SAMPLING WITH MOLLIFIED INTERACTION ENERGY DESCENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Sampling from a target measure whose density is only known up to a normalization constant is a fundamental problem in computational statistics and machine learning. In this paper, we present a new optimization-based method for sampling called *mollified interaction energy descent* (MIED). MIED minimizes a new class of energies on probability measures called *mollified interaction energies* (MIEs). These energies rely on mollifier functions—smooth approximations of the Dirac delta originated from PDE theory. We show that as the mollifier approaches the Dirac delta, the MIE converges to the chi-square divergence with respect to the target measure and the gradient flow of the MIE agrees with that of the chi-square divergence. Optimizing this energy with proper discretization yields a practical first-order particle-based algorithm for sampling in both unconstrained and constrained domains. We show experimentally that for unconstrained sampling problems our algorithm performs on par with existing particle-based algorithms like SVGD, while for constrained sampling problems our method readily incorporates constrained optimization techniques to handle more flexible constraints with strong performance compared to alternatives.

1 INTRODUCTION

Sampling from an unnormalized probability density is a ubiquitous task in statistics, mathematical physics, and machine learning. While Markov chain Monte Carlo (MCMC) methods (Brooks et al., 2011) provide a way to obtain unbiased samples at the price of potentially long mixing times, variational inference (VI) methods (Blei et al., 2017) approximate the target measure with simpler (e.g., parametric) distributions at a lower computational cost. In this work, we focus on a particular class of VI methods that approximate the target measure using a collection of interacting particles. A primary example is Stein variational gradient descent (SVGD) proposed by Liu & Wang (2016), which iteratively applies deterministic updates to a set of particles to decrease the KL divergence to the target distribution.

While MCMC and VI methods have found great success in sampling from unconstrained distributions, they often break down for distributions supported in a constrained domain. Constrained sampling is needed when the target density is undefined outside a given domain (e.g., the Dirichlet distribution), when the target density is not integrable in the entire Euclidean space (e.g., the uniform distribution), or when we only want samples that satisfy certain inequalities (e.g., fairness constraints in Bayesian inference (Liu et al., 2021)). A few recent approaches (Brubaker et al., 2012; Byrne & Girolami, 2013; Liu & Zhu, 2018; Shi et al., 2021) extend classical sampling methods like Hamiltonian Monte Carlo (HMC) or SVGD to constrained domains. These extensions, however, typically contain expensive numerical subroutines like solving nonlinear systems of equations, and require explicit formulas for quantities such as Riemannian metric tensors or mirror maps to be derived on a case-by-case basis from the constraints.

We present an optimization-based method called *mollified interaction energy descent* (MIED) that minimizes *mollified interaction energies* (MIEs) for both unconstrained and constrained sampling. An MIE takes the form of a double integral of the quotient of a *mollifier*—smooth approximation of δ_0, the Dirac delta at the origin—over the target density properly scaled. Intuitively, minimizing an MIE balances two types of forces: attractive forces that drive the current measure towards the target density, and repulsive forces from the mollifier that prevents collapsing. We show that as the

mollifier converges to δ_0, the MIE converges to the χ^2 divergence to the target measure up to an additive constant and the gradient flow of the MIE agrees with that of the χ^2 divergence, providing theoretical basis for using MIED as a sampling method.

While mollifiers can be interpreted as kernels with diminishing bandwidths, our analysis is fundamentally different from that of SVGD where a fixed-bandwidth kernel is used to define a reproducing kernel Hilbert space (RKHS) on which the Stein discrepancy has a closed-form (Gorham & Mackey, 2017). Deriving a version of the Stein discrepancy for constrained domains is far from trivial and requires special treatment (Shi et al., 2021; Xu, 2021). In contrast, our energy has the same form for constrained and unconstrained domains and approximates the χ^2 divergence so long as the bandwidth is sufficiently small so that short-range interaction dominates the energy.

Algorithmically, we use first-order optimization to minimize MIEs discretized using particles. We introduce a log-sum-exp trick to neutralize the effect of arbitrary scaling of the mollifiers and the target density; this form also improves numerical stability significantly. Since we turn sampling into optimization, we can readily apply existing constrained sampling techniques such as reparameterization using a differentiable (not necessarily bijective) map or the dynamic barrier method by Gong & Liu (2021) to handle a generic differentiable inequality constraint. Our method is effective as it only uses first-order derivatives of both the target density and the inequality constraint (or the reparameterization map), enabling large-scale applications in machine learning (see e.g. Figure 4). For unconstrained sampling problems, we show MIED achieves comparable performance to particle-based algorithms like SVGD, while for constrained sampling problems, MIED demonstrates strong performance compared to alternatives while being more flexible with constraint-handling.

2 Related Works

KL gradient flow and its discretization for unconstrained sampling. The Wasserstein gradient flow of the Kullback-Leibler (KL) divergence has been extensively studied, and many popular sampling algorithms can be viewed as discretizations of the KL-divergence gradient flow. Two primary examples are Langevin Monte Carlo (LMC) and Stein variational gradient descent (SVGD). LMC simulates Langevin diffusion and can be viewed as a forward-flow splitting scheme for the KL-divergence gradient flow (Wibisono, 2018). At each iteration of LMC, particles are pulled along $-\nabla \log p$ where p is the target density, while random Gaussian noise is injected. In contrast with LMC, SVGD is a deterministic algorithm that updates a collection of particles using a combination of an attractive force involving $-\nabla \log p$ and a repulsive force among the particles; it can be viewed as a kernelized gradient flow of the KL divergence (Liu, 2017) or of the χ^2 divergence (Chewi et al., 2020). The connection to the continuous gradient flow in the Wasserstein space is fruitful for deriving sharp convergence guarantees for these sampling algorithms (Durmus et al., 2019; Balasubramanian et al., 2022; Korba et al., 2020; Salim et al., 2022).

Sampling in constrained domains. Sampling in constrained domains is more challenging compared to the unconstrained setting. Typical solutions are rejection sampling and reparameterization to an unconstrained domain. However, rejection sampling can have a high rejection rate when the constrained domain is small, while reparameterization maps need to be chosen on a case-by-case basis with a determinant-of-Jacobian term that can be costly to evaluate. Brubaker et al. (2012) propose a constrained version of HMC for sampling on implicit submanifolds, but their algorithm is expensive as they need to solve a nonlinear system equations for every integration step in each step of HMC. Byrne & Girolami (2013) propose geodesic Hamiltonian Monte Carlo for sampling on embedded manifolds, but they require explicit geodesic formulae. Zhang et al. (2020); Ahn & Chewi (2021) propose discretizations of the mirror-Langevin diffusion for constrained sampling provided that a mirror map is given to capture the constraint. Similarly, Shi et al. (2021) propose mirror SVGD that evolves particles using SVGD-like updates in the dual space defined by a mirror map. While these two methods have strong theoretical convergence guarantees, they are limited by the availability of mirror maps that capture the constraints. Liu et al. (2021) extend SVGD to incorporate a population constraint over the samples obtained; their setting is different from ours where the constraint is applied to every sample.

Pairwise interaction energies on particles. Pairwise interaction energies on particles take the form $E(\{x_i\}_{i=1}^N) = \sum_{i \neq j} W(x_i, x_j)$ for a kernel $W(x, y)$. The gradient flow of E on N particles can give rise to phenomena such as flocking and swarming, and a long line of mathematical research

focuses on mean-field convergence of the particles following the gradient flow of E to continuous solutions of certain PDEs as $N \to \infty$ (Carrillo et al., 2014; Serfaty, 2020). Of particular interest to sampling, a separate line of works summarized by Borodachov et al. (2019) demonstrates that for the hypersingular Riesz kernel $W(x, y) = \|x - y\|_2^{-s}$ with sufficiently big s, minimizing E over a compact set yields uniform samples as $N \to \infty$. Moreover, W can depend on p to obtain non-uniform samples distributed according to p. As the hypersingular Riesz kernel is not integrable, their analysis is entirely based on geometric measure theory and avoids any variational methods. In contrast, we take a different approach by using integrable kernels in MIEs through mollifiers.This enables us to apply variational analysis to show our convergence results. We compare our formulation with theirs further in Appendix B. Recently, Korba et al. (2021) propose sampling via descending on kernel Stein discrepancy which can also be viewed as a type of interaction energy, but like SVGD, it is limited to unconstrained sampling and can be slow as higher-order derivatives of $\log p$ are required. Current work Craig et al. (2022) propose approximating the χ^2 divergence with a functional different from ours that also involves a mollifier. The resulting algorithm needs to evaluate an integral containing the target density over the whole domain at each step which can be costly. Consequently their method is only applied to special target densities for which the integral has a closed-form. In comparison, our method can handle general target densities with cheap update in each step.

3 Mollified interaction energy

Notation. Let $X \subset \mathbf{R}^n$ be the domain we work in. We use $\mathcal{L}_n$ to denote the Lebesgue measure on $\mathbf{R}^n$ and write $\int f(x)\mathrm{d}x$ to denote integration with respect to $\mathcal{L}_n$. We will assume all measures are Borel. We use $\mathcal{H}_d$ to denote the d-dimensional Hausdorff measure. We use ω_N to denote a set of points $\{x_1, \ldots, x_N\} \subset \mathbf{R}^n$. For $x \in \mathbf{R}^n$, let δ_x be the Dirac delta measure at x and δ_{ω_N} be the empirical measure $\frac{1}{N}\sum_{k=1}^N \delta_{x_k}$. We denote $L^p(X) := \{f : X \to \mathbf{R}$ Lebesgue measurable with $\int_X f(x)^p \mathrm{d}x < 0\}$, and for $f \in L^p(X)$ we let $\|f\|_p := \left(\int_X f(x)^p \mathrm{d}x\right)^{1/p}$. We use $\|f\|_\infty$ to denote the essential supremum of $|f|$. For $f, g : \mathbf{R}^n \to \mathbf{R}$, we define convolution $(f * g)(x) := \int f(y)g(x-y)\mathrm{d}y$ provided this integral exists. We use $C^k(X)$ to denote continuously k-differentiable functions and $C^\infty(X)$ to indicate the smooth functions. The space of continuous k-differentiable functions with compact support on X is $C_c^k(X)$. We denote by $\mathcal{P}_2(X)$ the set of probability measures with bounded second moments. Given a Lebesgue measurable map $T : X \to X$ and $\mu \in \mathcal{P}_2(X)$, $T_\# \mu$ is the pushforward measure of μ by T.

Setup. The domain $X \subset \mathbf{R}^n$ where we constrain samples in is assumed to be full-dimensional, i.e., $\mathcal{L}_n(X) > 0$; the unconstrained case corresponds to $X = \mathbf{R}^n$. The target density, always denoted as p, is assumed to satisfy $p \in C^1(X)$, $p(x) > 0$ for all $x \in X$, and $0 < \int_X p(x)\mathrm{d}x < \infty$. Let μ^* be the target probability measure $\mu^*(B) := \int_B p(x)\mathrm{d}x / \int_X p(x)\mathrm{d}x$ for Borel $B \subset X$. Our goal is to sample from μ^*.

3.1 Mollifiers

Our framework relies on the notion of *mollifiers* originally introduced in Friedrichs (1944).

Definition 3.1. *We say* $\{\phi_\epsilon\}_{\epsilon>0} \subset C^1(\mathbf{R}^n)$ *is a family of* mollifiers *if it satisfies*

(a) For any $\epsilon > 0$, $x \in \mathbf{R}^n$, $\phi_\epsilon(x) \geq 0$ *and* $\phi_\epsilon(x) = \phi_\epsilon(-x)$.
(b) For any $\epsilon > 0$, $\|\phi_\epsilon\|_1 = 1$ *and* $\phi_\epsilon \in L^\infty(\mathbf{R}^n)$.
(c) For any $\delta > 0$, $p \in \{1, \infty\}$, $\lim_{\epsilon \to 0} \left\|\mathbf{1}_{\mathbf{R}^n \setminus B_\delta(0)} \phi_\epsilon\right\|_p = 0$.

