(De)-regularized Maximum Mean Discrepancy Gradient Flow

Zonghao Chen Aratrika Mustafi Pierre Glaser Anna Korba
Arthur Gretton Bharath K. Sriperumbudur

August 15, 2025

Accepted (minor revision) to JMLR

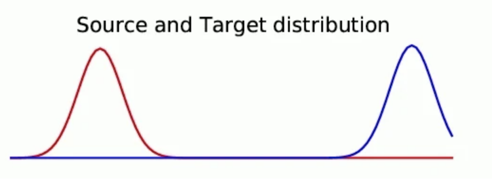
About Me

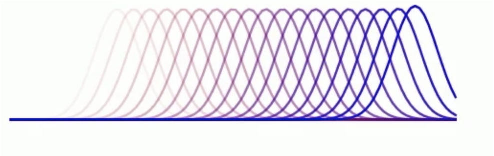
- Zonghao Chen
- 3rd year PhD Student at University College London (UCL)
  - Foundational AI Centre
  - Gatsby Computational Neuroscience Unit (Founded by Hinton in 1998)
- Graduated from Tsinghua University in 2022
  - Department of EE
- Kernel (nonparametric) methods, causal inference, statistical learning theory
- Visiting RIKEN AIP, Tokyo (Summer 2025)

Background

Problem: How to learn a target probability distribution π on $\mathbb{R}^d$.

- Sampling (e.g: $\pi \propto \exp(-V)$ is the posterior distribution in Bayesian inference).
- Optimizing neural networks (e.g: π is the mean-field limit over parameters of a neural network).
- Generative models (e.g: π is the distribution of an image dataset).





Background: Optimization in the space of probability measures

Problem: How to learn a target probability distribution π on $\mathbb{R}^d$.

- This problem can be written as an optimization problem on $\mathcal{P}_2(\mathbb{R}^d)$.

$$\arg\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} D(\mu \mid \pi).$$

- Here D is a similarity metric or distance, e.g. Kullback–Leibler divergence.
  - $D(\mu|\pi) = 0$ if and only if $\mu = \pi$.
- $\mathcal{P}_2(\mathbb{R}^d)$ denotes the space of probability measures with a finite second moment.
  - We primarily consider $\mathcal{P}_2(\mathbb{R}^d)$ rather than $\mathcal{P}(\mathbb{R}^d)$ for the nice geometrical properties.
- How to find the minimum? Gradient descent!

Background: Euclidean gradient flow

- Euclidean gradient flow of an objective $F : \mathbb{R}^d \to \mathbb{R}$

$$\partial_t x_t = -v(x_t), \quad v = \nabla F.$$

- ∇F denotes the gradient of F.
- This is the continuous-time analogue of gradient descent:

$$x_{n+1} = x_n - \gamma v(x_n),$$

  where $\gamma > 0$ is the step size.
- Gradient flow / descent is widely used to find minimizers of F:

$$x^* = \arg\min_{x \in \mathbb{R}^d} F(x).$$

  - Train large scale deep learning models.
- When F is both strongly convex and smooth, Euclidean gradient descent converges exponentially fast [Boyd and Vandenberghe, 2004].

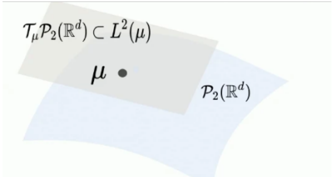
Background: Wasserstein gradient flow

Challenge: How to find $\arg\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} D(\mu \mid \pi)$?

