. Let us define $\tilde{\Sigma}_\infty = \beta^{-1}\Sigma$, motivated by the fact that the PRWPO of the stationary covariance is the target covariance. Further recall that $\tilde{\Sigma}_k$ is the covariance of the PRWPO of $\mathcal{N}(0, \Sigma_k)$, in closed form in Proposition 5.1. Recall the discrete covariance update for PBRWP from Corollary 5.2:

$$\Sigma_{k+1} = (I - \eta M\Sigma^{-1} + \eta\beta^{-1}M\tilde{\Sigma}_k^{-1})\Sigma_k(I - \eta M\Sigma^{-1} + \eta\beta^{-1}M\tilde{\Sigma}_k^{-1})^\top. \tag{B.14}$$

In terms of the M-conjugated matrices $\Xi_k = \sqrt{M}^{-1}\Sigma_k\sqrt{M}^{-1}$, $\tilde{\Xi}_k = \sqrt{M}^{-1}\tilde{\Sigma}_k\sqrt{M}^{-1}$, this becomes

$$\Xi_{k+1} = (I - \eta\beta^{-1}\tilde{\Xi}_\infty^{-1} + \eta\beta^{-1}\tilde{\Xi}_k^{-1})\Xi_k(I - \eta\beta^{-1}\tilde{\Xi}_\infty^{-1} + \eta\beta^{-1}\tilde{\Xi}_k^{-1}). \tag{B.15}$$

We define the following auxiliary matrix expressions for convenience.

$$K_+ \coloneqq I + T\sqrt{M}\Sigma^{-1}\sqrt{M}, \quad K_- \coloneqq I - T\sqrt{M}\Sigma^{-1}\sqrt{M}. \tag{B.16}$$

Then, the (conjugated) covariance of the Wasserstein proximal $\tilde{\Xi}_k = \sqrt{M}^{-1}\tilde{\Sigma}_k\sqrt{M}^{-1}$, $\tilde{\Sigma}_k = \mathrm{Cov}(\mathrm{WProx}^M_{T,V}\mathcal{N}(0, \Sigma_k))$ has the following iterations

$$\tilde{\Xi}_k = 2\beta^{-1}TK_+^{-1} + K_+^{-1}\Xi_k K_+^{-1}. \tag{B.17}$$

At the stationary point, we can compute the inverse PRWPO as

$$\Xi_\infty = K_-\tilde{\Xi}_\infty K_+ = K_+\tilde{\Xi}_\infty K_-, \quad \tilde{\Xi}_\infty = \beta^{-1}\sqrt{M}^{-1}\Sigma\sqrt{M}^{-1}. \tag{B.18}$$

They satisfy the following:

$$\tilde{\Xi}_k - \tilde{\Xi}_\infty = K_+^{-1}(\Xi_k - \Xi_\infty)K_+^{-1}, \tag{B.19}$$

$$\tilde{\Xi}_k = K_+^{-1}(\Xi_k + 2\beta^{-1}TK_+)K_+^{-1}. \tag{B.20}$$

The update of the $\tilde{\Xi}$ can be written as follows, using (B.15) and (B.20),

$$\begin{aligned}\tilde{\Xi}_{k+1} &= K_+^{-1}(\Xi_{k+1} + 2\beta^{-1}TK_+)K_+^{-1} \\ &= K_+^{-1}\left[(I - \eta\beta^{-1}\tilde{\Xi}_\infty^{-1} + \eta\beta^{-1}\tilde{\Xi}_k^{-1})\Xi_k(I - \eta\beta^{-1}\tilde{\Xi}_\infty^{-1} + \eta\beta^{-1}\tilde{\Xi}_k^{-1}) + 2\beta^{-1}TK_+\right]K_+^{-1}.\end{aligned} \tag{B.21}$$

Claim. We will show that $\tilde{\Xi}_k \to \tilde{\Xi}_\infty$, and accordingly that $\Xi_k \to \Xi_\infty$.

By the assumption that Σ_0 commutes with Σ, we have that Ξ_k, $\tilde{\Xi}_k$, $K_\pm$, $\tilde{\Xi}_\infty^{-1}$ all commute. We recall a useful Taylor expansion first.

Proposition B.3. *Let A be square with eigenvalues $\lambda_i \geq 0$. Then*

$$\log\det(I + A) = \mathrm{Tr}(A) + R \tag{B.22}$$

where

$$0 \geq R = \sum_i (\log(\lambda_i + 1) - \lambda_i) \geq -\sum \frac{\lambda_i^2}{1 + \lambda_i} \geq -\sum_i \lambda_i^2. \tag{B.23}$$

The inequalities other than the last one also hold for $\lambda_i > -1$.

Proof. Using $\det(A) = \prod_i \lambda_i$ and $\mathrm{Tr}(A) = \sum_i \lambda_i$. ∎

Recall that for Gaussians, the KL divergence to $\pi \sim \mathcal{N}(0, \beta^{-1}\Sigma) = \mathcal{N}(0, \tilde{\Sigma}_\infty)$ takes the particular form [20]:

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(\tilde{\rho}_k \| \pi) &= \int \tilde{\rho}_k(x) \log \frac{\tilde{\rho}_k(x)}{\pi(x)}\,\mathrm{d}x \\ &= \frac{1}{2}\left[\log\det\tilde{\Sigma}_\infty - \log\det\tilde{\Sigma}_k - d + \mathrm{Tr}\left(\tilde{\Sigma}_\infty^{-1}\tilde{\Sigma}_k\right)\right].\end{aligned}$$

The difference between two consecutive iterations is

$$\begin{aligned}&\mathrm{D}_{\mathrm{KL}}(\tilde{\rho}_{k+1}\|\pi) - \mathrm{D}_{\mathrm{KL}}(\tilde{\rho}_k\|\pi) \\ &= \frac{1}{2}\left[\log\frac{|\tilde{\Sigma}_k|}{|\tilde{\Sigma}_{k+1}|} + \mathrm{Tr}\left(\tilde{\Sigma}_\infty^{-1}(\tilde{\Sigma}_{k+1} - \tilde{\Sigma}_k)\right)\right] \\ &= \frac{1}{2}\left[\log\frac{|\tilde{\Xi}_k|}{|\tilde{\Xi}_{k+1}|} + \mathrm{Tr}\left(\tilde{\Xi}_\infty^{-1}(\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\right)\right].\end{aligned} \tag{B.24}$$