As $\epsilon \to 0$, the distribution with density ϕ_ϵ converges to δ_0, and if $f \in L^p(\mathbf{R}^n)$ and is continuous, then convolution $\phi_\epsilon * f$ converges to f both pointwise (Proposition A.4) and in L^p (Proposition A.7).

Our definition is different from the one used in PDE theory (Hörmander, 2015), where some $\phi \in C_c^\infty(\mathbf{R}^n)$ is fixed and the family of mollifiers is generated by $\phi_\epsilon(x) := \epsilon^{-n}\phi(x/\epsilon)$. While in PDE theory mollifiers are typically used to improve the regularity of non-smooth functions, in our framework they are used to construct interaction energies that approximate the χ^2 divergence.

We will use the following mollifiers. In following definitions we include normalizing constants Z_ϵ that ensure $\|\phi_\epsilon\| = 1$. We do not write out Z_ϵ explicitly as they might not admit clean forms

and they are only relevant in theory but not in our practical algorithm. For $s > n$, the *s-Riesz family of mollifiers* is defined as $\phi^s_\epsilon(x) := (\|x\|_2^2 + \epsilon^2)^{-s/2}/Z^s_\epsilon$. The *Gaussian family of mollifiers* is defined as $\phi^g_\epsilon(x) := \exp\left(-\frac{\|x\|_2^2}{2\epsilon^2}\right)/Z^g_\epsilon$. The *Laplace family of mollifiers* is defined as $\phi^l_\epsilon(x) := \exp\left(-\frac{\|x\|_2}{\epsilon}\right)/Z^l_\epsilon$. Since $\{\phi^g_\epsilon\}$ and $\{\phi^l_\epsilon\}$ correspond to the densities of centered Gaussian and Laplace random variables, they satisfy Definition 3.1. Proposition A.1 proves that $\{\phi^s_\epsilon\}$ also satisfies Definition 3.1.

3.2 Mollified Interaction Energies

Definition 3.2. *Given a family of mollifiers $\{\phi_\epsilon\}_{\epsilon>0}$, define $W_\epsilon : X \times X \to \mathbf{R}_{\geq 0}$ by*

$$W_\epsilon(x, y) := \phi_\epsilon(x - y)(p(x)p(y))^{-1/2}. \tag{1}$$

Define the mollified interaction energy *(MIE), $\mathcal{E}_\epsilon : \mathcal{P}_2(X) \to [0, \infty]$, to be*

$$\mathcal{E}_\epsilon(\mu) := \iint W_\epsilon(x, y)\, \mathrm{d}\mu(x)\, \mathrm{d}\mu(y). \tag{2}$$

Intuitively, minimizing $\mathcal{E}_\epsilon(\mu)$ balances two forces: a repulsive force from $\phi_\epsilon(x - y)$ makes sure μ has a good spread, and an attractive force from $(p(x)p(y))^{-1/2}$ helps μ concentrate on high-density regions. The exponent $-1/2$ is chosen to balance the two forces so that, as we will show in Theorem 3.3, $\mathcal{E}_\epsilon(\mu)$ approximates the χ^2 divergence with respect to μ^* for small ϵ.

3.3 Convergence to χ^2-divergence

Recall that the χ^2-divergence between probability measures P, Q is defined by

$$\chi^2(Q \parallel P) := \begin{cases} \int \left(\frac{\mathrm{d}Q}{\mathrm{d}P} - 1\right)^2 \mathrm{d}P & \text{if } Q \ll P \\ \infty & \text{otherwise,} \end{cases} \tag{3}$$

where $Q \ll P$ denotes absolute continuity of Q with respect to P with Radon-Nikodym density $\mathrm{d}Q/\mathrm{d}P$.

Theorem 3.3. *Suppose $\mu \in \mathcal{P}_2(X)$ satisfies $\chi^2(\mu \parallel \mu^*) < \infty$. Then, $\mathcal{E}_\epsilon(\mu) < \infty$ for any $\epsilon > 0$. Furthermore,*

$$\lim_{\epsilon \to 0} \mathcal{E}_\epsilon(\mu) = \mathcal{E}(\mu) := \chi^2(\mu \parallel \mu^*) + 1.$$

We prove Theorem 3.3 in Appendix A.2 using the fact that convolution with mollifiers converges to the identity operator as $\epsilon \to 0$. In Remark A.8 we discuss extension of Theorem 3.3 to cases when X is not full-dimensional; in particular Theorem 3.3 still holds when X is "flat", i.e., has Hausdorff dimension $d < n$ and is contained in a d-dimensional linear subspace.

3.4 Convexity and Γ-convergence on compact X

We next study properties of the minimizers of MIEs: Does $\min_{\mu \in \mathcal{P}_2(X)} \mathcal{E}_\epsilon$ admit a unique minimum? If so, do the minima of $\{\mathcal{E}_\epsilon\}_{\epsilon>0}$ converge to μ^* as $\epsilon \to 0$?

A symmetric lower semicontinuous (l.s.c.) kernel K on $X \times X$ is *integrally strictly positive definite* (i.s.p.d.) (Borodachov et al., 2019) if for every finite signed Borel measure ν on X, the energy $\mathcal{E}_K(\nu) := \iint K(x, y)\, \mathrm{d}(\nu \times \nu)\,(x, y)$ is well-defined (i.e., the integrals over the negative and positive parts of ν are not both infinite), and $\mathcal{E}_K(\nu) \geq 0$ where the equality holds only if $\nu = 0$ on all Borel sets of X.

Proposition 3.4. *Suppose X is compact and $k_\epsilon(x, y) := \phi_\epsilon(x - y)$ is i.s.p.d. on X. Then,*

- *(a) The kernel $W_\epsilon(x, y)$ defined in (1) restricted to X is also i.s.p.d.*
- *(b) The functional $\mathcal{E}_\epsilon$ is strictly convex on the space of signed measures and attains a unique minimum on $\mathcal{P}_2(X)$.*

(c) *For each* $\epsilon > 0$*, let* $\mu_\epsilon^* = \arg\min_{\mu \in \mathcal{P}_2(X)} \mathcal{E}_\epsilon(\mu)$*. Suppose for every* $\epsilon > 0$*,* $\chi^2(\mu_\epsilon^* \parallel \mu^*) < \infty$ *so that* μ_ϵ^* *has density* p_ϵ*. Suppose further that the sequence* $\{p_\epsilon\}$ *is bounded, i.e.,* $\sup_{\epsilon>0} |p_\epsilon(x)| < M(x) < \infty$*, and is* asymptotically equicontinuous *(a.e.c.), meaning*

$$\forall t > 0, x \in X, \exists \delta(x,t) > 0, \eta(x,t) > 0 \text{ such that } \sup_{\substack{\epsilon \in (0, \eta(x,t)) \\ y \in B_{\delta(x,t)}(x)}} |p_\epsilon(y) - p_\epsilon(x)| \le t.$$

Then $\mu_\epsilon^* \to \mu^*$ *weakly and* $\lim_k \mathcal{E}_\epsilon(\mu_\epsilon^*) = \mathcal{E}(\mu^*)$.

We prove Proposition 3.4 in Appendix A.3. Proposition 3.4 provides basis for minimizing $\mathcal{E}_\epsilon$ for a small ϵ, since its unique minimum will be a good approximation of μ^*.

3.5 Differential calculus of $\mathcal{E}_\epsilon$ in $\mathcal{P}_2(\mathbf{R}^n)$

We next study the gradient flow of $\mathcal{E}_\epsilon$ in Wasserstein space $\mathcal{P}_2(\mathbf{R}^n)$ for $X = \mathbf{R}^n$. Understanding the gradient flow of a functional often provides insights on convergence of algorithms that simulates gradient flow with time discretization (Ambrosio et al., 2005, Chapter 11) or spatial discretization (Chizat, 2022). Proofs of this section are given in Appendix A.4.

Let $\mathcal{F} : \mathcal{P}_2(\mathbf{R}^n) \to \mathbf{R}$ be a functional. The *Wasserstein gradient flow* of $\mathcal{F}$ (Ambrosio et al., 2005, Definition 11.1.1) is defined as the solution $\{\mu_t\}_{t \ge 0}$ of the PDE: $\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot (\mu_t \boldsymbol{w}_{\mathcal{F},\mu_t})$ where $\boldsymbol{w}_{\mathcal{F},\mu} \in L^2(\mu; \mathbf{R}^n)$ is a *Frechét subdifferential* of $\mathcal{F}$ at μ. Intuitively, gradient flows capture the evolution of the variable being optimized if we were to do gradient descent[1] on $\mathcal{F}$ when the step sizes go to 0. We next show that the gradient flow of $\mathcal{E}_\epsilon$ agrees with that of $\chi^2(\cdot \parallel \mu^*)$ in the sense that their subdifferentials coincide as $\epsilon \to 0$.

Proposition 3.5. *Assume* $\mu \in \mathcal{P}_2(\mathbf{R}^n)$ *has density* $q \in C_c^1(\mathbf{R}^n)$[2]*. Then any strong Frechét subdifferential* (16) *of* $\mathcal{E}_\epsilon$ *at* μ *takes the form*

$$\boldsymbol{w}_{\epsilon,\mu}(x) = 2\nabla \left(p(x)^{-1/2} (\phi_\epsilon * q/\sqrt{p})(x) \right), \text{ for } \mu\text{-a.e. } x \in \mathbf{R}^n. \tag{4}$$

On the other hand, any strong Frechét subdifferential of $\chi^2(\cdot \parallel \mu^*)$ *at* μ *takes the form*

$$\boldsymbol{w}_{\chi^2,\mu}(x) = 2\nabla \frac{q(x)}{p(x)}, \text{ for } \mu\text{-a.e. } x \in \mathbf{R}^n. \tag{5}$$

Moreover, if for sufficiently small ϵ*,* ϕ_ϵ *has compact support independent of* ϵ*, then*

$$\lim_{\epsilon \to 0} \boldsymbol{w}_{\epsilon,\mu}(x) = \boldsymbol{w}_{\chi^2,\mu}(x), \text{ for } \mu\text{-a.e. } x \in \mathbf{R}^n. \tag{6}$$

While simulating χ^2 divergence gradient flow is often intractable (Trillos & Sanz-Alonso, 2020, Section 3.3), our MIE admits a straightforward practical algorithm (Section 4). Chewi et al. (2020) show that SVGD can be viewed as a kernelized gradient flow of χ^2 divergence. However it is unclear how to choose the kernel to recover precisely the χ^2-divergence gradient flow. On the other hand, they show the exponential decay of KL divergence of χ^2-divergence gradient flow if the target measure satisfies a Poincaré inequality.

We next show that $\mathcal{E}_\epsilon$ is *displacement convex* at $\mu^* \in \mathcal{P}_2(\mathbf{R}^n)$ as $\epsilon \to 0$. This hints that gradient flow initialized near μ^* will have fast convergence.

Proposition 3.6. *Suppose* $p \in C_c^2(\mathbf{R}^n)$*. Suppose that for sufficiently small* ϵ*,* ϕ_ϵ *has compact support independent of* ϵ*. Assume* $k_\epsilon(x,y) := \phi_\epsilon(x-y)$ *is i.s.p.d. Let* $\boldsymbol{\xi} \in C_c^\infty(\mathbf{R}^n; \mathbf{R}^n)$ *and* $\mu_t := (\boldsymbol{I} + t\boldsymbol{\xi})_\# \mu^*$*. Then* $\lim_{\epsilon \to 0} \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=0} \mathcal{E}_\epsilon(\mu_t) \ge 0$.

[1] The more precise notion is of minimizing movements (Ambrosio et al., 2005, Chapter 2).

[2] We assume that μ has compact support because W_ϵ can be unbounded and cause integrability issues due to the presence of $(p(x)p(y))^{-1/2}$ term.

4 A PRACTICAL SAMPLING ALGORITHM

We now present a first-order particle-based algorithm for constrained and unconstrained sampling by minimizing a discrete version of (2). Substituting an empirical distribution for μ in (2) gives the discrete mollified interaction energy

$$E_\epsilon(\omega_N) := \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi_\epsilon(x_i - x_j)(p(x_i)p(x_j))^{-1/2} \tag{7}$$

for N particles $\omega_N = \{x_1, \dots, x_N\}$.

