Noise-Free Sampling Algorithms via Regularized Wasserstein Proximals

Hong Ye Tan*

Joint work with: Stanley Osher†, Wuchen Li‡

28th May 2024

SIAM IS24: SDE- and PDE-based Sampling Methods for Imaging Inverse Problems

Motivation

- Sample from the target distribution ρ over $\mathbb{R}^d$ (for bounded $\mathcal{C}^1$ potential V)

$$\rho(x) \sim \exp(-V(x))$$

  - Applications: global optimization, Bayesian neural networks, generative modelling etc.
- V is known, but sampling is difficult (normalizing constant, high dimensionality...)
- Common method: **Markov Chain Monte-Carlo (MCMC)**

Fokker-Planck Equation

The Fokker-Planck equation is a PDE evolution in the density space.

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V) + \beta \Delta \rho, \quad \rho(x, 0) = \rho_0(x). \qquad \text{(Fokker-Planck)}$$

Steady state:

$$\rho_\infty(x) \sim \exp\bigl(-\beta^{-1} V(x)\bigr).$$

Fokker-Planck Equation

The Fokker-Planck equation is a PDE evolution in the density space.

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V) + \beta \Delta \rho, \quad \rho(x, 0) = \rho_0(x). \qquad \text{(Fokker-Planck)}$$

Steady state:

$$\rho_\infty(x) \sim \exp\big(-\beta^{-1} V(x)\big).$$

Equivalent particle-based evolutions for $t \in [0, +\infty)$:

1. SDE

$$dX(t) = -\nabla V(X(t))dt + \sqrt{2\beta} dW(t)$$

Classical SDE-based formulation

$$dX(t) = -\nabla V(X(t))dt + \sqrt{2\beta}dW(t) \tag{1}$$

Need to solve SDE using some discretization.

1. Forward Euler-Maruyama discretization $\longrightarrow$ Unadjusted Langevin Algorithm[1]

$$X_{k+1} = X_k - \eta\nabla V(X_k) + \sqrt{2\beta\eta}Z_k \tag{ULA}$$

2. Adding a correction step $\longrightarrow$ Metropolis-adjusted Langevin Algorithm (MALA)

Ergodicity from noise. Convergence from ergodic theory: evolution defines an ergodic Markov chain, which converges to the invariant distribution.

[1] $Z_k \sim \mathcal{N}(0, I)$ i.i.d. normal

Fokker-Planck Equation

The Fokker-Planck equation is a PDE evolution in the density space.

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V) + \beta \Delta \rho, \quad \rho(x, 0) = \rho_0(x). \qquad \text{(Fokker-Planck)}$$

Steady state:

$$\rho_\infty(x) \sim \exp\big(-\beta^{-1} V(x)\big).$$

Equivalent particle-based evolutions for $t \in [0, +\infty)$:

1. SDE

$$dX(t) = -\nabla V(X(t))dt + \sqrt{2\beta} dW(t)$$

2. Score-based ODE (where $X(t) \sim \rho(t, \cdot)$ is the density at time t)

$$\frac{dX}{dt} = -\nabla V(X) - \beta \nabla \log \rho(t, X)$$

Score-based ODE formulation

$$\frac{dX}{dt} = -\nabla V(X) - \beta \nabla \log \rho(t, X)$$

- Difficulty: what is $\rho(t, \cdot)$?
- Kernel density estimation: using samples to approximate $\rho(t, X)$
  - Caveats: mode collapse, choice of kernel, hyperparameter choices
- Directly learning the score using neural networks
  - Empirically works in high dimensions, see diffusion models e.g. DALL-E

Score-based ODE formulation

$$\frac{dX}{dt} = -\nabla V(X) - \beta \nabla \log \rho(t, X)$$

- Difficulty: what is $\rho(t, \cdot)$?
- Kernel density estimation: using samples to approximate $\rho(t, X)$
  - Caveats: mode collapse, choice of kernel, hyperparameter choices
- Directly learning the score using neural networks
  - Empirically works in high dimensions, see diffusion models e.g. DALL-E
- Our proposed method: natural "choice of kernel" based on the Wasserstein proximal (among other things).

Definition

Let ρ_0 be a probability density function with finite second moment, 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

$$\rho_T = \mathrm{WProx}_{TV}(\rho_0) \coloneqq \underset{q \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min} \int_{\mathbb{R}^d} V(x) q(x)\, dx + \frac{\mathcal{W}(\rho_0, q)^2}{2T}, \tag{2}$$

where $\mathcal{W}(\rho_0, q)$ is the Wasserstein-2 distance between ρ_0 and q, and $\mathcal{P}_2$ is the set of probability density functions q with finite second moment.