Applying the Taylor expansion to the first term,

$$\begin{aligned}\log\frac{|\tilde{\Xi}_k|}{|\tilde{\Xi}_{k+1}|} &= \log\frac{|\tilde{\Xi}_k|}{|\tilde{\Xi}_k + (\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)|} \\ &= -\log|I + (\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1}| \\ &= -\mathrm{Tr}\left((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1}\right) - R_1,\end{aligned} \tag{B.25}$$

where

$$R_1 = \sum_i \left[\log\Big(1 + \lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\Big) - \lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\right]. \tag{B.26}$$

Therefore we have

$$\mathrm{D_{KL}}(\tilde{\rho}_{k+1}\|\pi) - \mathrm{D_{KL}}(\tilde{\rho}_k\|\pi) = \frac{1}{2}\left[\mathrm{Tr}\Big((\tilde{\Xi}_\infty^{-1} - \tilde{\Xi}_k^{-1})(\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\Big) - R_1\right]. \tag{B.27}$$

To bound R_1, we need to control the eigenvalues of $(\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1}$. Since all the matrices in question commute, from (B.21), we have

$$\tilde{\Xi}_{k+1} - \tilde{\Xi}_k = (2\eta\beta^{-1}(\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})\Xi_k + \eta^2\beta^{-2}(\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})^2\Xi_k)K_+^{-2}. \tag{B.28}$$

We now control the first term of (B.27). This expression is

$$\begin{aligned}
&\mathrm{Tr}\Big((\tilde{\Xi}_\infty^{-1} - \tilde{\Xi}_k^{-1})(\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\Big) \\
&= \eta\beta^{-1}\,\mathrm{Tr}\Big((\tilde{\Xi}_\infty^{-1} - \tilde{\Xi}_k^{-1})K_+^{-2}\left[2(\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})\Xi_k + \eta\beta^{-1}(\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})\Xi_k(\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})\right]\Big) \\
&= -2\eta\beta^{-1}\,\mathrm{Tr}\Big((\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})^2K_+^{-2}\Xi_k\Big) + \eta^2\beta^{-2}\,\mathrm{Tr}\Big((\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})^3K_+^{-2}\Xi_k\Big),
\end{aligned}$$

where we use (B.19), (B.15), and finally collect terms. To control all of (B.27), we use the sum-form of trace, and are required to upper-bound the expressions

$$\lambda_i\left((\tilde{\Xi}_\infty^{-1} - \tilde{\Xi}_k^{-1})(\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\right) - \log\Big(1 + \lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\Big) + \lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1}).$$

Summing this expression over i yields (B.27).

Step-size condition: assume $\eta\beta^{-1} \le (\max_i|\lambda_i(\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})|)^{-1}/2$. We abuse notation now to write λ_i as the eigenvalue of a matrix with respect to the i-th shared eigenvector v_i.

Case 1: suppose the eigenvalue of $(\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})$ is positive. Using (B.28), we have that $\lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1}) > 0$, so we can control

$$\begin{aligned}
&\log\Big(1 + \lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\Big) - \lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1}) \\
&\qquad \ge -\lambda_i^2((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1}) \\
&= -\lambda_i((\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})^2K_+^{-4}\Xi_k^2\tilde{\Xi}_k^{-2})[2\eta\beta^{-1} + \eta^2\beta^{-2}\lambda_i(\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})]^2.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\lambda_i((\tilde{\Xi}_\infty^{-1}-\tilde{\Xi}_k^{-1})(\tilde{\Xi}_{k+1}-\tilde{\Xi}_k))-\log\Big(1+\lambda_i((\tilde{\Xi}_{k+1}-\tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\Big)+\lambda_i((\tilde{\Xi}_{k+1}-\tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\\
&\le \lambda_i((\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})^2K_+^{-2}\Xi_k)\\
&\quad\cdot\Big[-2\eta\beta^{-1}+\eta^2\beta^{-2}\lambda_i(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})+\lambda_i(K_+^{-2}\Xi_k\tilde{\Xi}_k^{-2})[2\eta\beta^{-1}+\eta^2\beta^{-2}\lambda_i(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})]^2\Big]\\
&\le \lambda_i((\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})^2K_+^{-2}\Xi_k)\\
&\quad\cdot\Big[-2\eta\beta^{-1}+\eta^2\beta^{-2}\lambda_i(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})+\lambda_i(\tilde{\Xi}_k^{-1})[2\eta\beta^{-1}+\eta^2\beta^{-2}\lambda_i(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})]^2\Big]\\
&\le \lambda_i((\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})^2K_+^{-2}\Xi_k)\\
&\quad\cdot\Big[-2\eta\beta^{-1}+\eta^2\beta^{-2}\lambda_i(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})+\frac{1}{2}\beta T^{-1}(1+Tc^{-1})\eta^2\beta^{-2}[2+\eta\beta^{-1}\lambda_i(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})]^2\Big],
\end{aligned}$$

where we use that $\tilde{\Xi}_k \succeq K_+^{-2}\Xi_k$ in the second inequality, and then that $\tilde{\Xi}_k \succeq 2\beta^{-1}TK_+^{-1} \succeq 2\beta^{-1}T(1+Tc^{-1})^{-1}I$ in the third inequality.