Denote $\omega^*_{N,\epsilon} \in \arg\min_{\omega_N \subset X} E_\epsilon(\omega_N)$ and $\mu^*_\epsilon = \arg\min_{\mu\in\mathcal{P}_2(X)} \mathcal{E}_\epsilon(\mu)$, If X is compact, by Borodachov et al. (2019, Corollary 4.2.9), we have weak convergence $\delta_{\omega^*_{N,\epsilon}} \to \mu^*_\epsilon$ as $N \to \infty$. If in addition the assumptions of Proposition 3.4 hold, we have weak convergence $\mu^*_\epsilon \to \mu^*$. This shows that minimizing (7) with a large N and a small ϵ will result in an empirical distribution of particles that approximates μ^*. Our sampling method, mollified interaction energy descent (MIED), is simply running gradient descent on (7) over particles. Below we address a few practical concerns.

Optimization in the logarithmic domain. In practical applications, we only have access to the unnormalized target density p. The normalizing constant of the mollifier ϕ_ϵ can also be hard to compute (e.g., for the Riesz family). While these normalization constants do not affect the minima of in (7), they can still affect gradient step sizes during optimization. Moreover, ϕ_ϵ can be very large when ϵ is small, and in many Bayesian applications p can be tiny and only $\log p$ is numerically significant. To address these issues, we optimize the logarithm of (7) using the log-sum-exp trick (Blanchard et al., 2021) to improve numerical stability and to get rid of the arbitrary scaling of the normalizing constants:

$$\log E_\epsilon(\omega_N) := \log\sum_{i=1}^{N}\sum_{j=1}^{N}\exp\left(\log\phi_\epsilon(x_i - x_j) - \frac{1}{2}(\log p(x_i) + \log p(x_j))\right) - 2\log N. \tag{8}$$

Special treatment of the diagonal terms. Since ϕ_ϵ goes to the Dirac delta as $\epsilon \to 0$, the discretization of (2) on a neighborhood of the diagonal $\{(x, y) : x = y\}$ needs to be handled with extra care. In (7) the diagonal appears as $\sum_{i=1}^N \phi_\epsilon(0)p(x_i)^{-1}$ which can dominant the summation when ϵ is small and then $\phi_\epsilon(0)$ becomes too large. For mollifiers we consider, we use a different diagonal term $\sum_{i=1}^N \phi_\epsilon(h_i/\kappa_n)p(x_i)^{-1}$ where $h_i := \min_{j\neq i}\|x_i - x_j\|$ and $\kappa_n \geq 1$ is a constant depending only on the dimension n. Since $0 \leq \phi_\epsilon(h_i/\kappa) \leq \phi_\epsilon(0)$, the energy obtained will be bounded between (7) and the version of (7) without the diagonal terms. Hence by the proof of Theorem 4.2.2 of Borodachov et al. (2019), the discrete minimizers of E_ϵ still converge to μ^*_ϵ as $N \to \infty$ and $\epsilon \to 0$. Empirically we found the choice $\kappa_n = (1.3n)^{1/n}$ work well for the Riesz family of mollifiers and we use the Riesz family primarily for our experiments.

Constraint handling. If $X \neq \mathbf{R}^n$, we need to minimize (8) subject to constraints $\omega_N = \{x_1, \dots, x_N\} \subset X$. Since the constraint is the same for each particle x_i, we want our algorithm to remain parallelizable across particles.

We consider two types of constraints.

(a) There exists a differentiable map $f : \mathbf{R}^n \to X$ that so that $\mathcal{L}_n(X \setminus f(\mathbf{R}^n)) = 0$.
(b) The set X is given by $\{x \in \mathbf{R}^n : g(x) \leq 0\}$ for a differentiable $g : \mathbf{R}^n \to \mathbf{R}$.

For (a), we reduce the problem to unconstrained optimization in $\mathbf{R}^n$ using objective $\log E_\epsilon(f(\omega_N))$ with $\omega_N = \{x_1, \dots, x_N\} \subset \mathbf{R}^n$ and $f(\omega_N) := \{f(x_1), \dots, f(x_N)\}$. For (b), we apply the dynamic barrier method by Gong & Liu (2021) to particles in parallel. In our implementation, for both types of constraints, we use Adam optimizer (Kingma & Ba, 2014) that modulates the update directions for each particle using history of the moments of the gradients.

5 EXPERIMENTS

We compare MIED with recent alternatives on unconstrained and constrained sampling problems. Unless mentioned otherwise, we choose the s-Riesz family of mollifiers $\{\phi^s_\epsilon\}$ with $s = n+10^{-4}$ and

$\epsilon = 10^{-8}$: we found minimizing the MIE with such mollifiers results in well-separated particles so that we can take ϵ to be very small as our theory recommends. This is not the case for the Gaussian or the Laplace family as setting ϵ too small can cause numerical issues even when particles are well-separated.

5.1 UNCONSTRAINED SAMPLING

Gaussians in varying dimensions. We first compare MIED with SVGD (Liu & Wang, 2016) and KSDD (Korba et al., 2021) on Gaussians of varying dimensions and with different number of particles. In Figure 1, we see that as the number of dimensions increases, MIED results in best samples in terms of W_2 distance (Wasserstein-2 distance computed using linear programming) with respect to 10^4 i.i.d. samples, while SVGD yields lower energy distance (Székely & Rizzo, 2013). We think this is because MIED results in more evenly spaced samples (Figure 5) so the W_2 distance is lower, but it is biased since $\epsilon > 0$ so for energy distance SVGD performs better. More details can be found in Appendix C.1.

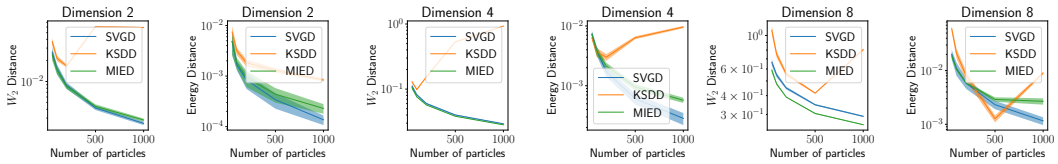

Figure 1: Gaussian experiments results in varying dimensions and different numbers of particles. For each method we plot the metric (W_2 distance or energy distance) versus the number of particles, averaged over 10 trials (shaded region indicates standard deviation).

Product of two Student's t-distributions. Next, we consider a 2-dimensional distribution constructed as the product of two independent t-distributions with degree of freedom $\nu = 2$ composed with a linear transform. Unlike Gaussians, Student's t-distributions have heavy tails. On the left of Figure 2, we visualize the results of each method with 1000 samples after 2000 iterations. MIED captures the heavy tail while SVGD fails. Quantitatively, on the right of Figure 2, while SVGD captures the distribution in $[-a, a]^2$ better for $a \leq 3$, MIED yields lower metrics for a bigger a.

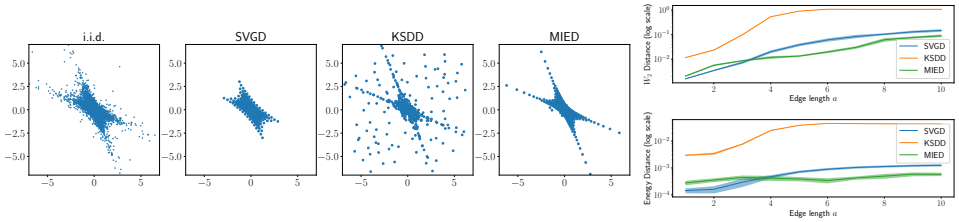

Figure 2: Left: visualization of samples from each method for the product of two Student's t-distribution (composed with a linear transform). Right: metrics of each method when restricting to $[-a, a]^2$. As t-distributions have heavy tail, if we draw i.i.d. samples from the true product distribution, a small number of them will have large norms, making the computation of metrics unstable. Thus we restrict both i.i.d. samples and the resulting samples from each method to $[-a, a]^2$ before computing the metrics for $a \in [2, 10]$.

Bayesian logistic regression. We compare MIED with SVGD and KSDD for the Bayesian logistic regression setting in (Liu & Wang, 2016; Korba et al., 2021). We further include a naïve baseline, independent particle descent (IPD), which simply runs gradient descent independently on each particle to maximize the posterior probability. In addition to using test accuracy as the metric, we include W_2 distance and energy distance between the samples from each method and 10^4 samples from NUTS (Hoffman et al., 2014) after sufficient burn-in steps. We summarize the results in Table 5.1. All methods, including the naïve baseline IPD, are comparable in terms of test accuracy. In other words, accuracy is not a good metric for comparing the quality of posterior samples. Bayesian

inference is typically preferred over maximum a posteriori estimation for its ability to capture uncertainty. When evaluating the quality of the uncertainty of the samples using distances between distributions, MIED provides the best approximation in terms of the W_2 distance and all methods (except IPD) are comparable in terms of the energy distance.

Dataset	NUTS	IPD	SVGD	KSDD	MIED
banana ($d = 3$)	0.55	-6.14/-3.58/**0.55**	-7.81/-5.24/**0.55**	**-8.24/-5.76/0.55**	-7.37/-5.06/**0.55**
breast cancer ($d = 10$)	0.64	-1.51/-1.03/**0.60**	-1.62/-2.06/**0.60**	-1.71/**-2.23/0.60**	**-1.99**/-2.18/**0.60**
diabetis ($d = 9$)	0.78	-2.18/-1.55/**0.77**	-3.09/-3.42/**0.77**	-2.91/**-3.90/0.77**	**-3.11**/-3.13/**0.77**
flare solar ($d = 10$)	0.59	3.30/2.65/0.48	6.91/4.09/0.52	**1.77/-0.08/0.55**	7.09/4.25/0.48
german ($d = 21$)	0.65	-1.80/-1.25/**0.65**	-1.89/-2.63/0.64	-1.27/**-2.83/0.65**	**-1.96**/-2.80/**0.65**
heart ($d = 14$)	0.87	-0.40/-0.56/**0.87**	-0.41/-1.50/**0.87**	-0.10/**-1.76/0.87**	**-0.92**/-1.67/**0.87**
image ($d = 19$)	0.82	6.53/4.31/**0.83**	7.17/4.01/**0.83**	2.16/-0.50/0.82	**1.14/-1.88**/0.82
ringnorm ($d = 21$)	0.77	-3.82/-2.45/**0.77**	**-4.11/-5.98/0.77**	1.07/-2.21/0.76	-4.03/-5.70/**0.77**
splice ($d = 61$)	0.85	**-1.47**/-1.18/**0.85**	-1.22/**-2.65/0.85**	2.04/-0.05/0.84	1.45/0.70/0.82
thyroid ($d = 6$)	0.84	1.95/0.53/**0.84**	1.17/-0.00/**0.84**	2.42/1.57/0.74	**0.84/-0.37/0.84**
titanic ($d = 4$)	0.40	**-1.59**/-0.16/**0.40**	-0.46/-0.31/**0.40**	-0.63/-0.39/**0.40**	-1.00/**-0.45/0.40**
twonorm ($d = 21$)	0.97	-1.21/-1.13/**0.97**	-1.32/-2.78/**0.97**	1.55/-0.62/**0.97**	**-1.44/-3.21/0.97**
waveform ($d = 22$)	0.77	-2.67/-1.87/**0.78**	-2.98/**-5.23/0.78**	-2.60/-4.18/0.77	**-3.09**/-3.17/**0.78**

Table 1: Bayesian logistic regression results with 1000 particles. We include the test accuracy for NUTS in the second column. Three numbers A/B/C in the following columnes are logarithmic W_2 distance, logarithmic energy distance, and test accuracy. Bold indicates the best numbers. We use $80\%/20\%$ training/test split. All methods are run with identical initialization and learning rate 0.01. Results are reported after 10^4 iterations.