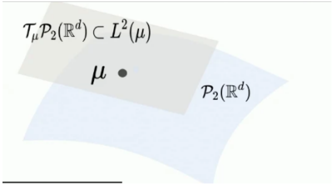
- Gradient descent! Wait... How do we define gradients in $\mathcal{P}_2(\mathbb{R}^d)$?
- Endow $\mathcal{P}_2(\mathbb{R}^d)$ with the Wasserstein-2 distance W_2.

$$W_2^2(\nu, \mu) = \int \|T(x) - x\|^2 \mathrm{d}\nu(x) = \|T - \mathrm{Id}\|^2_{L^2(\nu)}.$$

  - $W_2^2(\nu, \mu)$ means the minimal energy takes to transport mass from ν to μ.
  - $T : \mathbb{R}^d \to \mathbb{R}^d$ is the optimal transport map from ν to μ.
- $(\mathcal{P}_2, W_2)$ can be 'treated' as a Riemann manifold under the Otto's calculus [Otto, 2001].
- The tangent space $\mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d)$ at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is a dense subset of $L^2(\mu)$.



Background: Wasserstein gradient flow



Definition (Wasserstein gradient)

Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a regular functional. The Wasserstein gradient of $\mathcal{F}$ evaluated at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is the unique function $\nabla_{W_2}\mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}^d$, s.t. for any $T \in \mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d)$,

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[\mathcal{F}\left((\mathrm{Id} + \epsilon T)_{\#}\mu\right) - \mathcal{F}(\mu)\right] = \int [\nabla_{W_2}\mathcal{F}(\mu)](x)^\top T(x) \, \mathrm{d}\mu(x) = \langle \nabla_{W_2}\mathcal{F}, T \rangle_{L^2(\mu)} .$$

- The gradient is defined along a 'curve' $(\mathrm{Id} + \epsilon T)_{\#}\mu$ in $\mathcal{P}_2(\mathbb{R}^d)$.

Background: Wasserstein gradient flow

Definition (Wasserstein gradient flow)

Let $(v_t : \mathbb{R}^d \to \mathbb{R}^d)_{t\geq 0}$ be a family of vector fields and suppose that the random process $(x_t)_{t\geq 0}$ evolve according to $\dot{x}_t = v_t(x_t)$. Then, the law μ_t of x_t evolves according to the continuity equation (in the sense of distributions)

$$\partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0.$$

In particular, $(\mu_t)_{t\geq 0}$ is called the Wasserstein gradient flow of $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ if $v_t = -\nabla_{W_2}\mathcal{F}(\mu_t)$.

Euclidean Gradient Flow

- State space: $\mathbb{R}^d$
- Objective $F : \mathbb{R}^d \to \mathbb{R}$.
- Update scheme: $\dot{x}_t = v_t(x_t)$, with $v_t = -\nabla F$.

Wasserstein Gradient Flow

- State space: $\mathcal{P}_2(\mathbb{R}^d)$
- Objective $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$:
- Update scheme $\dot{x}_t = v_t(x_t)$, with $v_t = -\nabla_{W_2}\mathcal{F}(\mu_t)$.

Example 1: Langevin diffusion [Jordan et al., 1998]

- Given the target distribution $\pi \propto \exp(-V)$ with $V : \mathbb{R}^d \to \mathbb{R}$.
- The functional $\mathcal{F}_{\mathrm{KL}} = \mathrm{KL}(\cdot\|\pi)$ and its Wasserstein gradient

$$[\nabla_{W_2}\mathcal{F}_{\mathrm{KL}}(\mu)](\cdot) = \nabla V(\cdot) + \nabla \log \mu_t(\cdot).$$

- The Wasserstein gradient flow of $\mathcal{F}_{\mathrm{KL}}$

$$\partial_t \mu_t = \nabla \cdot \left(\mu_t \left(\nabla V + \nabla \log \mu_t\right)\right).$$

- It is equivalent to the Fokker–Planck equation of the Langevin diffusion [Särkkä and Solin, 2019]:

$$\mathrm{d}x_t = -\nabla V(x_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \quad \mu_t = \mathrm{Law}(x_t).$$

- A standard time-discretization (Euler–Maruyama scheme) is:

$$x_{n+1} = x_n - \gamma \nabla V(x_n) + \sqrt{2\gamma}\,\xi_n, \quad \xi_n \sim \mathcal{N}(0, I_d).$$

Example 2: MMD gradient flow [Arbel et al., 2019]

- Given M i.i.d samples $\{y_i\}_{i=1}^M$ from a target distribution π.
- The functional $\mathcal{F}_{\mathrm{MMD}} = \frac{1}{2}\mathrm{MMD}^2(\cdot\|\pi)$

$$\mathrm{MMD}(\mu\|\pi) := \|\int k(x,\cdot)\mathrm{d}\mu(x) - \int k(x,\cdot)\mathrm{d}\pi(x)\|_{\mathcal{H}},$$

where $\mathcal{H}$ is the RKHS associated with a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.