Case 2: the eigenvalue of $(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})$ is negative. Then

$$\begin{aligned}
&\log\Big(1+\lambda_i((\tilde{\Xi}_{k+1}-\tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\Big)-\lambda_i((\tilde{\Xi}_{k+1}-\tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\\
&\qquad\ge -\frac{\lambda_i^2((\tilde{\Xi}_{k+1}-\tilde{\Xi}_k)\tilde{\Xi}_k^{-1})}{1+\lambda_i((\tilde{\Xi}_{k+1}-\tilde{\Xi}_k)\tilde{\Xi}_k^{-1})}\\
&= -\frac{\lambda_i^2\left[(2\eta\beta^{-1}(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})\Xi_k+\eta^2\beta^{-2}(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})^2\Xi_k)K_+^{-2}\tilde{\Xi}_k^{-1}\right]}{1+\lambda_i\left[(2\eta\beta^{-1}(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})\Xi_k+\eta^2\beta^{-2}(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})^2\Xi_k)K_+^{-2}\tilde{\Xi}_k^{-1}\right]}\\
&\ge -\frac{\lambda_i^2\left[(2\eta\beta^{-1}(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})\Xi_k)K_+^{-2}\tilde{\Xi}_k^{-1}\right]}{1+\lambda_i\left[(2\eta\beta^{-1}(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})\Xi_k+\eta^2\beta^{-2}(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})^2\Xi_k)K_+^{-2}\tilde{\Xi}_k^{-1}\right]},
\end{aligned}$$

where we use the step-size condition to make sure the denominator is positive, then discard the second order term in the numerator. This uses (B.17), which gives the control $\lambda_i(\Xi_kK_+^{-2}\tilde{\Xi}_k^{-1}) \le \lambda_i(\tilde{\Xi}_k\tilde{\Xi}_k^{-1}) = 1$. Using the step-size condition on $\eta\beta^{-1}$ again and the quadratic structure, the denominator is at least $1/4$, and we have

$$\begin{aligned}
&\log\Big(1+\lambda_i((\tilde{\Xi}_{k+1}-\tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\Big)-\lambda_i((\tilde{\Xi}_{k+1}-\tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\\
&\qquad\ge -4\lambda_i^2\left[(2\eta\beta^{-1}(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})\Xi_k)K_+^{-2}\tilde{\Xi}_k^{-1}\right]
\end{aligned}$$

Therefore

$$\begin{aligned}
&\lambda_i((\tilde{\Xi}_\infty^{-1}-\tilde{\Xi}_k^{-1})(\tilde{\Xi}_{k+1}-\tilde{\Xi}_k))-\log\Big(1+\lambda_i((\tilde{\Xi}_{k+1}-\tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\Big)+\lambda_i((\tilde{\Xi}_{k+1}-\tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\\
&\le \lambda_i((\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})^2K_+^{-2}\Xi_k)\left[-2\eta\beta^{-1}+\eta^2\beta^{-2}\lambda_i(\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})+4\lambda_i(\Xi_kK_+^{-2}\tilde{\Xi}_k^{-2})(2\eta\beta^{-1})^2\right]\\
&\le \lambda_i((\tilde{\Xi}_k^{-1}-\tilde{\Xi}_\infty^{-1})^2K_+^{-2}\Xi_k)\left[-2\eta\beta^{-1}+4\lambda_i(\tilde{\Xi}_k^{-1})(2\eta\beta^{-1})^2\right]
\end{aligned}$$

Taking $\eta\beta^{-1} \le \min((\max_i |\lambda_i(\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})|)^{-1}/2, 3\lambda_{\min}(\tilde{\Xi}_k)/32)$, we have in **Case 1**:

$$\begin{aligned}
&\lambda_i((\tilde{\Xi}_\infty^{-1} - \tilde{\Xi}_k^{-1})(\tilde{\Xi}_{k+1} - \tilde{\Xi}_k) - \log\Big(1 + \lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\Big) + \lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1})) \\
&\le \lambda_i((\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})^2 K_+^{-2}\Xi_k) \\
&\quad \cdot \Big[-2\eta\beta^{-1} + \eta^2\beta^{-2}\lambda_i(\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1}) + \lambda_i(\tilde{\Xi}_k^{-1})[2\eta\beta^{-1} + \eta^2\beta^{-2}\lambda_i(\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})]^2\Big] \\
&\le \eta\beta^{-1}\lambda_i((\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})^2 K_+^{-2}\Xi_k)\left[-2 + 1/2 + \left(\frac{5}{2}\right)^2 \cdot \frac{3}{32}\right] \\
&\le -\frac{1}{2}\eta\beta^{-1}\lambda_i((\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})^2 K_+^{-2}\Xi_k)
\end{aligned}$$

Case 2:

$$\begin{aligned}
&\lambda_i((\tilde{\Xi}_\infty^{-1} - \tilde{\Xi}_k^{-1})(\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)) - \log\Big(1 + \lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\Big) + \lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1}) \\
&\qquad \le \lambda_i((\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})^2 K_+^{-2}\Xi_k)\left[-2\eta\beta^{-1} + 4\lambda_i(\tilde{\Xi}_k^{-1})(2\eta\beta^{-1})^2\right] \\
&\qquad \le \eta\beta^{-1}\lambda_i((\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})^2 K_+^{-2}\Xi_k)\left[-2 + 16 \cdot \frac{3}{32}\right] \\
&\qquad = -\frac{1}{2}\eta\beta^{-1}\lambda_i((\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})^2 K_+^{-2}\Xi_k)
\end{aligned}$$

In either case we have that

$$\lambda_i((\tilde{\Xi}_\infty^{-1} - \tilde{\Xi}_k^{-1})(\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)) - \log\Big(1 + \lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\Big) + \lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1}) \tag{B.29}$$

$$\le -\frac{1}{2}\eta\beta^{-1}\lambda_i((\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})^2 K_+^{-2}\Xi_k) \tag{B.30}$$