5.2 Constrained Sampling

Uniform sampling in 2D. We consider uniform sampling in the square $[-1, 1]^2$ with 500 particles. We reparameterize our particles using $\tanh$ to eliminate the constraint and show results with various choices of mollifiers—we always choose the smallest ϵ while the optimization remains stable. We compare our method with SVMD and MSVGD proposed for constrained sampling by Shi et al. (2021) using mirror map $\phi(\theta) = \sum_{i=1}^{n} \left(\log \frac{1}{1-\theta_i} + \log \frac{1}{1+\theta_i} \right)$ as in Ahn & Chewi (2021). To demonstrate that SVGD and KSDD break down in constrained domains, we implement these two methods adapted to the constrained setting using the same reparameterization as our method. The initial particles are drawn uniformly from $[-0.5, 0.5]^2$ for all methods. The left plot of Figure 3

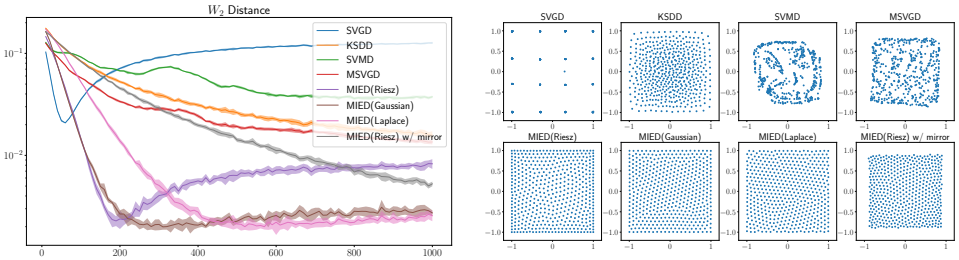

Figure 3: Convergence of metrics and visualization of samples for each method for uniform sampling from a 2D box.

shows that quantitatively our method achieves much lower energy distance and W_2 distance (compared to a 5000 i.i.d. samples uniformly drawn in $[-1, 1]^2$.). On the right of Figure 3, we visualize the samples from each method. We see that samples from SVGD collapse and samples from KSDD are non-uniform (we choose kernel size 0.1 for KSDD that produces the best result; otherwise it will collapse like SVGD). SVMD and MSVGD fail to draw samples near the boundary because they use mapping $\nabla\phi^*(\eta) = (\sqrt{1+\eta^2}-1)/\eta$ (coordinate-wise arithmetic) to push unconstrained samples to $[-1, 1]^2$, yet coordinates of η need to go to ∞ to make $\nabla\phi^*(\eta)$ land close to the boundary. We verify this phenomenon by using $\nabla\phi^*(\eta)$ as the reparametrization map in MIED: indeed as shown the the bottom right of Figure 3, with such reparameterization MIED also fails to capture samples near the boundary. When using $\tanh$ as the reparameterization map, MIED produces decent results for all three families of mollifiers (Riesz, Gaussian, Laplace). This highlight the flexibility of our

method with constraint handling, whereas for SVMD and MSVGD, the mirror map has to chosen carefully to be the gradient of a strongly convex function while capturing the constraints.

In Appendix C.2, we conduct a similar comparison for sampling from a 20-dimensional Dirichlet distribution using the same setup as Shi et al. (2021). In scenarios where a good choice of the mirror map is available, SVMD and MSVGD can obtain better performance compared to MIED.

Fairness Bayesian neural networks. We train fair bayesian neural networks to predict whether the annual income of a person is at least $\$50,000$ with gender as the protected attribute following the same setup as in Liu et al. (2021) where we are given a dataset $\mathcal{D} = \{x^{(i)}, y^{(i)}, z^{(i)}\}_{i=1}^{|\mathcal{D}|}$ consisting of feature vectors $x^{(i)}$, labels y^i (whether the income is $\geq \$50,000$), and genders $z^{(i)}$ (protected attribute). The target density is taken to be the posterior of logistic regression with a two-layer Bayesian neural network with weights θ, and we put a standard Gaussian prior on each entry of θ independently. Given $t > 0$, the fairness constraint is $g(\theta) = (\mathrm{cov}_{(x,y,z)\sim\mathcal{D}}[z, \hat{y}(x;\theta)])^2 - t \leq 0$. We use $80\%/20\%$ training/test split as in Liu et al. (2021). On the left of Figure 4, we plot the trade-off curve of the result obtained using our method and the methods from Liu et al. (2021) for $t \in \{10^{-5}, 10^{-4}, 0.0001, 0.001, 0.002, 0.005, 0.01\}$. We use the source code provided by Liu et al. (2021) with default hyperparameters.[3] All methods use 2000 iterations for training, and for our method we use learning rate 0.001. One of their four methods (Control+SVGD) got stuck at initialization (with accuracy around 0.75), so we omit its result from the plot. Our method recovers a much larger Pareto front compared to the alternatives. On the right of Figure 4, we visualize the curves of the energy and the covariance versus the number of training iterations: as expected we see a smaller t results in bigger MIE (lower log-likelihood) and smaller covariance between the prediction and the protected attribute.

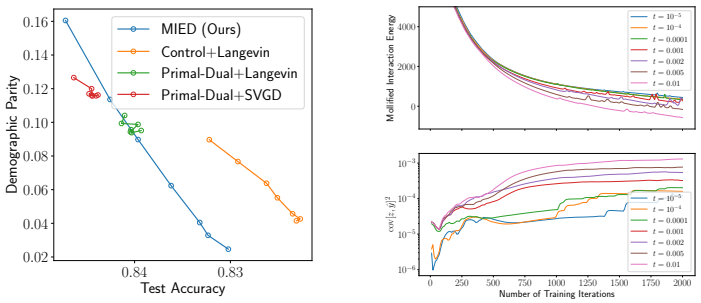

Figure 4: Left: trade-off curves of demographic parity versus test accuracy for MIED and methods from Liu et al. (2021). Right: MIEs and $(\mathrm{cov}_{(x,y,z)\sim\mathcal{D}}[z, \hat{y}(x;\theta)])^2$ (measured on the training data) versus the number of training iterations, for various t.

6 Conclusion

We present a new sampling method by minimizing MIEs discretized as particles for unconstrained and constrained sampling. This is motivated by the insight that MIEs converge to χ^2 divergence with respect to the target measure as the mollifier parameter goes to 0. The proposed method achieves promising results on the sampling problems we consider.

Below we highlight three limitations. First, as discussed in Remark A.8, our theory only applies when the domain is full-dimensional or flat. Extending our theory to handle cases where the domain is an arbitrary d-rectifiable set is an important next step as it allows handling of more complicated constraints such as nonlinear equality constraints. Secondly, when $\epsilon > 0$, the minimizer of MIE can be different from the target mesaure. Finding a way to debias MIE (e.g., like how Sinkhorn distances are debiased (Feydy et al., 2019)) is an interesting direction. Lastly, the connection between the minimizers of the discretized MIE (7) and those of the continuous MIE (2) is only established in the limit as $N \to \infty$. We hope to investigate how well the empirical distribution of particles minimizing (7) approximates the target measure when N is finite.

[3]The source code provided by the authors of Liu et al. (2021) does not implement the calculation of $g(\theta)$ faithfully as written in the formula, so we modified it.

Reproducibility statement. We provide self-contained proofs in Appendix A for the theoretical results stated in the main text.

We have included the complete source code in the supplementary necessary to reproduce all our experiments.

REFERENCES

Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-langevin algorithm. *Advances in Neural Information Processing Systems*, 34:28405–28418, 2021.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Krishna Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for langevin monte carlo. In *Conference on Learning Theory*, pp. 2896–2923. PMLR, 2022.

Pierre Blanchard, Desmond J Higham, and Nicholas J Higham. Accurately computing the log-sum-exp and softmax functions. *IMA Journal of Numerical Analysis*, 41(4):2311–2330, 2021.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Sergiy V Borodachov, Douglas P Hardin, and Edward B Saff. *Discrete energy on rectifiable sets*. Springer, 2019.

Andrea Braides et al. *Gamma-convergence for Beginners*, volume 22. Clarendon Press, 2002.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

Marcus Brubaker, Mathieu Salzmann, and Raquel Urtasun. A family of mcmc methods on implicitly defined manifolds. In *Artificial intelligence and statistics*, pp. 161–172. PMLR, 2012.

Simon Byrne and Mark Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

José Antonio Carrillo, Young-Pil Choi, and Maxime Hauray. The derivation of swarming models: mean-field limit and wasserstein distances. In *Collective dynamics from bacteria to crowds*, pp. 1–46. Springer, 2014.

Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, and Philippe Rigollet. Svgd as a kernelized wasserstein gradient flow of the chi-squared divergence. *Advances in Neural Information Processing Systems*, 33:2098–2109, 2020.

Lenaic Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 194(1):487–532, 2022.

Katy Craig, Karthik Elamvazhuthi, Matt Haberland, and Olga Turanova. A blob method method for inhomogeneous diffusion with applications to multi-agent control and sampling. *arXiv preprint arXiv:2202.12927*, 2022.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of langevin monte carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690. PMLR, 2019.

Kurt Otto Friedrichs. The identity of weak and strong extensions of differential operators. *Transactions of the American Mathematical Society*, 55(1):132–151, 1944.

Chengyue Gong and Xingchao Liu. Bi-objective trade-off with dynamic barrier gradient descent. *NeurIPS 2021*, 2021.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pp. 1292–1301. PMLR, 2017.

Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Lars Hörmander. *The analysis of linear partial differential operators I: Distribution theory and Fourier analysis*. Springer, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for stein variational gradient descent. *Advances in Neural Information Processing Systems*, 33:4672–4682, 2020.

Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin. Kernel stein discrepancy descent. In *International Conference on Machine Learning*, pp. 5719–5730. PMLR, 2021.

Chang Liu and Jun Zhu. Riemannian stein variational gradient descent for bayesian inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Qiang Liu. Stein variational gradient descent as gradient flow. *Advances in neural information processing systems*, 30, 2017.

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

Xingchao Liu, Xin Tong, and Qiang Liu. Sampling with trusthworthy constraints: A variational gradient framework. *Advances in Neural Information Processing Systems*, 34:23557–23568, 2021.

Luc Pronzato and Anatoly Zhigljavsky. Minimum-energy measures for singular kernels. *Journal of Computational and Applied Mathematics*, 382:113089, 2021.

Adil Salim, Lukang Sun, and Peter Richtarik. A convergence theory for svgd in the population limit under talagrand's inequality t1. In *International Conference on Machine Learning*, pp. 19139–19152. PMLR, 2022.

Sylvia Serfaty. Mean field limit for coulomb-type flows. *Duke Mathematical Journal*, 169(15): 2887–2935, 2020.

Jiaxin Shi, Chang Liu, and Lester Mackey. Sampling with mirrored stein operators. *arXiv preprint arXiv:2106.12506*, 2021.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

Nicolas Garcia Trillos and Daniel Sanz-Alonso. The bayesian update: variational formulations and gradient flows. *Bayesian Analysis*, 15(1):29–56, 2020.

Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pp. 2093–3027. PMLR, 2018.

Wenkai Xu. Generalised kernel stein discrepancy (gksd): A unifying approach for non-parametric goodness-of-fit testing. *arXiv preprint arXiv:2106.12105*, 2021.

Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror langevin monte carlo. In *Conference on Learning Theory*, pp. 3814–3841. PMLR, 2020.

A Detailed Analysis

A.1 Preliminaries on mollifiers

Proposition A.1. *For $s > n$, the s-Riesz family of mollifiers, defined as $\phi_\epsilon^s(x) := (\|x\|_2^2 + \epsilon^2)^{-s/2}/Z_\epsilon^s$, satisfies Definition 3.1.*

In order to prove Proposition A.1, we first prove a lemma.

Lemma A.2. *Let $b \geq 0$. Then for any $\epsilon > 0$,*

$$Z_\epsilon^s \int_{B_\epsilon(0)} \phi_\epsilon^s(y)\|y\|_2^b \mathrm{d}y = \mathcal{H}_{n-1}(S^{n-1}) \left(\int_0^1 \frac{t^{n+b-1}}{(t^2+1)^{s/2}} \right) \epsilon^{n+b-s}, \tag{9}$$

where $\mathcal{H}_{n-1}(S^{n-1})$ is the volume of the $(n-1)$-dimensional sphere.