- Its Wasserstein gradient

$$\begin{aligned}[\nabla_{W_2}\mathcal{F}_{\mathrm{MMD}}(\mu)](\cdot) &= \nabla\left(\int k(x,\cdot)\,\mathrm{d}\mu(x) - \int k(x,\cdot)\,\mathrm{d}\pi(x)\right) \\ &\approx \int \nabla_2 k(x,\cdot)\,\mathrm{d}\mu(x) - \frac{1}{M}\sum_{i=1}^M \nabla_2 k(y_i,\cdot).\end{aligned}$$

- The Wasserstein gradient flow of $\mathcal{F}_{\mathrm{MMD}}$,

$$\partial_t\mu_t = \nabla\cdot\left(\mu_t\nabla_{W_2}\mathcal{F}_{\mathrm{MMD}}(\mu_t)\right), \quad \mathrm{d}x_t = -[\nabla_{W_2}\mathcal{F}_{\mathrm{MMD}}(\mu_t)](x_t)\,\mathrm{d}t.$$

- A standard time-discretization (Euler–Maruyama scheme) is:

$$x_{n+1} = x_n - \gamma\nabla_{W_2}\mathcal{F}_{\mathrm{MMD}}(\mu_n)(x_n).$$

Convexity and Smoothness

Question: When does Wasserstein gradient flow find $\arg\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} D(\mu \mid \pi)$?

Definition (Wasserstein Hessian [Villani et al., 2009])

Given any $T \in \mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d)$ and a curve (constant-speed geodesic) $\rho_t = (\mathrm{Id} + tT)_\# \mu$ for $0 \le t \le 1$, the Wasserstein Hessian of a functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ at μ, denoted as Hess $\mathcal{F}_{|\mu}$, is an operator from $L^2(\mu)$ to $L^2(\mu)$:

$$\left\langle \mathrm{Hess}\, \mathcal{F}_{|\mu} T, T \right\rangle_{L^2(\mu)} = \frac{d^2}{dt^2}\bigg|_{t=0} \mathcal{F}(\rho_t).$$

- A functional $\mathcal{F}$ is said to be (geodesically) M-smooth at μ if

$$\left\langle \mathrm{Hess}\, \mathcal{F}_{|\mu} T, T \right\rangle_{L^2(\mu)} \le M \| T \|_{L^2(\mu)}.$$

- A functional $\mathcal{F}$ is said to be (geodesically) Λ-convex at μ if

$$\left\langle \mathrm{Hess}\, \mathcal{F}_{|\mu} T, T \right\rangle_{L^2(\mu)} \ge \Lambda \| T \|_{L^2(\mu)}.$$

- If $\mathcal{F}$ is both (geodesically) M-smooth and Λ-convex, then $M \ge \Lambda$.

Convexity and Smoothness

- Let $(\mu_t)_{t\geq 0}$ be the Wasserstein gradient flow of $\mathcal{F}$
- If $\mathcal{F}$ is Λ-convex with $\Lambda > 0$,

$$\mathcal{F}(\mu_t) - \min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu) \leq \exp(-2\Lambda t)\Big(\mathcal{F}(\mu_0) - \min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu)\Big).$$

- Let $(\mu_n)_{n\in\mathbb{N}}$ be the Wasserstein gradient descent of $\mathcal{F}$.
- If $\mathcal{F}$ is Λ-convex with $\Lambda > 0$ and M-smooth, and the step size $0 < \gamma \leq \frac{1}{M}$,

$$\mathcal{F}(\mu_n) - \min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu) \leq \exp(-\gamma\Lambda n)\Big(\mathcal{F}(\mu_0) - \min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu)\Big).$$

- This is the same as convex optimization in Euclidean space [Boyd and Vandenberghe, 2004].