and summing over i we have

$$\begin{aligned}
&\mathrm{D_{KL}}(\tilde{\rho}_{k+1}\|\pi) - \mathrm{D_{KL}}(\tilde{\rho}_k\|\pi) \\
&= \frac{1}{2}\sum_i \Big[\lambda_i((\tilde{\Xi}_\infty^{-1} - \tilde{\Xi}_k^{-1})(\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)) \\
&\qquad\qquad - \log\Big(1 + \lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\Big) + \lambda_i((\tilde{\Xi}_{k+1} - \tilde{\Xi}_k)\tilde{\Xi}_k^{-1})\Big] \\
&\le -\frac{1}{4}\eta\beta^{-1}\operatorname{Tr}\Big((\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})^2 K_+^{-2}\Xi_k\Big) \\
&\le -\frac{1}{4}\eta\beta^{-1}\lambda_{\min}(K_+^{-2}\Xi_k\tilde{\Xi}_k^{-1})\operatorname{Tr}\Big((\tilde{\Xi}_k^{-1} - \tilde{\Xi}_\infty^{-1})^2\tilde{\Xi}_k\Big) \\
&= -\frac{1}{4}\eta\beta^{-1}\lambda_{\min}(K_+^{-2}\Xi_k\tilde{\Xi}_k^{-1})\mathcal{I}_\pi^M(\tilde{\rho}_k) \\
&\le -\frac{1}{2C}\eta\beta^{-1}\lambda_{\min}(K_+^{-2}\Xi_k\tilde{\Xi}_k^{-1})\mathrm{D_{KL}}(\tilde{\rho}_k\|\pi).
\end{aligned}$$

where we use that the log-Sobolev constant of $\tilde{\Xi}_\infty = \sqrt{M}^{-1}\Sigma\sqrt{M}^{-1}$ is C^{-1} as in Appendix B.1. Finally note that

$$\begin{aligned}\lambda_{\min}(K_+^{-2}\Xi_k\tilde{\Xi}_k^{-1}) &\geq \frac{\lambda_{\min}(K_+^{-2}\Xi_k)}{2\beta^{-1}T(1+TC^{-1})^{-1}+\lambda_{\min}(K_+^{-2}\Xi_k)}\\ &\geq \frac{(1+Tc^{-1})^{-2}\lambda_{\min}(\Xi_k)}{2\beta^{-1}T(1+TC^{-1})^{-1}+(1+Tc^{-1})^{-2}\lambda_{\min}(\Xi_k)}\\ &= \frac{1}{1+2\beta^{-1}T(1+TC^{-1})^{-1}(1+Tc^{-1})^2\lambda_{\min}(\Xi_k)^{-1}},\end{aligned}$$

where in the second inequality, we use that $x/(c+x)$ is increasing and $\lambda_{\min}(K_+^{-2}\Xi_k) \geq (1+Tc^{-1})^{-2}\lambda_{\min}(\Xi_k)$.

Uniqueness: now we have established existence of a stationary solution as the Gaussian $\mathcal{N}(0,\Sigma_\infty)$ satisfying $\mathrm{WProx}^M_{T,V}(\mathcal{N}(0,\Sigma_\infty)) = \mathcal{N}(0,\beta^{-1}\Sigma)$. From the modified Fokker–Planck equation (4.9), stationarity at a distribution ν is equivalent to satisfying $\mathrm{WProx}^M_{T,V}\,\nu \propto \exp(-\beta V) = \mathcal{N}(0,\beta^{-1}\Sigma)$. Consider the definition of the regularized Wasserstein proximal as the coupled Fokker–Planck and Hamilton–Jacobi equation, plus the Cole–Hopf transformation. Since the forward- and backwards- anisotropic heat equations are well defined and are injective, the inverse of the Wasserstein proximal is unique. ∎

Appendix C. Additional results.

C.1. Sufficient conditions for a unique inverse proximal. We wish to find sufficient conditions on a distribution ρ such that the PRWPO (3.10) is invertible, that is, there exists a unique distribution ρ_0 for which $\rho = \mathrm{WProx}^M_{T,V}\,\rho_0$ is the preconditioned regularized Wasserstein proximal. Equivalently, from (3.4), we find sufficient conditions that the backward heat equation for $\hat{\eta}$ from time T to 0 is well defined. The forward heat equation is always well defined.

Recall that the solution to the anisotropic (forward) heat equation

$$\partial_t f = \beta^{-1}\nabla\cdot(M\nabla f)$$

is given in Fourier space by

$$f(t,x) = \int_{\mathbb{R}^d} \hat{f}(\xi)e^{i\langle\xi,x\rangle - t\beta^{-1}\xi^\top M\xi}\mathrm{d}\xi. \tag{C.1}$$

In order for $f(-T,x)$ to be well defined by Plancherel's theorem, we simply need that the Fourier transform $\hat{f}(\xi)$ decays faster than $e^{-T\beta^{-1}\xi^\top M\xi}$. Since the terminal condition for $\hat{\eta}$ is $\hat{\eta}(T,x) = \rho(x)e^{-\beta\Phi(T,M^{-1}x)/2} = \rho(T,x)e^{\beta V(x)/2}$, which evolves through the backward heat equation, we want $\rho(x)e^{\beta V(x)/2}$ to have Fourier transform decaying (at least) faster than $e^{-T\beta^{-1}\xi^\top M\xi}$.

C.1.1. Application: maximum value of T. Consider the quadratic potential case where $V = x^\top\Sigma^{-1}x/2$, and we wish to find ρ_0 whose regularized Wasserstein proximal is proportional to $\exp(-\beta V)$. The Fourier transform of a normal distribution $\rho\sim\mathcal{N}(\mu,\Sigma)$ (up to some factors of $\sqrt{2\pi}$) is

$$\hat{\rho} = \mathbb{E}_{x\sim\rho}\left[e^{-i\langle\xi,x\rangle}\right] = e^{i\langle\xi,\mu\rangle - \frac{1}{2}\xi^\top\Sigma\xi}. \tag{C.2}$$

In this case, $\rho \exp(\beta V/2) \propto \exp(-\beta V/2)$ is proportional to a Gaussian with covariance $2\beta^{-1}\Sigma$, and thus the decay of the Fourier transform is $\sim e^{-\beta^{-1}\xi^\top \Sigma \xi}$. This is at least faster than $e^{-T\beta^{-1}\xi^\top M\xi}$ if and only if $\Sigma \succeq TM$.