Furthermore, assuming $s > b + n$, then for any $\epsilon > 0, \delta > 0$,

$$Z_\epsilon^s \int_{\mathbf{R}^n \setminus B_\delta(0)} \phi_\epsilon^s(y)\|y\|_2^b \mathrm{d}y \leq \mathcal{H}_{n-1}(S^{n-1}) \frac{\delta^{n+b-s}}{s-(n+b)}. \tag{10}$$

Proof. For (9), we compute

$$\begin{aligned} Z_\epsilon^s \int_{B_\epsilon(0)} \phi_\epsilon^s(y)\|y\|_2^b \mathrm{d}y &= \int_{B_\epsilon(0)} \frac{\|y\|_2^b}{(\|y\|_2^2+\epsilon^2)^{s/2}} = \mathcal{H}_{n-1}(S^{n-1}) \int_0^\epsilon \frac{r^{n+b-1}}{(r^2+\epsilon^2)^{s/2}} \mathrm{d}r \\ &= \mathcal{H}_{n-1}(S^{n-1}) \int_0^1 \frac{(\epsilon t)^{n+b-1}}{(\epsilon^2 t^2+\epsilon^2)^{s/2}} \epsilon \mathrm{d}t \\ &= \mathcal{H}_{n-1}(S^{n-1}) \left(\int_0^1 \frac{t^{n+b-1}}{(t^2+1)^{s/2}} \mathrm{d}t \right) \epsilon^{n+b-s}, \end{aligned}$$

where we use substitution $r = \epsilon t$ on the second line.

If $s > b + n$, then

$$\begin{aligned} Z_\epsilon^s \int_{\mathbf{R}^n \setminus B_\delta(0)} \phi_\epsilon^s(0,y)\|y\|_2^b \mathrm{d}y &= \int_{\mathbf{R}^n \setminus B_\delta(0)} \frac{\|y\|_2^b}{(\|y\|_2^2+\epsilon^2)^{s/2}} \\ &\leq \int_{\mathbf{R}^n \setminus B_\delta(0)} \frac{\|y\|_2^b}{\|y\|_2^s} = \mathcal{H}_{n-1}(S^{n-1}) \int_\delta^\infty r^{n-1+b-s} \\ &= \mathcal{H}_{n-1}(S^{n-1}) \frac{\delta^{n+b-s}}{s-(n+b)}, \end{aligned}$$

where the last integral uses $s > b + n$. □

Proof of Proposition A.1. It is clear that (a) of Proposition A.1 holds for $\phi_\epsilon^s(x)$. Since $\phi_\epsilon^s(x)$ is bounded we have $\phi_\epsilon^s \in L^\infty(\mathbf{R}^n)$. For any $\delta > 0$, we have, for $\epsilon \leq \delta$,

$$Z_\epsilon^s \int_{B_\delta(0)} \phi_\epsilon^s(x)\mathrm{d}x \geq Z_\epsilon^s \int_{B_\epsilon(0)} \phi_\epsilon^s(x)\mathrm{d}x = C\epsilon^{n-s},$$

where we use (9) with $b = 0$ and C is a constant depending only on n, s. On the other hand, using (10) with $b = 0$,

$$Z_\epsilon^s \int_{\mathbf{R}^n \setminus B_\delta(0)} \phi_\epsilon^s(x)\mathrm{d}x \leq C'\delta^{n-s},$$

where C' depends only on n, s. With $\delta = \epsilon$, we see that $Z_\epsilon^s \phi_\epsilon^s \in L^1(\mathbf{R}^n)$ so (b) is satisfied.

Since δ is fixed and $s > n$, we see that

$$\lim_{\epsilon \to 0} \frac{\int_{B_\delta(0)} \phi_\epsilon^s(x)\mathrm{d}x}{\int_{\mathbf{R}^n \setminus B_\delta(0)} \phi_\epsilon^s(x)\mathrm{d}x} \geq \lim_{\epsilon \to 0} \frac{C\epsilon^{n-s}}{C'\delta^{n-s}} = \infty.$$

Since $\int_{B_\delta(0)} \phi_\epsilon^s(x)\mathrm{d}x + \int_{\mathbf{R}^n \setminus B_\delta(0)} \phi_\epsilon^s(x)\mathrm{d}x = 1$, we have shown (c). □

In the rest of this section, we assume $\{\phi_\epsilon\}_{\epsilon>0}$ is a family of mollifiers satisfying Definition 3.1.

Lemma A.3. *For any $p \in [1,\infty]$, for any $\delta > 0$, $\phi_\epsilon \in L^p(\mathbf{R}^n)$ and $\lim_{\epsilon\to 0} \left\|\mathbf{1}_{\mathbf{R}^n\setminus B_\delta(0)}\phi_\epsilon\right\|_p = 0$. holds.*

Proof. Assume $p \notin \{1,\infty\}$ since both cases are covered in the assumptions. Then for any $\delta > 0$, by Hölder's inequality,

$$\|\phi_\epsilon\|_p^p = \left\|\phi_\epsilon \cdot \phi_\epsilon^{p-1}\right\|_1 \le \|\phi_\epsilon\|_1 \left\|\phi_\epsilon^{p-1}\right\|_\infty = \|\phi_\epsilon\|_1 \|\phi_\epsilon\|_\infty^{p-1} < \infty.$$

Similarly,

$$\begin{aligned}\left\|\mathbf{1}_{\mathbf{R}^n\setminus B_\delta(0)}\phi_\epsilon\right\|_p^p &= \left\|\mathbf{1}_{\mathbf{R}^n\setminus B_\delta(0)}\phi_\epsilon^p\right\|_1 = \left\|\phi_\epsilon \cdot \mathbf{1}_{\mathbf{R}^n\setminus B_\delta(0)}\phi_\epsilon^{p-1}\right\|_1 \\ &\le \|\phi_\epsilon\|_1 \left\|\mathbf{1}_{\mathbf{R}^n\setminus B_\delta(0)}\phi_\epsilon^{p-1}\right\|_\infty = \left\|\mathbf{1}_{\mathbf{R}^n\setminus B_\delta(0)}\phi_\epsilon\right\|_\infty^{p-1}.\end{aligned}$$

Letting $\epsilon \to 0$ and applying (c) gives $\lim_{\epsilon\to 0} \left\|\mathbf{1}_{\mathbf{R}^n\setminus B_\delta(0)}\phi_\epsilon\right\|_p = 0$. □

Proposition A.4. *Let $f \in L^p(\mathbf{R}^n)$, $p \in [1,\infty]$. Then for every $\epsilon > 0$, the integral*

$$\int f(x-y)\phi_\epsilon(y)\mathrm{d}y$$

exists, so that $(f * \phi_\epsilon)(x)$ is finite. Moreover, if f is continous at x, then

$$\lim_{\epsilon\to 0}(f * \phi_\epsilon)(x) = f(x). \tag{11}$$

Proof. Observe that by Hölder's inequality, for q such that $1/p + 1/q = 1$ (allowing infinity),

$$\int |f(x-y)\phi_\epsilon(y)|\mathrm{d}y \le \|f\|_p \|\phi_\epsilon\|_q < \infty.$$

Hence $(f * \phi_\epsilon)(x)$ is integrable and finite.

For any $\epsilon > 0$, note that

$$|(f * \phi_\epsilon)(x) - f(x)| = \left|\int f(x-y)\phi_\epsilon(y)\mathrm{d}y - f(x)\right|.$$

Since $\int \phi_\epsilon(x)\mathrm{d}x = 1$, we have

$$|(f * \phi_\epsilon)(x) - f(x)| = \left|\int \left(f(x-y) - f(x)\right)\phi_\epsilon(y)\mathrm{d}y\right| \le \int |f(x-y) - f(x)|\phi_\epsilon(y)\mathrm{d}y.$$

Fix $t > 0$. Continuity of f at x implies there exists $\delta > 0$ such that $|f(x-y) - f(x)| < t$ for all $y \in B_\delta(0)$. Then

$$\int_{B_\delta(0)} |f(x-y) - f(x)|\phi_\epsilon(y)\mathrm{d}y \le t\int_{B_\delta(0)} \phi_\epsilon(y)\mathrm{d}y \le t.$$

On the other hand, by Hölder's inequality,

$$\int_{\mathbf{R}^n\setminus B_\delta(x)} |f(x-y) - f(x)|\phi_\epsilon(y)\mathrm{d}y \le \|\tau_x f - f\|_p \left\|\mathbf{1}_{\mathbf{R}^n\setminus B_\delta(0)}\phi_\epsilon\right\|_q \le 2\|f\|_p \left\|\mathbf{1}_{\mathbf{R}^n\setminus B_\delta(0)}\phi_\epsilon\right\|_q.$$

Hence

$$|(f * \phi_\epsilon)(x) - f(x)| \le t + 2\|f\|_p \left\|\mathbf{1}_{\mathbf{R}^n\setminus B_\delta(0)}\phi_\epsilon\right\|_q.$$

By Lemma A.3, since $\|f\|_p < \infty$, taking $\epsilon \to 0$ we get, for any $t > 0$,

$$\limsup_{\epsilon\to 0} |(f * \phi_\epsilon)(x) - f(x)| \le t.$$

Now let $t \to 0$ we get (11). □

Corollary A.5. *Let $f \in L^p(\mathbf{R}^n)$, $p \in [1, \infty]$, be a continuous function. If $\{f_\epsilon\}$ is a sequence of functions that converge uniformly to f, then for every x,*

$$\lim_{\epsilon \to 0} (f_\epsilon * \phi_\epsilon)(x) = f(x). \tag{12}$$

Proof. Note that

$$|(\phi_\epsilon * f_\epsilon)(x) - f(x)| \leq |(\phi_\epsilon * f_\epsilon)(x) - (\phi_\epsilon * f)(x)| + |(\phi_\epsilon * f)(x) - f(x)|.$$

Proposition A.4 shows the second term goes to 0 as $\epsilon \to 0$. For the first term, we have

$$\begin{aligned} |(\phi_\epsilon * f_\epsilon)(x) - (\phi_\epsilon * f)(x)| &= \left| \int (f_\epsilon(x-y) - f(x-y))\, \phi_\epsilon(y) \mathrm{d}y \right| \\ &\leq \sup_x |f_\epsilon(x) - f(x)| \to 0 \end{aligned}$$

by uniform convergence. □

Lemma A.6. *For $f \in L^p(\mathbf{R}^n)$, $p \in [1, \infty)$, we have*

$$\lim_{y \to 0} \|\tau_y f - f\|_p = 0,$$

where we use $\tau_a f$ to denote the translated function $\tau_a f(x) := f(x - a)$.

Proof. Fix $\epsilon > 0$. It is a standard fact that $C_c(\mathbf{R}^n)$ is dense in $L^p(\mathbf{R}^n)$. Hence there exists $g \in C_c(\mathbf{R}^n)$ such that $\|f - g\|_p < \epsilon$. Since g is continuous with compact support, it is uniform continuous. Then there exists $\delta > 0$ with $|g(x-y) - g(x)| < \epsilon^{1/p}/\mathcal{L}_n(K)$ if $y \in B_\delta(x)$. Hence for such y we have $\|\tau_y g - g\|_p^p < \epsilon$ with $\mathcal{L}_n(K) < \infty$. Thus

$$\|\tau_y f - f\|_p \leq \|\tau_y f - \tau_y g\|_p + \|\tau_y g - g\|_p + \|g - f\|_p \leq 3\epsilon.$$

□

Proposition A.7. *Let $f \in L^p(\mathbf{R}^n)$, $p \in (1, \infty)$. Then for every $\epsilon > 0$,*

$$\|f * \phi_\epsilon\|_p \leq 3\|f\|_p.$$

In particular, $f * \phi_\epsilon \in L^p(\mathbf{R}^n)$. Moreover,

$$\lim_{\epsilon \to 0} \|f * \phi_\epsilon - f\|_p = 0.$$

Proof. For every $\epsilon > 0$, we have

$$\begin{aligned} \|f * \phi_\epsilon - f\|_p^p &= \int \left| \int (f(x-y) - f(x))\phi_\epsilon(y) \mathrm{d}y \right|^p \mathrm{d}x \\ &\leq \int \int |f(x-y) - f(x)|^p \phi_\epsilon(y) \mathrm{d}y \mathrm{d}x \\ &= \int \left(\int |f(x-y) - f(x)|^p \mathrm{d}x \right) \phi_\epsilon(y) \mathrm{d}y \\ &= \int \|\tau_y f - f\|_p^p \phi_\epsilon(y) \mathrm{d}y, \end{aligned}$$

where we use Jensen's inequality with the observation that $\phi_\epsilon(y)\mathrm{d}y$ is a probability measure and Tonelli's theorem to exchange the order of integration. To show the first claim, note that,

$$\|f * \phi_\epsilon - f\|_p^p \leq \int \|\tau_y f - f\|_p^p \phi_\epsilon(y) \mathrm{d}y \leq \int (2\|f\|_p)^p \phi_\epsilon(y) \mathrm{d}y = (2\|f\|_p)^p.$$

Hence $\|f * \phi_\epsilon\|_p \leq \|f\|_p + \|f * \phi_\epsilon - f\|_p \leq 3\|f\|_p$.