Convexity and Smoothness

Wasserstein gradient flow of $\mathcal{F}_{\mathrm{KL}}$

- $\pi \propto \exp(-V)$
- Sampling
- When $\pi \propto \exp(-V)$ is strongly log-concave, i.e., $\mathbf{H}V \geq \alpha \mathrm{Id}$, then $\mathcal{F}_{\mathrm{KL}}$ is α-convex.
- Convergence in discrete time [Vempala and Wibisono, 2019]

$$\mathcal{F}_{\mathrm{KL}}(\mu_n \| \pi) \leq \exp(-\alpha\gamma n)\mathcal{F}_{\mathrm{KL}}(\mu_n \| \pi) + \frac{\gamma n \beta^2}{\alpha}.$$

- β is the Lipschitz continuity of V.
- It takes $\mathcal{O}(\frac{1}{\alpha\delta} \log \frac{1}{\delta})$ to reach δ error.

Wasserstein gradient flow of $\mathcal{F}_{\mathrm{MMD}}$

- $\{y_i\}_{i=1}^M$ i.i.d samples from π
- Generative modelling
- When k is bounded and has bounded derivatives, then $\mathcal{F}_{\mathrm{MMD}}$ is M-smooth and $-M$-convex.
- Convergence in discrete time [Arbel et al., 2019]

$$\mathcal{F}_{\mathrm{MMD}}(\mu_n, \pi) \leq \frac{W_2^2(\mu_0, \pi)}{\gamma n} + \bar{K}.$$

- $\bar{K}$ is a positive barrier term that does not vanish.
- $\lim_{n\to\infty} \mathcal{F}_{\mathrm{MMD}}(\mu_n, \pi) \neq 0$!

Non-convexity of MMD prevents global convergence

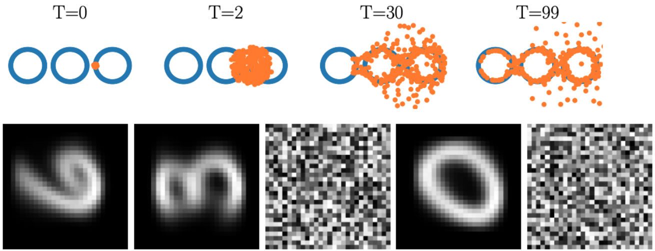


Figure: Belhadji et al. [2025]

- Arbel et al. [2019] proves convergence of MMD flow, yet under noise injection and stringent conditions on the scale of noise.
- Disclaimer: There exists many papers where MMD flow empirically generate high-quality images on with adversarial training of kernels [Galashov et al., 2025], or with non-smooth kernels plus deep neural network distillation [Hertrich et al., 2024, Altekrüger et al., 2023].

Convexity and Smoothness

Question: Can we find a new objective $\mathcal{F}$ such that it enjoys (geodesic) convexity, similar to $\mathcal{F}_{\mathrm{KL}}$, in the generative modelling setting where only samples are available?

Proposition 1 (MMD and χ^2-divergence)

Suppose μ is absolutely continuous with respect to π, i.e., $\mu \ll \pi$. Then

$$\mathrm{MMD}^2(\mu\|\pi) = \left\| T_\pi^{\frac{1}{2}} \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1 \right) \right\|_{L^2(\pi)}^2 \quad \text{and} \quad \chi^2(\mu\|\pi) = \left\| \frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1 \right\|_{L^2(\pi)}^2 .$$

Here, $T_\pi : L^2(\pi) \to L^2(\pi)$ is the kernel integral operator defined as

$$T_\pi f(\cdot) = \int k(x, \cdot) f(x) \, \mathrm{d}\pi(x).$$

- Similar to $\mathcal{F}_{\mathrm{KL}}$, $\mathcal{F}_{\chi^2}(\cdot) = \chi^2(\cdot\|\pi)$ is (geodesic) strong convex when π is strongly log-concave [Ohta and Takatsu, 2011].
- An interpolation between MMD^2 and χ^2?