Inspiration: JKO[2] scheme

The iterations

$$\rho_{T+1} = \underset{q \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min} \int_{\mathbb{R}^d} V(x)q(x) + \beta q \log q \; dx + \frac{\mathcal{W}(\rho_T, q)^2}{2h} \tag{3}$$

converge (weakly) to the solution of the Fokker-Planck equation as $h \to 0$.

Similar to a proximal descent method in variational analysis.

[2] Jordan, Kinderlehler, Otto. *The variational formulation of the Fokker–Planck equation*. SIMA 1998.

Benamou-Brenier PDE formulation of the Wasserstein proximal:

$$\begin{cases} \partial_t \rho(t,x) + \nabla_x \cdot (\rho(t,x) \nabla_x \Phi(t,x)) = 0 & \text{(4a)} \\ \partial_t \Phi(t,x) + \dfrac{1}{2} \|\nabla_x \Phi(t,x)\|^2 = 0 & \text{(4b)} \\ \rho(0,x) = \rho_0(x), \quad \Phi(T,x) = -V(x). & \text{(4c)} \end{cases}$$

Φ: Kantorovich dual variable.

Benamou-Brenier PDE formulation of the regularized Wasserstein proximal:

$$\begin{cases} \partial_t \rho(t,x) + \nabla_x \cdot (\rho(t,x) \nabla_x \Phi(t,x)) = \beta \Delta_x \rho(t,x) & \text{(4a)} \\ \partial_t \Phi(t,x) + \frac{1}{2} \|\nabla_x \Phi(t,x)\|^2 = -\beta \Delta_x \Phi(t,x) & \text{(4b)} \\ \rho(0,x) = \rho_0(x), \quad \Phi(T,x) = -V(x). & \text{(4c)} \end{cases}$$

Φ: Kantorovich dual variable.

Later: regularized Wasserstein proximal $\rho(T,x)$ has a closed form, as opposed to the (non-regularized) Wasserstein proximal.

Proposed Method

3 step approximation:

1. Approximate Fokker-Planck equation with regularized Fokker-Planck equation
    - One-step time approximation using Wasserstein proximal
2. Backwards Euler time-discretization of ODE
3. Per-step approximation of density using empirical measure

This allows us to use the deterministic computation methods:

1. Deterministic computation of score using kernel formulation
2. Convolution as Monte-Carlo sampling

Magic Ingredient 1: Backwards Discretization

Standard score-based ODE:

$$\frac{dX}{dt} = -\nabla V(X) - \beta \nabla \log \rho(t, X).$$

Regularized score-based ODE (Liouville's equation):

$$\frac{dX}{dt} = \nabla \Phi(t, X) - \beta \nabla \log \rho(t, X).$$

Backwards (one-step) discretization (where Φ and $\rho_{k,T}$ use initial condition $X_k \sim \rho_{k,0}$):

$$X_{k+1} = X_k + \eta \nabla \Phi(T, X_k) - \eta \beta \nabla \log \rho_k(T, X_k).$$

Magic Ingredient 1: Backwards Discretization

Standard score-based ODE:

$$\frac{dX}{dt} = -\nabla V(X) - \beta \nabla \log \rho(t, X).$$

Regularized score-based ODE (Liouville's equation):

$$\frac{dX}{dt} = \nabla \Phi(t, X) - \beta \nabla \log \rho(t, X).$$

Backwards (one-step) discretization (where Φ and $\rho_{k,T}$ use initial condition $X_k \sim \rho_{k,0}$):

$$X_{k+1} = X_k + \eta \nabla \Phi(T, X_k) - \eta \beta \nabla \log \rho_k(T, X_k).$$

Magic step: $\Phi(T, x) = -V(x)$ by definition

$$X_{k+1} = X_k - \eta \nabla V(X_k) - \eta \beta \nabla \log \rho_{k,T}(X_k).$$

Magic Ingredient 1: Backwards Discretization

Standard score-based ODE:

$$\frac{dX}{dt} = -\nabla V(X) - \beta \nabla \log \rho(t, X).$$

Regularized score-based ODE (Liouville's equation):

$$\frac{dX}{dt} = \nabla \Phi(t, X) - \beta \nabla \log \rho(t, X).$$

Backwards (one-step) discretization (where Φ and $\rho_{k,T}$ use initial condition $X_k \sim \rho_{k,0}$):

$$X_{k+1} = X_k + \eta \nabla \Phi(T, X_k) - \eta \beta \nabla \log \rho_k(T, X_k).$$