For the second claim, by Lemma A.6, the function $y \mapsto \|\tau_y f - f\|_p^p$ is continuous at $y = 0$ with limit 0. Hence by Proposition A.4, we are done by taking $\epsilon \to 0$. □

A.2 CONVERGENCE TO χ^2-DIVERGENCE

Proof of Theorem 3.3. Since $\chi^2(\mu \parallel \mu^*) < \infty$, we know $\mu \ll \mu^* \ll \mathcal{L}_n$, so let $q(x)$ the density of μ. We set $q(x) = 0$ for $x \in \mathbf{R}^n \setminus X$. We compute using Tonelli's theorem (since our integrand is positive):

$$\begin{aligned}\mathcal{E}_\epsilon(\mu) &= \iint \phi_\epsilon(x-y)(p(x)p(y))^{-1/2} q(x)q(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int \left(\phi_\epsilon * \frac{q}{\sqrt{p}}\right)(x)\frac{q}{\sqrt{p}}(x)\,\mathrm{d}x = \int (\phi_\epsilon * g)(x)g(x)\,\mathrm{d}x,\end{aligned}$$

where we define $g(x) := q(x)/\sqrt{p(x)}$ on X and $g(x) := 0$ for $x \notin X$. Moreover,

$$\begin{aligned}\chi^2(\mu \parallel \mu^*) + 1 &= \int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\mu^*}(x) - 1\right)^2 \mathrm{d}\mu^*(x) + 1 = \int_X \left(\frac{q(x)}{p(x)} - 1\right)^2 p(x)\,\mathrm{d}x + 1 \\ &= \int_X \frac{q(x)^2}{p(x)}\mathrm{d}x - 2\int_X q(x)\,\mathrm{d}x + 2 = \int g(x)^2\,\mathrm{d}x.\end{aligned}$$

The assumption $\chi^2(\mu \parallel \mu^*) < \infty$ then implies $g \in L^2(\mathbf{R}^n)$. By Proposition A.7, we have $\phi_\epsilon * g \in L^2(\mathbf{R}^n)$. Next notice that

$$\left|\mathcal{E}_\epsilon(\mu) - \left(\chi^2(\mu \parallel \mu^*) + 1\right)\right| \le \int |(\phi_\epsilon * g)(x) - g(x)|g(x)\,\mathrm{d}x \le \|\phi_\epsilon * g - g\|_2 \|g\|_2 < \infty.$$

This shows the first claim. Finally, the last expression goes to 0 as $\epsilon \to 0$ since $\|\phi_\epsilon * g - g\|_2 \to 0$ by Proposition A.7. □

Remark A.8. *It is possible to prove versions of Theorem 3.3 when X has Hausdorff dimension $d < n$: in such cases χ^2-divergence still makes sense as does (2). When X is "flat", i.e., with Hausdorff dimension d and contained in a d-dimensional linear subspace of $\mathbf{R}^n$, e.g., when X is defined by a set of linear inequalities, then Theorem 3.3 follows if we adapt the assumptions of Definition 3.1 and calculation in the proof of Theorem 3.3 to be in the subspace. For a general d-dimensional X, a similar calculation yields $\mathcal{E}_\epsilon(\mu) = \int_X \left(\int_X \phi_\epsilon(x-y)\left(\frac{q}{\sqrt{p}}\right)(y)\mathrm{d}\mathcal{H}_d(y)\right)\left(\frac{q}{\sqrt{p}}\right)(x)\,\mathrm{d}\mathcal{H}_d(x)$. For a similar argument to go through, we will need $\int_X \phi_\epsilon(x-y)\mathrm{d}\mathcal{H}_d(y)$ to be independent of x. This is true when X is a d-dimensional sphere, but for a general X the last integral will depend on the base point x. Proving a version of Theorem 3.3 for X with Hausdorff dimension $d < n$ is an interesting future direction.*

A.3 CONVEXITY AND Γ-CONVERGENCE ON COMPACT X

Recall that we say a functional $\mathcal{F} : \mathcal{P}_2(X) \to (-\infty, +\infty]$ is *proper* if there exists $\mu \in \mathcal{P}_2(X)$ such that $\mathcal{F}(\mu) < \infty$, and is *lower semicontinuous* (l.s.c.) if for any sequence $\mu_k \to \mu$ (in W_2 metric), we have $\liminf_{k\to\infty} \mathcal{F}(\mu_k) \ge \mathcal{F}(\mu)$.

Lemma A.9. *For any $\epsilon > 0$, the functional $\mathcal{E}_\epsilon : \mathcal{P}_2(X) \to (-\infty, +\infty]$ is proper and l.s.c. Moreover, if X is compact, the minimum $\min_{\mu\in\mathcal{P}_2(X)} \mathcal{E}_\epsilon(\mu)$ is attained by some measure in $\mathcal{P}_2(X)$.*

Proof. The continuity of $k_\epsilon(x, y) = \phi_\epsilon(x - y)$ and $p(x)$ implies that W_ϵ defined in (1) is continuous. Then by the nonnegativity of $\mathcal{E}_\epsilon$, the argument from Example 9.3.4 of Ambrosio et al. (2005) concludes that $\mathcal{E}_\epsilon$ is proper and l.s.c.

The set of probability distributions $\mathcal{P}_2(X)$ is tight by the compactness of X. It is closed since if $\{\mu_k\} \subset \mathcal{P}_2(X)$ weakly converges to μ, then $\mu(X) \ge \limsup \mu_k(X) = 1$ so that $\mu \in \mathcal{P}_2(X)$. Hence by Prokhorov's theorem (Theorem 5.1.3 (Ambrosio et al., 2005)), $\mathcal{P}_2(X)$ is compact. It is then an elementary result that any l.s.c. function attains its minimum on a compact set. □

Proof of Proposition 3.4(a)(b). Let μ be any finite signed Borel measure. The compactness assumption and the fact that $p \in C^1(X)$ imply $p(x) > \delta$ for some $\delta > 0$. Hence $p(x)^{-1/2} \le \delta^{-1/2}$, so the weighted measure $\tilde{\mu}$ defined by $\mathrm{d}\tilde{\mu} := p^{-1/2} \cdot \mathrm{d}\mu$ is also finite. By the definition of i.s.p.d., we

have $\mathcal{E}_{W_\epsilon}(\mu) = \mathcal{E}_{k_\epsilon}(\tilde{\mu}) > 0$ if $\tilde{\mu}$ is not the zero measure, which is equivalent to μ not being zero since $p > 0$. Thus W_ϵ is i.s.p.d. Also note that $(p(x)p(y))^{-1/2}$ for all $x, y \in X$, so $\mathcal{E}_{W_\epsilon}$ is always finite. By Lemma 1.1 of Pronzato & Zhigljavsky (2021), we conclude that $\mathcal{E}_{W_\epsilon}$ is strictly convex in the space of signed measures, and in particular on $\mathcal{P}_2(X)$. Hence combined with the existence result from Lemma A.9 we conclude $\mathcal{E}_\epsilon$ attains a unique minimum in $\mathcal{P}_2(X)$. □

Proof of Proposition 3.4(c) . We will show that any subsequence of $\{\mu_\epsilon^*\}$ (resp. $\{\mathcal{E}_\epsilon(\mu_\epsilon^*)\}$) has a further convergence subsequence converging to μ^* (resp. $\mathcal{E}(\mu^*)$), which will give the claim. With a slight abuse of notation, we assume $\{\mu_\epsilon^*\}$ is already some subsequence of the original sequence. As argued in the proof of Lemma A.9, $\mathcal{P}_2(X)$ is compact with respect to weak convergence. Hence $\{\mu_\epsilon^*\}$ has a weakly convergence subsequence $\mu_{\epsilon_k}^* \to \nu$ for some ν. By Theorem 1 of Braides et al. (2002), the boundedness and a.e.c. assumptions imply that p_{ϵ_k} uniformly converges to a continous function p_ν that is the density of ν. Hence by Fatou's lemma, we compute

$$\begin{aligned}
\liminf_k \mathcal{E}_{\epsilon_k}(\mu_{\epsilon_k}^*) &= \liminf_k \iint_{X\times X} \phi_{\epsilon_k}(x-y)(p(x)p(y))^{-1/2} p_{\epsilon_k}(x)p_{\epsilon_k}(y)\mathrm{d}x\mathrm{d}y \\
&\geq \int_X \liminf_k \left(\int_X \phi_{\epsilon_k}(x-y)p(y)^{-1/2}p_{\epsilon_k}(y)\mathrm{d}y\right) p(x)^{-1/2}p_\nu(x)\mathrm{d}x \\
&= \int_X p(x)^{-1}p_\nu(x)^2\mathrm{d}x = \mathcal{E}(\nu),
\end{aligned}$$

where we use the fact that $\liminf_k a_k b_k \geq (\liminf_k a_k)(\liminf_k b_k)$ for non-negative numbers $\{a_k\}, \{b_k\}$, and Corollary A.5 since $p^{-1/2}p_{\epsilon_k} \in L^2(X)$ because $\chi^2(\mu_{\epsilon_k}^* \parallel \mu^*) < \infty$ (see the proof of Theorem 3.3). On the other hand, the previous inequality shows $\mathcal{E}(\nu) < \infty$, so by Theorem 3.3,

$$\mathcal{E}(\nu) = \lim_k \mathcal{E}_{\epsilon_k}(\nu) \geq \limsup_k \mathcal{E}_{\epsilon_k}(\mu_{\epsilon_k}^*),$$

where the last inequality comes from the fact that $\mu_{\epsilon_k}^* \in \arg\min_{\mu\in\mathcal{P}_2(X)} \mathcal{E}_{\epsilon_k}(\mu)$. Thus we have shown $\liminf_k \mathcal{E}_{\epsilon_k}(\mu_{\epsilon_k}^*) \geq \limsup_k \mathcal{E}_{\epsilon_k}(\mu_{\epsilon_k}^*)$, which means all inequalities are equalities, so that

$$\lim_k \mathcal{E}_{\epsilon_k}(\mu_{\epsilon_k}^*) = \mathcal{E}(\nu).$$

Optimality of $\mu_{\epsilon_k}^*$ implies

$$\mathcal{E}(\nu) = \lim_k \mathcal{E}_{\epsilon_k}(\mu_{\epsilon_k}^*) \leq \liminf_k \mathcal{E}_{\epsilon_k}(\mu^*) = \mathcal{E}(\mu^*).$$

By the optimality of μ^* we have $\mathcal{E}(\nu) = \mathcal{E}(\mu^*)$, and ν minimizes $\mathcal{E}$. In this case, $\chi^2(\nu \parallel \mu^*) = 0$. This can only happen when ν and μ^* agree on all Borel sets, so we conclude $\nu = \mu^*$. □

A.4 Differential calculus of $\mathcal{E}_\epsilon$ in $\mathcal{P}_2(\mathbf{R}^n)$

Lemma A.10. *Let $x_0 \in \mathbf{R}$, $h > 0$, and $\Omega \subset \mathbf{R}^m$ be a compact set. Suppose $f : (x_0 - h, x_0 + h) \times \Omega \to \mathbf{R}$ is jointly continuous and the derivative $\frac{\partial}{\partial x} f : (x_0 - h, x_0 + h) \times \Omega \to \mathbf{R}$ exists and is jointly continuous. Then $\int_\Omega f(x,\omega)\mathrm{d}\omega$ is differentiable for $x \in (x_0 - h, x_0 + h)$, and*

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_\Omega f(x,\omega)\mathrm{d}\omega = \int_\Omega \frac{\partial}{\partial x} f(x,\omega)\mathrm{d}\omega$$

where the integration is with respect to the Lebesgue measure $\mathcal{L}_m$ on Ω.