DrMMD: An interpolation of MMD and χ^2-divergence

Definition (De-regularized Maximum Mean Discrepancy (DrMMD))

Suppose $\mu \ll \pi$ where $\mu, \pi \in \mathcal{P}_2(\mathbb{R}^d)$. Then the (de)-regularized maximum mean discrepancy (DrMMD) between μ, π is defined as, for $\lambda > 0$,

$$\mathrm{DrMMD}(\mu \| \pi) = (1+\lambda) \left\| \left((T_\pi + \lambda \mathrm{Id})^{-1} T_\pi \right)^{\frac{1}{2}} \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1 \right) \right\|_{L^2(\pi)}^2 .$$

Proposition 2 (Interpolation of MMD^2 and χ^2)

Suppose k is bounded, continuous, and c_0-universal.

$$\lim_{\lambda \to 0} \mathrm{DrMMD}(\mu \| \pi) = \chi^2(\mu \| \pi), \qquad \lim_{\lambda \to \infty} \mathrm{DrMMD}(\mu \| \pi) = \mathrm{MMD}^2(\mu \| \pi).$$

- Similar idea of spectral regularization has been done for kernel hypothesis testing [Mika et al., 1999, Harchaoui et al., 2007, Hagrass et al., 2024].
- This is known as Tikhonov regularization.

DrMMD: An interpolation of MMD and χ^2

Question: Does DrMMD inherit the advantages of MMD^2 and χ^2?

- Does $\mathcal{F}_{\mathrm{DrMMD}}$ admit finite sample implementation of its Wasserstein gradient flow?
- Is $\mathcal{F}_{\mathrm{DrMMD}}$ (geodesic) strongly convex when π is strongly log-concave?

Assumption 1

$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a continuous and c_0-universal kernel. The kernel is bounded by K, its first order derivatives bounded by K_{1d} and second order derivatives bounded by K_{2d}.

- This condition is satisfied by Gaussian kernels, Matérn kernels and inverse multiquadratic kernels.

Finite-sample estimate

- Let $\Sigma_\pi : \mathcal{H} \to \mathcal{H}$ denote the covariance operator $\Sigma_\pi = \mathbb{E}_\pi[k(x,\cdot) \otimes k(x,\cdot)]$.

$$\langle f, \Sigma_\pi f \rangle_{\mathcal{H}} = \mathbb{E}_\pi[f(X)^2].$$

Proposition 3 (Finite-sample estimate of the Wasserstein gradient of $\mathcal{F}_{\mathrm{DrMMD}}$)

The Wasserstein gradient of $\mathcal{F}_{\mathrm{DrMMD}}(\cdot) = \mathrm{DrMMD}(\cdot \| \pi)$ *at* μ *is* $(1+\lambda)\nabla h_{\mu,\pi}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$, *where*

$$h_{\mu,\pi} = \left(T_\pi + \lambda \mathrm{I}\right)^{-1} T_\pi \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\right) = \left(\Sigma_\pi + \lambda \mathrm{I}\right)^{-1} \left(\int k(x,\cdot)\mathrm{d}\mu - \int k(x,\cdot)\mathrm{d}\pi\right).$$

Given empirical distributions $\hat{\mu} = \frac{1}{N}\sum_{i=1}^N x_i$ *and* $\hat{\pi} = \frac{1}{M}\sum_{i=1}^M y_i$. *Given the Gram matrices* $K_{xx} \in \mathbb{R}^{N\times N}$, $K_{yy} \in \mathbb{R}^{M\times M}$, $K_{xy} \in \mathbb{R}^{N\times M}$.