Magic step: $\Phi(T, x) = -V(x)$ by definition

$$X_{k+1} = X_k - \eta \nabla V(X_k) - \eta \beta \nabla \log \rho_{k,T}(X_k).$$

Magic Ingredient 2: Kernel Formulation[3]

The regularized Wasserstein proximal can be written as a convolution.

$$\rho_T(x) = \int_{\mathbb{R}^d} K(x,y)\rho_0(y)\,dy, \tag{5}$$

$$K(x,y) = \frac{\exp\Big(-\frac{1}{2\beta}(V(x) + \frac{\|x-y\|^2}{2T})\Big)}{\int_{\mathbb{R}^d}\exp\Big(-\frac{1}{2\beta}(V(z) + \frac{\|z-y\|^2}{2T})\Big)\,dz}. \tag{6}$$

[3]W. Li, S. Liu, S. Osher. "A kernel formula for regularized Wasserstein proximal operators." Research in the Mathematical Sciences 10.4 (2023): 43.

Magic Ingredient 2: Kernel Formulation

The convolution is (relatively) easy to compute for empirical distributions

$$\rho_0(x) = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i} \tag{7}$$

$$\Rightarrow\ \rho_T(\mathbf{x}_i) = \frac{1}{N}\sum_{j=1}^{N}K(\mathbf{x}_i,\mathbf{x}_j) = \frac{1}{N}\sum_{j=1}^{N}\frac{\exp\left[-\frac{1}{2\beta}\left(V(\mathbf{x}_i)+\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2T}\right)\right]}{\mathcal{Z}(\mathbf{x}_j)}, \tag{8}$$

$$\mathcal{Z}(\mathbf{x}_j) \coloneqq \mathbb{E}_{z\sim\mathcal{N}(\mathbf{x}_j,2T\beta)}\left[\exp\left(-\frac{V(z)}{2\beta}\right)\right]. \tag{9}$$

Magic Ingredient 3: Empirical Approximations

- At each step, we have some samples. This defines an empirical measure (hopefully approximating the true measure)
- The regularized Wasserstein proximal applied to an empirical distribution has a simple closed form
- This allows us to compute a closed-form solution

Full algorithm

Backwards Regularized Wasserstein Proximal (BRWP) Algorithm

1. Approximate current measure using empirical measure:

$$\rho_{k,0} = \frac{1}{N}\sum_{i=1}^{N}\delta(\cdot|x_{k,i})$$

2. Compute score $\nabla \log \rho_{k,T}(x_{k,j})$ of regularized Wasserstein proximal of $\rho_{k,0}$
   - Utilise the kernel formulation
   - Three Monte Carlo integrals here: normalizing constant, ρ and $\nabla\rho$
3. Evolve the particles according to backwards Euler-discretized regularized Fokker-Planck equation

$$x_{k+1,j} = x_{k,j} - \eta\nabla V(x_{k,j}) - \eta\beta\nabla \log \rho_{k,T}(x_{k,j}), \quad j = 1, ..., N.$$

Algorithm

Algorithm 1 Backwards regularized Wasserstein proximal (BRWP) scheme

Input: Potential V, samples $(\mathbf{x}_{0,i})_{i=1}^N \sim \mu_0^{\otimes N}$, step-size $\eta > 0$, regularization parameters $T, \beta > 0$, Monte Carlo sample count P