C.2. Regularized Wasserstein proximal is a Wasserstein-2 contraction. From (5.4) or (B.11), we have that K_M is a composition of the Gaussian convolution Markov kernel $\delta_y \mapsto \mathcal{N}(y, (\frac{\beta\Sigma^{-1}}{2} + \frac{\beta M^{-1}}{2T})^{-1})$ with the (linear) map $\omega : \mathbb{R}^d \to \mathbb{R}^d$, defined as $y \mapsto (\frac{\Sigma^{-1}}{2} + \frac{M^{-1}}{2T})^{-1} \frac{M^{-1}}{2T} y$. Since $c^{-1}M^{-1} \succeq \Sigma^{-1} \succeq C^{-1}M^{-1}$, we have that $\frac{\beta\Sigma^{-1}}{2} + \frac{\beta M^{-1}}{2T} \succeq (\frac{\beta C^{-1}}{2} + \frac{\beta}{2T})M^{-1}$. Therefore the eigenvalues of the scaling matrix satisfy

$$\lambda_i\left(\frac{M^{-1}}{2T}\left(\frac{\Sigma^{-1}}{2} + \frac{M^{-1}}{2T}\right)^{-2}\frac{M^{-1}}{2T}\right) \leq \left(\frac{C^{-1}}{2} + \frac{1}{2T}\right)^{-2}\left(\frac{1}{2T}\right)^2 < 1.$$

In particular, the scaling map ω is contractive with constant $\zeta := (\frac{C^{-1}}{2} + \frac{1}{2T})^{-1}(\frac{1}{2T}) < 1$.

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be any probability measures with finite second moment, and let $\gamma \in \Pi(\mu, \nu)$. We wish to bound

$$\mathcal{W}_2(\mathrm{WProx}^M_{T,V}\,\mu, \mathrm{WProx}^M_{T,V}\,\nu).$$

Note that $\mathrm{WProx}^M_{T,V}\,\mu = \mathcal{N}(0, (\frac{\beta\Sigma^{-1}}{2} + \frac{\beta M^{-1}}{2T})^{-1}) * \omega_\# \mu$. Since ω is a contractive map with constant ζ, we have that

$$\mathcal{W}_2(\omega_\#\mu, \omega_\#\nu) \leq \zeta \mathcal{W}_2(\mu, \nu). \tag{C.3}$$

From [39, Lem. 5.2], since the kernel given by the pdf of $\mathcal{N}(0, (\frac{\beta\Sigma^{-1}}{2} + \frac{\beta M^{-1}}{2T})^{-1})$ is even, we have moreover that for any μ, ν probability measures with finite second moment, that

$$\mathcal{W}_2\left(\mathcal{N}\left(0, \left(\frac{\beta\Sigma^{-1}}{2} + \frac{\beta M^{-1}}{2T}\right)^{-1}\right) * \mu, \mathcal{N}\left(0, \left(\frac{\beta\Sigma^{-1}}{2} + \frac{\beta M^{-1}}{2T}\right)^{-1}\right) * \nu\right) \leq \mathcal{W}_2(\mu, \nu). \tag{C.4}$$

Therefore, we have that for quadratic V, and any μ, ν in Wasserstein-2 space, that

$$\mathcal{W}_2(\mathrm{WProx}^M_{T\cdot V}\,\mu, \mathrm{WProx}^M_{T\cdot V}\,\nu) \leq \zeta \mathcal{W}_2(\mu, \nu). \tag{C.5}$$

This also holds for any $p \in [1, \infty)$.

C.3. Mean reversion from high variance initialization. We list some simple definitions and properties before stating the inequality Proposition C.4. We fix $\beta = 1$ for simplicity.

Definition C.1 (Relative Smoothness/Convexity). *Let $\Psi : \mathcal{X} \to \mathbb{R}$ be a differentiable convex function, defined on a convex set $\mathcal{X}$ (with non-empty interior), which will be used as a reference. Let $f : \mathcal{X} \to \mathbb{R}$ be another differentiable convex function.*

*f is L-*smooth relative to *Ψ if for any $x, y \in int(\mathcal{X})$,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + L B_\Psi(y, x). \tag{C.6}$$

f is μ-strongly-convex relative to Ψ *if for any* $x, y \in int(\mathcal{X})$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \mu B_\Psi(y, x). \tag{C.7}$$

In the special case where $\Psi(x) = \frac{1}{2}x^\top Ax$ *for some (positive definite) symmetric* A*, we write that* f *is* L*-smooth relative to* A *and* μ*-strongly-convex relative to* A *respectively. In particular, we have that* $B_\Psi(x, y) = \frac{1}{2}(x - y)^\top A(x - y)$.

Proposition C.2 (Control on spectral radius). *For a row- (or column-) stochastic matrix* $A \in \mathbb{R}^{n\times n}$*, the spectral norm* $\|A\|_2$ *equals its largest singular value* $\sigma_1(A)$*, and moreover satisfies* $\|A\|_2 \leq \sqrt{n}$.

Proof. $\|A\|_2 \leq \sqrt{n}\|A\|_\infty = \sqrt{n}$, where $\|\cdot\|_\infty$ is the induced ∞-norm, equal to the maximum absolute row sum. ∎

Proposition C.3. *A function* $f : \mathbb{R}^n \to \mathbb{R}$ *is* μ*-strongly convex relative to* A *if and only if for all* $x, y \in \mathbb{R}^n$,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \langle A(x - y), x - y \rangle. \tag{C.8}$$

Proof. Note that μ-relative strong convexity is equivalent to $f - \mu\Psi$ being convex. Hence, f being μ-strongly convex relative to $\frac{1}{2}x^\top Ax$ is equivalent to the following by the monotone gradient condition,

$$\langle \nabla(f - \mu\Psi)(x) - \nabla(f - \mu\Psi)(y), x - y \rangle \geq 0.$$

$$\Leftrightarrow \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \langle \nabla\Psi(x) - \nabla\Psi(y), x - y \rangle = \mu \langle A(x - y), x - y \rangle.$$