$$\begin{aligned} h_{\hat{\mu},\hat{\pi}}(\cdot) = \; & \frac{1}{N\lambda} k\left(\cdot, x_{1:N}\right) \mathbb{1}_N - \frac{1}{M\lambda} k\left(\cdot, y_{1:M}\right) \mathbb{1}_M - \frac{1}{M\lambda} k\left(\cdot, y_{1:M}\right) \left(M\lambda \mathrm{I} + K_{yy}\right)^{-1} K_{yx} \mathbb{1}_N \\ & + \frac{1}{M\lambda} k\left(\cdot, y_{1:M}\right) \left(M\lambda \mathrm{I} + K_{yy}\right)^{-1} K_{yy} \mathbb{1}_M. \end{aligned}$$

Wasserstein Hessian and convexity

Proposition 4 (Wasserstein Hessian of $\mathcal{F}_{\chi^2}$)

Suppose k satisfies Assumption 1. Let $\mu, \pi \in \mathcal{P}_2(\mathbb{R}^d)$.

$$\left| \left\langle \operatorname{Hess} \mathcal{F}_{\mathrm{DrMMD}|\mu} T, T \right\rangle_{L^2(\mu)} \right| \leq 2(1+\lambda) \frac{2\sqrt{K K_{2d}} + K_{1d}}{\lambda} \|T\|^2_{L^2(\mu)}, \quad \forall T \in \mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d).$$

Let π be α-strongly log-concave, i.e., $\pi \propto \exp(-V)$, $\mathbf{H} V \succeq \alpha \mathrm{I}$, and assume additionally that $x \mapsto \mathbf{H} V(x)$ is continuous. Then for all μ such that $x \mapsto \nabla \log \mu(x)$ is continuous and $\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1 \in \mathcal{H}$,

$$\left\langle \operatorname{Hess} \mathcal{F}_{\mathrm{DrMMD}|\mu} T, T \right\rangle_{L^2(\mu)} \geq \alpha(1+\lambda) \int \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x) \|T(x)\|^2 \mathrm{d}\mu - R(\lambda, \mu, T),$$

where $\lim_{\lambda \to 0} R(\lambda, \mu, T) = 0$.

- DrMMD is more convex when $\lambda \to 0$ and more smooth when $\lambda \to \infty$.
- Unfortunately, $\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1 \in \mathcal{H}$ is too strong in practice.

Poincaré inequality

- Exponential convergence to the global minima (not necessarily unique) still hold under a Polyak-Łojasiewicz inequality, a strict relaxation of strong convexity.
- $\mathcal{F}_{\chi^2} = \chi^2(\cdot\|\pi)$ satisfies a (modified) PL inequality with α if, for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\chi^2(\mu\|\pi) \leq \frac{1}{2\alpha}\left\|\nabla_{W_2}\mathcal{F}_{\chi^2}(\cdot)\right\|^2_{L^2(\pi)} = \frac{1}{2\alpha}\left\|\nabla\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)(x)\right\|^2_{L^2(\pi)}.$$

- It is implied by π satisfying the Poincaré inequality ($f = \frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1$).

Definition (Poincaré inequality)

We say that π satisfies a Poincaré inequality with constant C_P if for all $f, \nabla f \in L^2(\pi)$,

$$\mathrm{Var}_\pi[f] \leq C_P \mathbb{E}_\pi\left[\|\nabla f\|^2\right].$$

Furthermore, π satisfies a Poincaré with constant α if π is α-log concave.

- Poincaré inequality is a strict relaxation of strong log concavity. It is satisfied by mixture of Gaussians. It is invariant under Lipschitz perturbations [Bakry et al.,

Poincaré inequality

Proposition 5 (Exponential convergence of $\mathcal{F}_{\chi^2}$ gradient flow [Chewi et al., 2020])

Suppose that π satisfies a Poincaré inequality with constant C_P. Let $(\mu_t)_{t\geq 0}$ be the Wasserstein gradient flow of $\mathcal{F}_{\chi^2}$. Then, for any $T \geq 0$,

$$\mathrm{KL}\left(\mu_T \| \pi\right) \leq \exp\left(-\frac{2T}{C_P}\right) \mathrm{KL}\left(\mu_0 \| \pi\right).$$