Output: Sequence of samples $(\mathbf{x}_{k,i})_{i=1}^N$ for $k = 1, 2, ...$

for $k \in \mathbb{N}$ **do**
  for $i = 1, ..., N$ **do**
    Sample $(\mathbf{z}_{k,i,p})_{p=1}^P \sim \mathcal{N}(\mathbf{x}_{k,i}, 2\beta T I)$
    $\mathcal{Z}_{k,i} = \frac{1}{P} \sum_{p=1}^P \exp\left(-\frac{V(\mathbf{z}_{k,i,p})}{2\beta}\right)$
  end for
  for $i, j = 1, ..., N$ **do** ▷ Compute pre-requisites for score
    $\mathcal{E}_{k,i,j} = \exp\left[-\frac{1}{2\beta}\left(V(\mathbf{x}_{k,i}) + \frac{\|\mathbf{x}_{k,i} - \mathbf{x}_{k,j}\|^2}{2T}\right)\right]$
    $\mathcal{V}_{k,i,j} = -\frac{1}{2\beta}\left(\nabla V(\mathbf{x}_{k,i}) + \frac{\mathbf{x}_{k,i} - \mathbf{x}_{k,j}}{T}\right)$
  end for
  for $i = 1, ..., N$ **do**
    $\nabla \log \rho_{k,T}(\mathbf{x}_{k,i}) = (\sum_j \mathcal{V}_{k,i,j} \mathcal{E}_{k,i,j} / \mathcal{Z}_{k,j}) / (\sum_j \mathcal{E}_{k,i,j} / \mathcal{Z}_{k,j})$ ▷ Compute score
    $\mathbf{x}_{k+1,i} = \mathbf{x}_{k,i} - \eta \nabla V(\mathbf{x}_{k,i}) - \eta\beta \nabla \log \rho_{k,T}(\mathbf{x}_{k,i})$ ▷ Perform the update
  end for
end for

- Basically 3 Monte Carlo integrals
- Quadratic scaling in number of samples(!)

Gaussian analysis

1D Ornstein-Uhlenbeck process ($V = ax^2/2$)

$$dX = -aX\,dt + \sqrt{2\beta}\,dW.$$

True stationary distribution of Fokker-Planck equation: $\mathcal{N}(0, \beta/a)$.

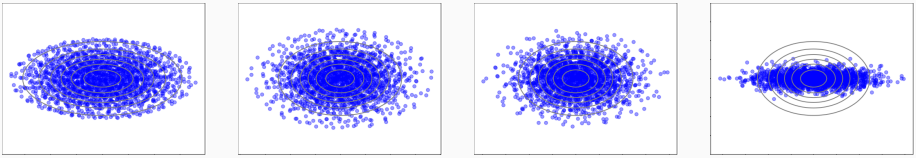
ULA	MALA	BRWP
$\mathcal{N}(0, \frac{2\beta}{(2-a\eta)a})$	$\mathcal{N}(0, \beta/a)$	$\mathcal{N}(0, \frac{\beta}{a}(1 - a^2T^2))$

Table 1: Stationary distribution of each MC

BRWP *decreases* variance (as opposed to ULA which increases variance)

Sampling Behavior (Gaussian)

T	0.05	0.25	0.5	0.999



Variance reduction phenomenon for 5-dimensional Gaussian with condition number $\kappa = 10$.

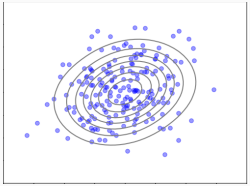
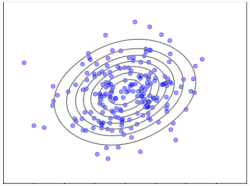
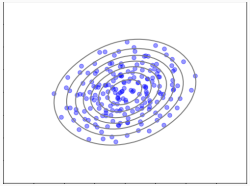
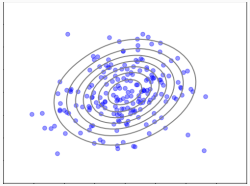
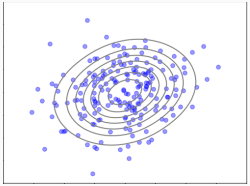
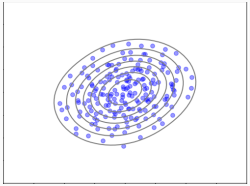
Mixing time

For Ornstein-Uhlenbeck process $V(x) = -\frac{1}{2}x^\top \Sigma^{-1} x$. Mixing time for a Gaussian with minimum eigenvalue L^{-1}, maximum eigenvalue m^{-1}, centred at x^*. $\kappa = L/m$.
Mixing time = time for distribution to be distance δ away from true distribution in total variation.