Proof. Fix $x \in (x_0 - h, x_0 + h)$ and let $t > 0$ be small enough such that $[x - t, x + t] \subset (x_0 - h, x_0 + h)$. Note that $\int_\Omega f(x,\omega)\mathrm{d}\omega$ is well-defined by the dominated convergence theorem since $\sup_{x\in[x-t,x+t],\omega\in\Omega} f(x,\omega) < \infty$ and $\mathcal{L}_m(\Omega) < \infty$. Define $\theta(\omega) := \sup_{x\in[x-t,x+t]} \left|\frac{\partial}{\partial x} f(x,\omega)\right|$ which is finite since $\frac{\partial}{\partial x} f(x,\omega)$ is continuous, so that $\theta(\omega) \geq \frac{\partial}{\partial x}|f(x,\omega)|$ for all $x \in [x - t, x + t]$ and θ is integrable since $\mathcal{L}_m(\Omega) < \infty$. Hence by the differentiation lemma (Theorem 6.28 of Klenke (2013)) we are done. □

The following lemma is similar to Korba et al. (2021, Proposition 3) but with different assumptions: we do not put any integrability assumptions on W_ϵ, but we do restrict measures to have compact support. For a differentiable $f : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$, we use $\nabla_1 f(x, y)$ to denote the gradient with respect to x. We also use $\mathrm{H}_1 f(x, y)$ to denote the Hessian of f with respect to x. We similarly denote $\nabla_2 f(x, y)$ and $\mathrm{H}_2 f(x, y)$.

Lemma A.11. *Assume $\mu \in \mathcal{P}_2(\mathbf{R}^n)$ has density $q \in C_c^1(\mathbf{R}^n)$. Let $\boldsymbol{\xi} \in C_c^\infty(\mathbf{R}^n; \mathbf{R}^n)$. Denote $\boldsymbol{s}(x, t) := x + t\boldsymbol{\xi}(x)$. Then for all $t > 0$ and all $\epsilon > 0$,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}_\epsilon(\mu_t) = 2 \iint \left(\boldsymbol{\xi}(x)^\top \nabla_1 W_\epsilon(\boldsymbol{s}(x, t), \boldsymbol{s}(y, t))\right) \mathrm{d}\mu(x)\mathrm{d}\mu(y). \tag{13}$$

In particularly,

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \mathcal{E}_\epsilon(\mu_t) = 2 \iint \left(\boldsymbol{\xi}(x)^\top \nabla_1 W_\epsilon(x, y)\right) \mathrm{d}\mu(x)\mathrm{d}\mu(y). \tag{14}$$

Moreover,

$$\left.\frac{\mathrm{d}^2}{\mathrm{d}t^2}\right|_{t=0} \mathcal{E}_\epsilon(\mu_t) = 2 \iint \left(\boldsymbol{\xi}(x)^\top \nabla_1 \nabla_2 W_\epsilon(x, y)\boldsymbol{\xi}(y) + \boldsymbol{\xi}(x)^\top \mathrm{H}_1 W_\epsilon(x, y)\boldsymbol{\xi}(x)\right) \mathrm{d}\mu(x)\mathrm{d}\mu(y). \tag{15}$$

Proof. We compute

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}_\epsilon(\mu_t) = \frac{\mathrm{d}}{\mathrm{d}t}\left(2 \iint W_\epsilon(\boldsymbol{s}(x, t), \boldsymbol{s}(y, t))\mathrm{d}\mu(x)\mathrm{d}\mu(y)\right).$$

Since μ has compact support and $(x, y, t) \mapsto W_\epsilon(\boldsymbol{s}(x, t), \boldsymbol{s}(y, t))q(x)q(y)$ is jointly continuous and its derivative with respect to t is also jointly continuous, by Lemma A.10, we can push $\frac{\mathrm{d}}{\mathrm{d}t}$ inside the double integral and we obtain (13). Another application of Lemma A.10 shows that if we take derivative with respect to t again on (13) and evaluate at 0 we obtain (15). □

Lemma A.12. *Let $f \in C^1(\mathbf{R}^n)$, $p \in [1, \infty]$. Assume ϕ_ϵ has compact support for some $\epsilon > 0$. Then for all $i = 1, \ldots, n$, $\phi_\epsilon * f$ is differentiable and*

$$\frac{\partial}{\partial x_i}(\phi_\epsilon * f)(x) = \left(\phi_\epsilon * \frac{\partial}{\partial x_i} f\right)(x).$$

Proof. Since $\mathrm{supp}(\phi_\epsilon)$ is compact and $f \in C^1(\mathbf{R}^n)$ is bounded on any compact set, we know $\phi_\epsilon * f$ is well-defined at every $x \in \mathbf{R}^n$. Note that

$$\begin{aligned}\frac{\partial}{\partial x_i}(\phi_\epsilon * f)(x) &= \frac{\partial}{\partial x_i}\left(\int f(x - y)\phi_\epsilon(y)\mathrm{d}y\right) \\ &\overset{(?)}{=} \int \frac{\partial f}{\partial x_i}(x - y)\phi_\epsilon(y)\mathrm{d}y = \left(\phi_\epsilon * \frac{\partial}{\partial x_i} f\right)(x).\end{aligned}$$

Since $\mathrm{supp}(\phi_\epsilon)$ is compact and $(x, y) \mapsto f(x - y)\phi(y)$ is C^1, by Lemma A.10 we justify the existence of the derivative and the exchange of differentiation (?). □

Recall that a vector field $\boldsymbol{w} \in L^2(\mu; \mathbf{R}^n)$ is a *strong Fréchet subdifferential* of a functional $\mathcal{F} : \mathcal{P}_2(\mathbf{R}^n) \to (-\infty, +\infty]$ if for all $\boldsymbol{T} \in L^2(\mu; \mathbf{R}^n)$, the following holds (Definition 10.1.1 of Ambrosio et al. (2005)):

$$\mathcal{F}(\boldsymbol{T}_\# \mu) - \mathcal{F}(\mu) \geq \int \boldsymbol{w}(x)^\top (\boldsymbol{T}(x) - x)\mathrm{d}\mu(x) + o\left(\|\boldsymbol{T} - \boldsymbol{I}\|_{L^2(\mu; \mathbf{R}^n)}\right). \tag{16}$$

Note that we cannot apply Ambrosio et al. (2005, Lemma 10.4.1) directly to prove (4) because interaction energies cannot be written as (10.4.1) in their setup.

Proof of Proposition 3.5. Let $\boldsymbol{\xi} \in C_c^\infty(\mathbf{R}^n; \mathbf{R}^n)$. By Lemma A.11, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} \mathcal{E}_\epsilon(\mu_t) &= 2\iint \left(\boldsymbol{\xi}(x)^\top \nabla_1 W_\epsilon(x,y)\right) \mathrm{d}\mu(x)\mathrm{d}\mu(y) \\ &= 2\iint \left(\boldsymbol{\xi}(x)^\top \nabla_1 \left(\phi_\epsilon(x-y)(p(x)p(y))^{-1/2}\right)\right) q(y)\mathrm{d}y\mathrm{d}\mu(x) \\ &= 2\int \boldsymbol{\xi}(x)^\top \nabla \left(p(x)^{-1/2}(\phi_\epsilon * q/\sqrt{p})(x)\right) \mathrm{d}\mu(x),\end{aligned}$$

where the last step follows from applying Lemma A.10 since q has compact support. Now suppose $\boldsymbol{w} \in L^2(\mu; \mathbf{R}^n)$ is a strong Fréchet subdifferential satisfying (16). For the sequence $\{\boldsymbol{T}_t\}$, we have by definition

$$\begin{aligned}\mathcal{E}_\epsilon(\mu_t) - \mathcal{E}_\epsilon(\mu) &\geq \int \boldsymbol{w}(x)^\top (\boldsymbol{T}_t(x) - x)\mathrm{d}\mu(x) + o\left(\|\boldsymbol{T}_t - \boldsymbol{I}\|_{L^2(\mu;\mathbf{R}^n)}\right) \\ &= \int \boldsymbol{w}(x)^\top (t\boldsymbol{\xi}(x))\mathrm{d}\mu(x) + o(t).\end{aligned}$$

Hence

$$\liminf_{t\downarrow 0} \frac{\mathcal{E}_\epsilon(\mu_t) - \mathcal{E}_\epsilon(\mu)}{t} \geq \int \boldsymbol{w}(x)^\top \boldsymbol{\xi}(x)\mathrm{d}\mu(x) \geq \liminf_{t\uparrow 0} \frac{\mathcal{E}_\epsilon(\mu_t) - \mathcal{E}_\epsilon(\mu)}{t}.$$

The previous calculation shows $\frac{\mathrm{d}}{\mathrm{d}t}|_{t=0}\mathcal{E}_\epsilon(\mu_t)$ exists, and hence it is equal to $\int \boldsymbol{w}(x)^\top \boldsymbol{\xi}(x)\mathrm{d}\mu(x)$. This proves for any $\boldsymbol{\xi} \in C_c^\infty(\mathbf{R}^n; \mathbf{R}^n)$,

$$\int \boldsymbol{w}(x)^\top \boldsymbol{\xi}(x)\mathrm{d}\mu(x) = \int \boldsymbol{\xi}(x)^\top \nabla \left(2p(x)^{-1/2}(\phi_\epsilon * q/\sqrt{p})(x)\right) \mathrm{d}\mu(x).$$

Hence we have shown (4).

By Ambrosio et al. (2005, Lemma 10.4.1) with $F(x, \rho(x)) = \left(\frac{\rho(x)}{p(x)} - 1\right)^2 p(x)$, we find the strong subdifferential of $\chi^2(\cdot \parallel \mu^*)$ is the one in (5). To show (6), we compute for μ-a.e. x,

$$\begin{aligned}&\left\|\boldsymbol{w}_\epsilon(x) - \boldsymbol{w}_{\chi^2}(x)\right\|_2 \\ =&2\left\|\nabla \left(p^{-1/2}(x) \int \phi_\epsilon(y) \left(q(x-y)p^{-1/2}(x-y) - q(x)p(x)^{-1/2}\right)\right)\right\|_2 \\ \overset{(?)}{=}&2\left\|\int \phi_\epsilon(y)\nabla \left(q(x-y)p^{-1/2}(x-y)p^{-1/2}(x) - q(x)p(x)^{-1}\right)\right\|_2 \\ \leq&2\int \phi_\epsilon(y)\left\|\nabla \left(q(x-y)p^{-1/2}(x-y)p^{-1/2}(x) - q(x)p(x)^{-1}\right)\right\|_2,\end{aligned} \tag{17}$$

where for (?) we applied Lemma A.12. Now taking $\epsilon \to 0$, by Proposition A.4 and the fact that $\mathrm{supp}(\phi_\epsilon)$ is compact, we see the last expression goes to zero. □

The statement of Proposition 3.6 is similar in form as Korba et al. (2021, Corollary 4) but in our case we do not have the second term in Equation (15) vanishing and we need to take the limit $\epsilon \to 0$. We also do not resort to RKHS theory in the proof.