∎

Proposition C.4 (Contraction-diffusion inequality). *Suppose that* V *is* μ*-strongly convex relative to* M^{-1}*, and that* V *is minimized at* $\hat{x} \in \mathbb{R}^d$*. Then the following holds for any* $\mathbf{X}$*:*

$$\langle \Delta, M^{-1}(\mathbf{X} - \hat{\mathbf{X}}) \rangle_{Frob} \leq \|\mathbf{X} - \hat{\mathbf{X}}\|_{2,M} \left(-\frac{\mu}{2}\|\mathbf{X} - \hat{\mathbf{X}}\|_{2,M} + \frac{(1 + \sqrt{N})}{2T}\|\mathbf{X} - \frac{1}{N}\mathbf{X}\mathbf{1}_N\mathbf{1}_N^\top\|_{2,M} \right), \tag{C.9}$$

where $\hat{\mathbf{X}}$ *is the tensorized minimizer*

$$\hat{\mathbf{X}} = \begin{bmatrix} \hat{x} & \dots & \hat{x} \end{bmatrix} = \hat{x}\mathbf{1}_N^\top \in \mathbb{R}^{d\times N}.$$

Proof. Let $\Delta = \Delta(\mathbf{X}^{(k)})$ be the descent direction defined as follows

$$\Delta = -\frac{1}{2}M\nabla V(\mathbf{X}) + \frac{1}{2T}(\mathbf{X} - \mathbf{X}\text{softmax}(W)^\top). \tag{C.10}$$

such that the PBRWP iteration (4.8) becomes $\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} - \eta\Delta$.

Compute, noting that $\hat{\mathbf{X}}\mathrm{softmax}(W)^\top = \hat{\mathbf{X}}$ since $\mathbf{1}_N$ is a left-eigenvector of $\mathrm{softmax}(W)^\top$ of eigenvalue 1:

$$\begin{aligned}\langle \Delta, M^{-1}(\mathbf{X}-\hat{\mathbf{X}})\rangle_{\mathrm{Frob}} &= -\frac{1}{2}\langle M\nabla V(\mathbf{X}) - M\nabla V(\hat{\mathbf{X}}), M^{-1}(\mathbf{X}-\hat{\mathbf{X}})\rangle_{\mathrm{Frob}} \\ &\quad + \frac{1}{2T}\langle (\mathbf{X}-\hat{\mathbf{X}})(I_{N\times N} - \mathrm{softmax}(W)^\top), M^{-1}(\mathbf{X}-\hat{\mathbf{X}})\rangle_{\mathrm{Frob}} \\ &= -\frac{1}{2}\sum_{i=1}^N \langle \nabla V(\mathbf{x}_i) - \nabla V(\hat{x}), \mathbf{x}_i - \hat{x}\rangle \\ &\quad + \frac{1}{2T}\langle (\mathbf{X}-\hat{\mathbf{X}})(I_{N\times N} - \mathrm{softmax}(W)^\top), M^{-1}(\mathbf{X}-\hat{\mathbf{X}})\rangle_{\mathrm{Frob}}\end{aligned}$$

where in the second equality, we use that $\langle AX, Y\rangle_{\mathrm{Frob}} = \langle X, A^\top Y\rangle_{\mathrm{Frob}}$. To control the second term we utilize Proposition C.2 for the column-stochastic matrix $\mathrm{softmax}(W)^\top$. We have, recalling $\hat{\mathbf{X}} = \hat{x}\mathbf{1}_N^\top$, and $\mathbf{1}_N^\top \mathrm{softmax}(W)^\top = \mathbf{1}_N^\top$,

$$\begin{aligned}&\mathbf{X} - \hat{\mathbf{X}} = (\mathbf{X} - \frac{1}{N}\mathbf{X}\mathbf{1}_N\mathbf{1}_N^\top) + \frac{1}{N}\mathbf{X}\mathbf{1}_N\mathbf{1}_N^\top + \hat{x}\mathbf{1}_N^\top \\ \Rightarrow\ &(\mathbf{X}-\hat{\mathbf{X}})(I_{N\times N} - \mathrm{softmax}(W)^\top) = (\mathbf{X} - \frac{1}{N}\mathbf{X}\mathbf{1}_N\mathbf{1}_N^\top)(I_{N\times N} - \mathrm{softmax}(W)^\top).\end{aligned}$$

$$\begin{aligned}&\langle (\mathbf{X}-\hat{\mathbf{X}})(I_{N\times N} - \mathrm{softmax}(W)^\top), M^{-1}(\mathbf{X}-\hat{\mathbf{X}})\rangle_{\mathrm{Frob}} \\ &= \langle M^{-1/2}(\mathbf{X} - \frac{1}{N}\mathbf{X}\mathbf{1}_N\mathbf{1}_N^\top)(I_{N\times N} - \mathrm{softmax}(W)^\top), M^{-1/2}(\mathbf{X}-\hat{\mathbf{X}})\rangle_{\mathrm{Frob}} \\ &\le \|M^{-1/2}(\mathbf{X} - \frac{1}{N}\mathbf{X}\mathbf{1}_N\mathbf{1}_N^\top)(I_{N\times N} - \mathrm{softmax}(W)^\top)\|_2 \|M^{-1/2}(\mathbf{X}-\hat{\mathbf{X}})\|_2 \\ &\le (1+\sqrt{N})\|M^{-1/2}(\mathbf{X} - \frac{1}{N}\mathbf{X}\mathbf{1}_N\mathbf{1}_N^\top)\|_2 \|M^{-1/2}(\mathbf{X}-\hat{\mathbf{X}})\|_2\end{aligned}$$

We use Proposition C.2 to control the spectral norm (induced 2-norm) of $\mathrm{softmax}(W)^\top$. In addition, we use the tracial matrix Hölder inequality $|\langle A, B\rangle_{\mathrm{Frob}}| = |\operatorname{Tr}(A^\top B)| \le \|A\|_p \|B\|_q$ with $p^{-1} + q^{-1} = 1$, and $\|A\|_p$ is the p-Schatten norm (equal to the Frobenius norm for $p = 2$).