Method	Initialization	Mixing time
ULA	$\mathcal{N}(x^*, m^{-1}I)$	$\mathcal{O}\left(\frac{d\kappa^2 \log(d\kappa/\delta)}{\delta^2}\right)$
ULA	$\mathcal{N}(x^*, L^{-1}I)$	$\mathcal{O}\left(\frac{(d^3 + d\log^2(1/\delta)}{\delta^2}\right)$
MALA	$\mathcal{N}(x^*, L^{-1}I)$	$\mathcal{O}\left(d^2\kappa \log\left(\frac{\kappa}{\delta}\right)\right)$
BRWP	$\mathcal{N}(x^*, L^{-1}(1 - L^{-2}T^2)^{-1}I)$	$\mathcal{O}\left(\kappa^{3/2}\log\left(\kappa\sqrt{d}/\delta\right)\right)$

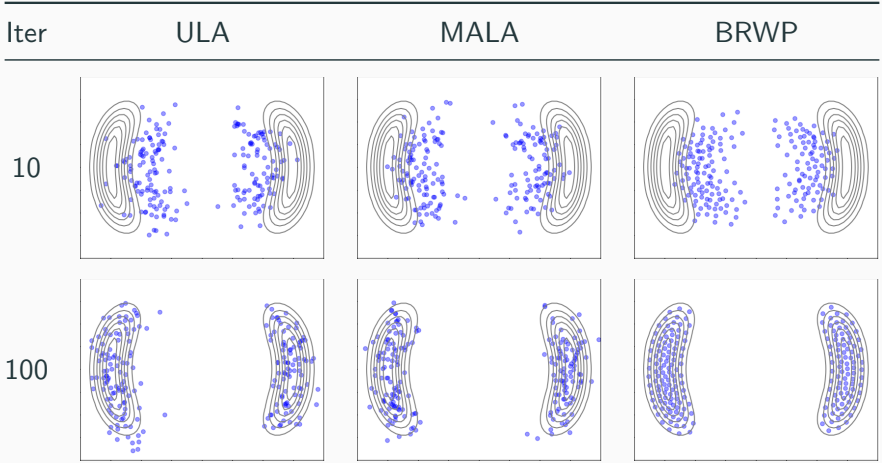
Better dimension dependence(?)

Sampling Behavior (Gaussian mixture)

Iter	ULA	MALA	BRWP
50			
500			

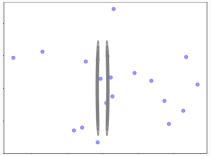
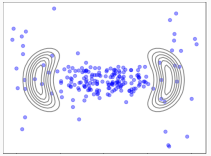
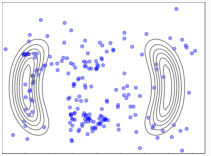
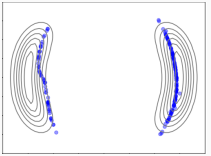
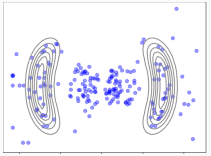
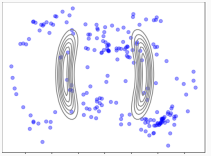
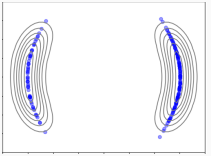
Evolution of particles under ULA, MALA and BRWP for the bimodal distribution, with step-size $\eta = 0.01$. The parameter of T was taken to be $T = 0.01$ for BRWP.

Sampling Behavior (Double Banana)



Evolution of particles under ULA, MALA and BRWP for the bimodal distribution, with step-size $\eta = 0.01$. The parameter of T was taken to be $T = 0.01$ for BRWP.

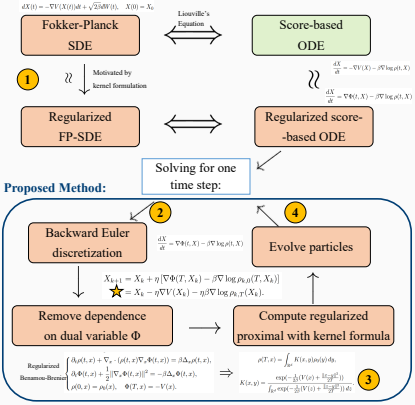
Large Step-Size Regime

Iter	ULA	MALA	BRWP ($T = 0.1$)	BRWP ($T = 0.2$)
10				
100	Diverged			

Convergent behavior reappears for large step-sizes when regularization parameter is large

Summary

- We propose a novel deterministic sampling method based on the regularized Wasserstein proximal
- Fully characterized stationary distribution, convergence behavior, and asymptotic bias for quadratic potentials
- Outstanding: asymptotic theory? General convergence rates? Approximation errors?