Proof of Proposition 3.6. By (15), we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\bigg|_{t=0} \mathcal{E}_\epsilon(\mu_t) = 2(\mathcal{F}_\epsilon + \mathcal{G}_\epsilon),$$

where

$$\begin{aligned}F_\epsilon &= \iint \left(\boldsymbol{\xi}(x)^\top \nabla_1\nabla_2 W_\epsilon(x,y)\boldsymbol{\xi}(y)\right) \mathrm{d}\mu^*(x)\mathrm{d}\mu^*(y) \\ G_\epsilon &= \iint \left(\boldsymbol{\xi}(x)^\top \mathrm{H}_1 W_\epsilon(x,y)\boldsymbol{\xi}(x)\right) \mathrm{d}\mu^*(x)\mathrm{d}\mu^*(y).\end{aligned}$$

We tackle F_ϵ first. Observe that successive application of integration by parts using the fact that $\boldsymbol{\xi}$ has compact support gives

$$\begin{aligned} F_\epsilon &= \sum_{i,j=1}^n \iint \boldsymbol{\xi}_i(x) \frac{\partial^2}{\partial x_i \partial y_i} W_\epsilon(x,y) \boldsymbol{\xi}_j(y) p(x) p(y) \mathrm{d}x\mathrm{d}y \\ &= -\sum_{i,j=1}^n \iint \frac{\partial}{\partial x_i}(\boldsymbol{\xi}_i(x)p(x)) \frac{\partial}{\partial y_i} W_\epsilon(x,y) \boldsymbol{\xi}_j(y) p(y) \mathrm{d}x\mathrm{d}y \\ &= \sum_{i,j=1}^n \iint \frac{\partial}{\partial x_i}(\boldsymbol{\xi}_i(x)p(x)) W_\epsilon(x,y) \frac{\partial}{\partial y_i}(\boldsymbol{\xi}_j(y)p(y)) \mathrm{d}x\mathrm{d}y. \end{aligned}$$

For each i, if we view $\frac{\partial}{\partial x_i}(\boldsymbol{\xi}_i(x)p(x))$ as the density of a signed measure (it is integrable since it has compact support), and since W_ϵ is i.s.p.d. by Proposition 3.4(a), we see that each double integral in the last expression is non-negative. Hence $F_\epsilon \geq 0$.

Next we show $\lim_{\epsilon\to 0} G_\epsilon = 0$. Since μ^* has compact support, by Fubini's theorem,

$$G_\epsilon = \int \boldsymbol{\xi}(x)^\top \mathrm{H}_1 \left(\int W_\epsilon(x,y) p(y) \mathrm{d}y \right) \boldsymbol{\xi}(x) p(x) \mathrm{d}x.$$

To expand the integral inside the Hessian operator, we have

$$\mathrm{H}_1 \left(\int W_\epsilon(x,y) p(y) \mathrm{d}y \right) = \mathrm{H} \left(p(x)^{-1/2} (\phi_\epsilon * \sqrt{p})(x) \right).$$

First by the chain rule and Lemma A.12, we have

$$\frac{\mathrm{d}}{\mathrm{d}x} \left(p(x)^{-1/2} (\phi_\epsilon * \sqrt{p})(x) \right) = \frac{\mathrm{d}}{\mathrm{d}x} \left(p(x)^{-1/2} \right) (\phi_\epsilon * \sqrt{p}) + p(x)^{-1/2} \left(\phi_\epsilon * \frac{\mathrm{d}}{\mathrm{d}x} \sqrt{p}(x) \right).$$

Differentiating again while applying Lemma A.12, we obtain after rearranging terms,

$$\begin{aligned} \mathrm{H}_1 \left(\int W_\epsilon(x,y) p(y) \mathrm{d}y \right) &= \left(\frac{\mathrm{d}^2}{\mathrm{d}x^2} p(x)^{-1/2} \right) (\phi_\epsilon * \sqrt{p})(x) \\ &\quad + 2 \left(\frac{\mathrm{d}}{\mathrm{d}x} p(x)^{-1/2} \right) \left(\phi_\epsilon * \frac{\mathrm{d}}{\mathrm{d}x} \sqrt{p} \right)(x) \\ &\quad + p(x)^{-1/2} \left(\phi_\epsilon * \frac{\mathrm{d}^2}{\mathrm{d}x^2} \sqrt{p} \right)(x). \end{aligned} \tag{18}$$

By Proposition A.4 and the fact that $p \in C_c^2(\mathbf{R}^n)$, we have

$$\begin{aligned} &\lim_{\epsilon\to 0} \mathrm{H}_1 \left(\int W_\epsilon(x,y) p(y) \mathrm{d}y \right) \\ =& \left(\frac{\mathrm{d}^2}{\mathrm{d}x^2} p(x)^{-1/2} \right) \sqrt{p}(x) + 2 \left(\frac{\mathrm{d}}{\mathrm{d}x} p(x)^{-1/2} \right) \frac{\mathrm{d}}{\mathrm{d}x} \sqrt{p}(x) + p(x)^{-1/2} \frac{\mathrm{d}^2}{\mathrm{d}x^2} \sqrt{p}(x) \\ =& \frac{\mathrm{d}^2}{\mathrm{d}x^2} \left(p(x)^{-1/2} \sqrt{p(x)} \right) = 0. \end{aligned}$$

Finally, we have

$$\begin{aligned} \lim_{\epsilon\to 0} G_\epsilon &= \lim_{\epsilon\to 0} \int \boldsymbol{\xi}(x)^\top \mathrm{H}_1 \left(\int W_\epsilon(x,y) p(y) \mathrm{d}y \right) \boldsymbol{\xi}(x) p(x) \mathrm{d}x \\ &= \int \boldsymbol{\xi}(x)^\top \left(\lim_{\epsilon\to 0} \mathrm{H}_1 \left(\int W_\epsilon(x,y) p(y) \mathrm{d}y \right) \right) \boldsymbol{\xi}(x) p(x) \mathrm{d}x \\ &= 0, \end{aligned}$$

where interchanging the limit and the integral is jusfitied by the dominated convergence theorem and the fact that p and ϕ_ϵ have compact support (we need compact support assumption of ϕ_ϵ to make sure convolutions appearing in (18) are uniformly bounded when ϵ is sufficiently small).

□

B WEIGHTED HYPERSINGULAR RIESZ ENERGY

We recall results most relevant to us from Borodachov et al. (2019). Suppose $X \subset \mathbf{R}^n$ is compact, of Hausdorff dimension d, and d-rectifiable, i.e., the image of a bounded set in $\mathbf{R}^d$ under a Lipschitz mapping. Given a non-vanishing continuous probability density p on X, define a measure $\mathcal{H}_d^p(B) := \int_B p(x)\mathrm{d}\mathcal{H}_d(x)$ for any Borel set $B \subset X$. The target measure h_d^p is defined to be $h_d^p(B) := \mathcal{H}_d^p(B)/\mathcal{H}_d^p(X)$. The *weighted s-Riesz energy* is defined as the interaction energy

$$E_s(\omega_N) := \sum_{i \neq j} \frac{(p(x_i)p(x_j))^{-s/2d}}{\|x_i - x_j\|^s}. \tag{19}$$

The following result states that minimizers of the weighted s-Riesz energy approximate the target measure when the s is sufficiently large.

Theorem B.1 (Theorem 11.1.2, Borodachov et al. (2019))**.** *Suppose $s > d$. For $\omega_N^* \in \arg\min E_s$ with $\omega_N^* = \{x_1^N, \ldots, x_N^N\}$, we have weak convergence $\delta_{\omega_N^*} \to h_d^p$ as $N \to \infty$.*

A similar result with a slightly different assumption holds for $s = d$ (Theorem 11.1.3 of Borodachov et al. (2019)).

Comparison of (19) with discretized MIE (7). Our theory is only valid for the full-dimensional case, i.e., $n = d$ (see a few exceptions discussed in Remark A.8). When this is the case and if the mollifier is taken to be from the s-Riesz family, (7) becomes, for $s > n$,

$$E_\epsilon(\omega_N) := \sum_{i,j=1}^{N} \frac{(p(x_i)p(x_j))^{-1/2}}{\left(\|x_i - x_j\|_2^2 + \epsilon^2\right)^{s/2}}.$$

Compared with (19), in our case there is an ϵ in the denominator, so that the continuous version of the energy $\mathcal{E}_\epsilon$ does not blow up all the time (this is in contrast with the continuous version of (19)—see Borodachov et al. (2019, Theorem 4.3.1)). Moreover, the exponential scaling on p is different: in our case we use $-1/2$ whereas in (19) it is $-s/2n$. The diagonal $i = j$ is included in our energy, but this is inconsequential as another valid discretization is to discard the diagonal term (Borodachov et al., 2019, Theorem 4.2.2). On the other hand, our theory allows a bigger class of mollifiers that are not necessarily of Riesz families. An interesting future research direction is to extend our theory to cases where $d < n$, e.g., when X is an embedded d-dimensional submanifold in $\mathbf{R}^n$.

C EXPERIMENT DETAILS AND ADDITIONAL RESULTS

C.1 GAUSSIANS IN VARYING DIMENSIONS

We generate the n-dimensional Gaussians used to produce Figure 1 as follows. We generate a matrix $\sqrt{A} \in \mathbf{R}^{n \times n}$ with i.i.d. entries uniformly in $[-1, 1]$. Then we set $A = \sqrt{A}\sqrt{A}^\top / \det\left(\sqrt{A}\right)$. This way $\det(A) = 1$. We then use A as the covariance matrix for the Gaussian (centered at 0). We use Adam with learning rate 0.01 for all methods for a total of 2000 iterations. This is enough for SVGD and MIED to converge, while for KSDD the convergence can be much slower.

We visualize the samples from each method for $n = 2$ in Figure 5. We notice that MIED is capable

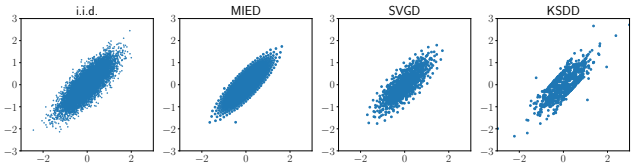

Figure 5: Visualization of samples from each method for a 2D skewed Gaussian.

of generating well-separated samples, while for SVGD there is a gap between the inner cluster of samples and a sparse outer ring. For KSDD we see the artifact where too many samples concentrate on the diagonal.

C.2 20-DIMENSIONAL DIRICHLET DISTRIBUTION

We sample from the 20-dimensional Dirichlet distribution in the same setup as in Shi et al. (2021) with 50 particles. Results and visualization are shown in Figure 6. We see that unlike sampling from a box Figure 3, both MSVGD and SVMD by Shi et al. (2021) perform well in this experiment. This is due to the fact that the entropic mirror map used here is a well-tested choice for simplex constraint, yet obtaining a good mirror map for a generic constraint, even if it is linear, can be challenging. Our method does not have such limitation as it can easily incorporate exisiting constrained optimization tools.

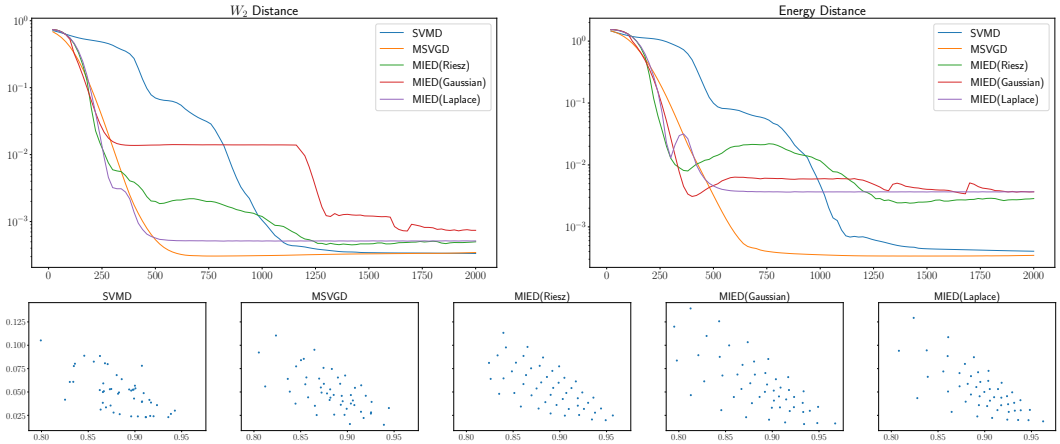

Figure 6: Top: Metrics vs. the number of iteration for sampling the 20-dimensional Dirichlet distribution. Bottom: visualization of samples from each method.