From the assumption on V and Proposition C.3, we have control on the other term as well:

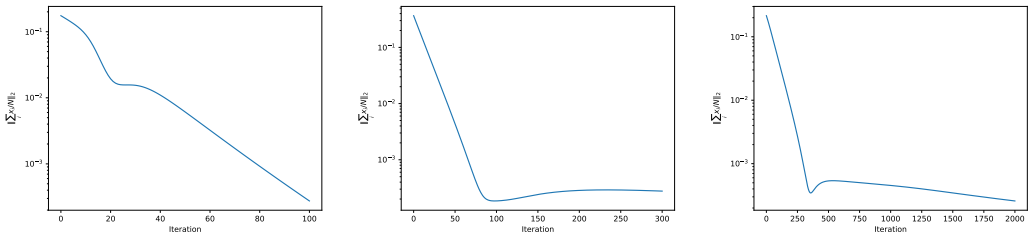
$$\langle \nabla V(x) - \nabla V(y), x - y\rangle \ge \mu \|x-y\|_M^2.$$

Putting everything together, we have

$$\langle \Delta, M^{-1}(\mathbf{X}-\hat{\mathbf{X}})\rangle_{\mathrm{Frob}} \tag{C.11}$$

$$\le -\frac{\mu}{2}\sum_{i=1}^N \|\mathbf{x}_i - \hat{x}\|_M^2 + \frac{(1+\sqrt{N})}{2T}\|M^{-1/2}(\mathbf{X} - \frac{1}{N}\mathbf{X}\mathbf{1}_N\mathbf{1}_N^\top)\|_2 \|M^{-1/2}(\mathbf{X}-\hat{\mathbf{X}})\|_2 \tag{C.12}$$

$$= \|\mathbf{X}-\hat{\mathbf{X}}\|_{2,M}\left(-\frac{\mu}{2}\|\mathbf{X}-\hat{\mathbf{X}}\|_{2,M} + \frac{(1+\sqrt{N})}{2T}\|\mathbf{X} - \frac{1}{N}\mathbf{X}\mathbf{1}_N\mathbf{1}_N^\top\|_{2,M}\right) \tag{C.13}$$



(a) $N = 6, \eta = 0.1$ (b) $N = 100, \eta = 0.1$ (c) $N = 100, \eta = 0.02$

Figure 11. *Plot of the norm of the average of the points* $\|\frac{1}{N}\sum_i \mathbf{x}_i\|$ *with respect to iteration, for the target distribution* $\mathcal{N}(0, I_2)$ *and identity preconditioner,* $T = 2$*, and various choices of* N *and step-size. The mean does not converge linearly, and instead has two separate regimes.*

where for a positive definite symmetric matrix $M \in \mathbb{R}^{d\times d}$, we define the scaled $(2, M)$ norm of a $d \times N$ matrix by

$$\|A\|_{2,M}^2 := \|M^{-1/2}A\|_2^2 = \mathrm{Tr}\Big(A^\top M^{-1} A\Big). \tag{C.14}$$

■

Interpretation: if the particles are far from the minimizer and have small standard deviation, then they will flow together (on average) to the minimizer. We note that we do not use the actual values of the interaction matrix W.

Figure 11 demonstrates the change of regime that occurs from a high variance initial distribution. The target distribution is the two-dimensional standard Gaussian with $M = I$.

C.4. Particle diffusion boundedness. We refer back to the Gaussian case where the target distribution is $\mathcal{N}(0, \beta^{-1}\Sigma)$.

Proposition C.5. *Suppose* $M = \Sigma$ *(and* $T < 1$*), and that* $\mathbf{x}_1$ *is an exterior point of the convex hull of* $\{\mathbf{x}_i\}$*. Suppose* $\delta > 0$ *is such that for all* $j \neq 1$*,*

$$\mathbf{x}_1^\top M^{-1}\mathbf{x}_j \leq \mathbf{x}_1^\top M^{-1}\mathbf{x}_1 - \delta\|\mathbf{x}_1\|_M, \tag{C.15}$$

Then, if $\mathbf{x}_1$ *has sufficiently large norm, in particular assuming,*

$$\delta\|\mathbf{x}_1\|_M \geq 2\beta^{-1}T\log\left(\frac{2(N-1)}{T}\right), \tag{C.16}$$

then for small step-size, the updated point satisfies $\|\mathbf{x}_1^{(k+1)}\|_M < \|\mathbf{x}_1^{(k)}\|_M$*.*

Proof. Recall the particle evolution

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

where in the Gaussian case, the interaction weight matrix W takes the special form

$$W_{ij} = \frac{\beta}{4T}\left(2\mathbf{x}_i^\top M^{-1}\mathbf{x}_j - \mathbf{x}_j^\top(M + TM\Sigma^{-1}M)^{-1}\mathbf{x}_j\right). \tag{C.17}$$

We control the values of W_{1j}. In this case $M = \Sigma$ we have

$$W_{ij} = \frac{\beta}{4T}\left(2\mathbf{x}_i^\top M^{-1}\mathbf{x}_j - (1+T)^{-1}\mathbf{x}_j^\top M^{-1}\mathbf{x}_j\right)$$

which controls

$$W_{11} = \frac{\beta}{4T}(2-(1+T)^{-1})\|\mathbf{x}_1\|_M^2$$

and for $j \neq 1$,

$$W_{1j} \leq \frac{\beta}{4T}(2\|\mathbf{x}_1\|_M^2 - (1+T)^{-1}\|\mathbf{x}_1\|_M^2 - 2\delta\|\mathbf{x}_1\|_M) \tag{C.18}$$

$$= W_{11} - \frac{\delta\beta}{2T}\|\mathbf{x}_1\|_M. \tag{C.19}$$

We have that

$$\sum_{i=1}^N \exp(W_{1j}) \leq \exp(W_{11})[1 + (N-1)\exp(-\delta\beta\|\mathbf{x}_1\|_M/(2T))] \tag{C.20}$$