For any $t > 0$,

$$\partial_t \mathrm{KL}\left(\mu_t \| \pi\right) = -2\mathbb{E}_{\mu_t}\left\langle \nabla \log \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}, \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} \right\rangle = -2\mathbb{E}_{\pi}\left[\left\| \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} \right\|^2\right]$$
$$\overset{(*)}{\leq} -\frac{2}{C_P}\chi^2\left(\mu_t \| \pi\right) \leq -\frac{2}{C_P}\mathrm{KL}\left(\mu_t \| \pi\right).$$

- $(*)$ holds by the Poincaré inequality.

DrMMD: An interpolation of MMD and χ^2

Question: Does DrMMD inherit the advantages of MMD^2 and χ^2 ?

- Does $\mathcal{F}_{\mathrm{DrMMD}}$ admit finite sample implementation of its Wasserstein gradient flow?
- ~~Is $\mathcal{F}_{\mathrm{DrMMD}}$ (geodesic) convex when π is log-concave?~~
- Does $\mathcal{F}_{\mathrm{DrMMD}}$ satisfy a (modified) PL condition when π satisfies a Poincaré inequality?

- Let $(\mu_t)_{t\geq 0}$ be the Wasserstein gradient flow of $\mathcal{F}_{\mathrm{DrMMD}}$ with a continuity equation

$$\partial_t \mu_t + \nabla \cdot (\mu_t (1+\lambda) \nabla h_{\mu_t,\pi}) = 0, \quad h_{\mu_t,\pi} = (T_\pi + \lambda \mathrm{I})^{-1} T_\pi \left(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} - 1 \right).$$

PL condition of $\mathcal{F}_{\mathrm{DrMMD}}$

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\mu_t \| \pi\right) \\
&= -\int \nabla h_{\mu_t,\pi}(x)^\top \nabla \log \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x)\, \mathrm{d}\mu_t \\
&= -\int \nabla h_{\mu_t,\pi}(x)^\top \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x)\, \mathrm{d}\pi \\
&= -\int \left(\nabla h_{\mu_t,\pi}(x) - \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x)\right)^\top \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x)\, \mathrm{d}\pi - \int \left\| \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x) \right\|^2 \mathrm{d}\pi \\
&= -\int \left(\nabla h_{\mu_t,\pi}(x) - \nabla \left(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x) - 1\right)\right)^\top \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x)\, \mathrm{d}\pi - \int \left\| \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x) \right\|^2 \mathrm{d}\pi.
\end{aligned}
$$

Apply integration by parts for the first term.

PL condition of $\mathcal{F}_{\mathrm{DrMMD}}$

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\mu_t \| \pi\right) \\
&= \int \left(h_{\mu_t,\pi}(x) - \left(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x) - 1 \right) \right) \nabla \cdot \left(\pi(x) \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x) \right) \mathrm{d}x - \int \left\| \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}(x) \right\|^2 \mathrm{d}\pi \\
&\leq \left\| h_{\mu_t,\pi} - \left(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} - 1 \right) \right\|_{L^2(\pi)} \left\| \frac{\nabla \cdot \left(\pi \nabla \frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} \right)}{\pi} \right\|_{L^2(\pi)} - \frac{1}{C_P}\mathrm{KL}\left(\mu_t \| \pi\right),
\end{aligned}
$$

- The first term is bounded by Cauchy-Schwartz inequality, and the second term is bounded by the Poincaré inequality with C_P.
- Suppose $\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} - 1 \in \mathrm{Ran}(T_\pi^r)$ with $r > 0$.

$$
\left\| h_{\mu_t,\pi} - \left(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} - 1 \right) \right\|_{L^2(\pi)} \leq \lambda^r \left\| q_t \right\|_{L^2(\pi)}, \text{ where } h_{\mu_t,\pi} = T_\pi^r q_t.
$$

- $h_{\mu_t,\pi} = (T_\pi + \lambda \mathrm{I})^{-1} T_\pi (\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} - 1)$
- Similar results have been established in kernel ridge regression [Cucker and Zhou, 2007].