arXiv:2308.14945

Definition of Wasserstein-2 Distance

Definition
For two probability density functions μ, η on $\mathbb{R}^d$ with finite second moment, the *Wasserstein-2* distance between μ and η is

$$\mathcal{W}(\mu, \eta) \coloneqq \left(\inf_{\gamma \in \Gamma(\mu,\eta)} \iint_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \gamma(x, y)\, dx\, dy \right)^{1/2},$$

where the norm is the Euclidean norm, and the infimum is taken over all couplings between μ, η, i.e. γ is a joint probability measure on $\mathbb{R}^d \times \mathbb{R}^d$ with

$$\int_{\mathbb{R}^d} \gamma(x, y)\, dy = \mu(x), \quad \int_{\mathbb{R}^d} \gamma(x, y)\, dx = \eta(y).$$

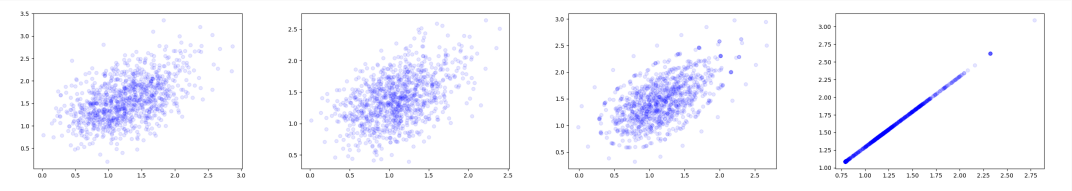
Gaussian: closed-form recurrence

Proposition
X_{k+1} is Gaussian with mean μ_{k+1} and covariance Σ_{k+1} given by

$$\tilde{\Sigma}_{k+1}^{-1} = \left(2\beta T(I + T\Sigma^{-1})^{-1} + (I + T\Sigma^{-1})^{-1}\Sigma_k(I + T\Sigma^{-1})^{-1}\right)^{-1}, \tag{10a}$$

$$\begin{aligned}\mu_{k+1} &= (I - \eta\Sigma^{-1})\mu_k + (\eta\beta\tilde{\Sigma}_{k+1}^{-1})(\mu_k - \tilde{\mu}_{k+1}) \\ &= \left(I - \eta\Sigma^{-1} + \eta\beta\left(2\beta TI + \Sigma_k(I + T\Sigma^{-1})^{-1}\right)^{-1}(T\Sigma^{-1})\right)\mu_k,\end{aligned} \tag{10b}$$

$$\Sigma_{k+1} = (I - \eta\Sigma^{-1} + \eta\beta\tilde{\Sigma}_{k+1}^{-1})\Sigma_k(I - \eta\Sigma^{-1} + \eta\beta\tilde{\Sigma}_{k+1}^{-1})^\top. \tag{10c}$$



(a) ULA **(b)** MALA **(c)** BRWP $T = 0.025$ **(d)** BRWP $T = 0.1$

Figure 1: Plots of the samples of θ after 4000 iterations, with $N = 1000$ samples. Parameters are $\alpha = 0.5, \eta = 0.05$. For this particular instantiation, we find that $\theta^* \approx (1.16, 1.45)$. We observe that for small T, we have a teardrop shaped structure. For large T, we have mode collapse in one direction.

Dataset	BRWP	AIG	WGF	SVGD
Boston	$3.309_{\pm 5.31e-1}$	$2.871_{\pm 3.41e-3}$	$3.077_{\pm 5.52e-3}$	$\mathbf{2.775_{\pm 3.78e-3}}$
Combined	$\mathbf{3.975_{\pm 3.94e-2}}$	$4.067_{\pm 9.27e-1}$	$4.077_{\pm 3.85e-4}$	$4.070_{\pm 2.02e-4}$
Concrete	$4.478_{\pm 2.05e-1}$	$\mathbf{4.440_{\pm 1.34e-1}}$	$4.883_{\pm 1.93e-1}$	$4.888_{\pm 1.39e-1}$
Kin8nm	$\mathbf{0.089_{\pm 6.06e-6}}$	$0.094_{\pm 5.56e-6}$	$0.096_{\pm 3.36e-5}$	$0.095_{\pm 1.32e-5}$
Wine	$0.623_{\pm 1.35e-3}$	$0.606_{\pm 1.40e-5}$	$0.614_{\pm 3.48e-4}$	$\mathbf{0.604_{\pm 9.89e-5}}$

Table 2: Test root-mean-square-error (RMSE). Bold indicates smallest in row.