Considering $\langle \mathbf{x}_1^{(k+1)} - \mathbf{x}_1^{(k)}, \mathbf{x}_1^{(k)}\rangle_M$, we want to show it is negative for big R. Then the norm of the iteration decreases. The iteration takes the form

$$\frac{1}{\eta}\langle \mathbf{x}_1^{(k+1)} - \mathbf{x}_1^{(k)}, \mathbf{x}_1^{(k)}\rangle_M \tag{C.21}$$

$$= -\frac{1}{2}\|\mathbf{x}_1\|_M^2 + \frac{1}{2T}\sum_{j=2}^N \mathrm{softmax}(W_{1\cdot})_j\langle \mathbf{x}_1 - \mathbf{x}_j, \mathbf{x}_1\rangle_M \tag{C.22}$$

By maximality we have

$$0 < \langle \mathbf{x}_1 - \mathbf{x}_j, \mathbf{x}_1\rangle_M < \begin{cases} \|\mathbf{x}_1\|_M^2 - \delta^2/2, & \langle \mathbf{x}_1, \mathbf{x}_j\rangle_M > 0; \\ 2\|\mathbf{x}_1\|_M^2, & \langle \mathbf{x}_1, \mathbf{x}_j\rangle_M < 0. \end{cases} \tag{C.23}$$

If $\sum_{j=2}^N \mathrm{softmax}(W_{1\cdot})_j \leq \frac{T}{2}$, then naturally

$$-\frac{1}{2}\|\mathbf{x}_1\|_M^2 + \frac{1}{2T}\sum_{j=2}^N \mathrm{softmax}(W_{1\cdot})_j\langle \mathbf{x}_1 - \mathbf{x}_j, \mathbf{x}_1\rangle_M$$
$$< -\frac{1}{2}\|\mathbf{x}_1\|_M^2 + \frac{1}{2T}\frac{T}{2}2\|\mathbf{x}_1\|_M^2 = 0.$$

So we wish to find a sufficient condition on $\|\mathbf{x}_1\|_M$ such that $\sum_{j=2}^N \mathrm{softmax}(W_{1\cdot})_j \leq \frac{T}{2}$. It is sufficient that $\mathrm{softmax}(W_{1\cdot})_1 \geq 1 - T/2$, which is satisfied if

$$[1 + (N-1)\exp(-\delta\beta\|\mathbf{x}_1\|_M/(2T))]^{-1} \geq 1 - \frac{T}{2} \tag{C.24}$$

$$\Leftrightarrow [1 + \exp(-\delta\beta\|\mathbf{x}_1\|_M/(2T) + \log(N-1))]^{-1} \geq 1 - \frac{T}{2} \tag{C.25}$$

Note $(1+x)^{-1} > 1-x$, so it is sufficient that

$$\exp(-\delta\beta\|\mathbf{x}_1\|_M/(2T) + \log(N-1)) \leq \frac{T}{2}$$
$$\Leftrightarrow \delta\|\mathbf{x}_1\|_M \geq 2\beta^{-1}T\log\left(\frac{2(N-1)}{T}\right).$$

■

Appendix D. Sensitivity to hyperparameters for imaging experiment. Tables 3 and 4 plot the deconvolution reconstructions using PBRWP and BRWP respectively over a variety of choices of T and step-size. Both methods are run with 40 particles up to 5000 iterations. We observe that both methods are not particularly sensitive to the choice of T except for an intermediate choice of $T = 10^{-2}$, for which we observe more variance. This can be explained by variance reduction phenomena: by the regularization of the Wasserstein proximal for large T, and by the finite number of particles for small T. For the smallest step-size $\eta = 10^{-5}$, 5000 iterations are insufficient to reach convergence and therefore the performance is worse.

Table 3

*PSNR (in dB) of the mean of 40 particles computed using **PBRWP**, taken with respect the MAP estimate for the deconvolution problem. Taken at iteration 5000, at which stepsizes 1e-3 and 5e-4 have converged.*

T \ Step-size	1e-3	5e-4	2e-4	1e-4	1e-5
1e-4	30.28	34.22	39.53	43.34	31.30
1e-3	30.28	34.24	39.45	42.76	31.29
1e-2	30.30	33.95	37.88	39.37	30.97
0.5	30.29	34.22	39.53	43.37	31.30
10	30.28	34.22	39.52	43.37	31.30

Table 4

*PSNR (in dB) of the mean of 40 particles computed using **BRWP**, taken with respect the MAP estimate for the deconvolution problem. Taken at iteration 5000, at which stepsizes 1e-3, 5e-4 and 2e-4 have converged.*

T \ Step-size	1e-3	5e-4	2e-4	1e-4	1e-5
1e-4	34.22	38.20	43.36	42.75	28.59
1e-3	34.22	38.20	43.08	42.47	28.59
1e-2	34.18	37.69	40.82	39.99	28.53
0.5	34.22	38.20	43.36	42.72	28.58
10	34.22	38.20	43.37	42.76	28.59

Appendix E. Choice of variable preconditioner for BNNs. To compute the diagonal variable preconditioner, we use the second moment estimates of Adam [24], which are used to precondition a moving average of the gradient. We discard the first moment estimate since we

are not optimizing using Adam directly. Algorithm 2 details how pointwise gradient estimates (corresponding to each particle) give rise to the variable preconditioning matrices $M^{(k)}$. By applying this algorithm to each particle $\mathbf{x}_j$, which are updated according to PBRWP, we get their corresponding preconditioning matrices $M_j^{(k)}$.

Algorithm 2: Adam-based Preconditioner

Data: Objective function f, exponential decay rates $\beta_2 = 0.999$, point sequence $(x^{(l)})_{l\geq 1}$, epsilon $\epsilon = 0.001$.

Result: Preconditioners $M^{(k)}$, where $M^{(k)} = M^{(k)}(x^{(1)}, ..., x^{(k)})$.

$v_0 \leftarrow 0$; // initialize second moment vector

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

  $g_k \leftarrow \nabla f(x^{(k)})$; // compute gradient

  $v_k \leftarrow \beta_2 v_{k-1} + (1-\beta_2) g_k^2$; // update second moment estimate

  $\hat{v}_k \leftarrow v_k/(1-\beta_2^k)$; // bias correction

  $M^{(k)} = \mathrm{diag}(1/(\sqrt{\hat{v}_k} + \epsilon))$; // construct preconditioning matrix

end

return $(M^{(k)})_{k=1}^K$.