PL condition of $\mathcal{F}_{\mathrm{DrMMD}}$

Proposition 6 (PL condition of $\mathcal{F}_{\mathrm{DrMMD}}$)

Let $(\mu_t)_{t\geq 0}$ be the Wasserstein gradient flow of $\mathcal{F}_{\mathrm{DrMMD}}$. Suppose the kernel satisfied Assumption 1. Suppose the target distribution π satisfies a Poincaré inequality with constant C_P. Suppose $\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} - 1 \in \mathrm{Ran}(T_\pi^r)$ with $r > 0$, i.e., there exists $q_t \in L^2(\pi)$ such that $\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} - 1 = T_\pi^r q_t$. Suppose $\|\nabla(\log \pi)^\top \nabla(\frac{d\mu_t}{d\pi})\|_{L^2(\pi)} \leq \mathcal{J}_t$ and $\|\Delta(\frac{d\mu_t}{d\pi})\|_{L^2(\pi)} \leq \mathcal{I}_t$. Then, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\mu_t \| \pi\right) \leq -\frac{1}{C_P}\mathrm{KL}\left(\mu_t \| \pi\right) + \lambda^r(\mathcal{J}_t + \mathcal{I}_t).$$

- When $\lambda = 0$, we recover the PL condition of χ^2 divergence.
- $\mathcal{J}_t$ and $\mathcal{I}_t$ are additional regularity conditions.
- Compared with the initial regularity condition $\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} - 1 \in \mathcal{H}$ required for the (geodesic) convexity of $\mathcal{F}_{\mathrm{DrMMD}}$, $\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} - 1 \in \mathrm{Ran}(T_\pi^r)$ is a strict relaxation when $0 < r < \frac{1}{2}$.

Interpolation space $\mathrm{Ran}(T_\pi^r)$

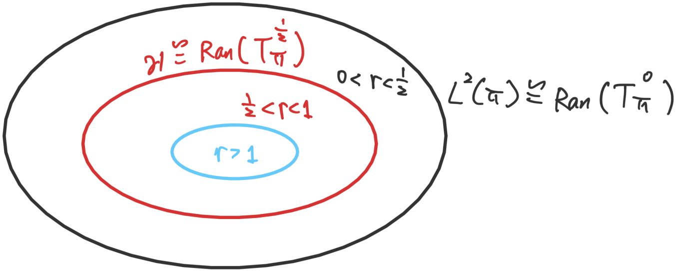


Figure: Visualization of $\mathrm{Ran}(T_\pi^r)$.

- Large r indicates larger smoothness.
- $\mathrm{Ran}(T_\pi^0) \cong L^2(\pi)$ and $\mathrm{Ran}(T_\pi^{\frac{1}{2}}) \cong \mathcal{H}$.

Convergence of DrMMD gradient flow

Theorem 1 (Convergence of DrMMD gradient flow)

In addition to the assumptions of the proposition of PL condition on $\mathcal{F}_{\mathrm{DrMMD}}$. If $\|q_t\|_{L^2(\pi)} \leq Q, \mathcal{J}_t \leq \mathcal{J}, \mathcal{I}_t \leq \mathcal{I}$ for all $0 \leq t \leq T$, where $Q, \mathcal{J}$, and $\mathcal{I}$ are positive constants independent of λ, then for any $T \geq 0$,

$$\mathrm{KL}\left(\mu_T \| \pi\right) \leq \exp\left(-\frac{2(1+\lambda)}{C_P} T\right) \mathrm{KL}\left(\mu_0 \| \pi\right) + \lambda^r C_P Q(\mathcal{J} + \mathcal{I}).$$

- When $\lambda = 0$, it recovers the exponential convergence of χ^2 flow.

$$\mathrm{KL}\left(\mu_T \| \pi\right) \leq \exp\left(-\frac{2}{C_P} T\right) \mathrm{KL}\left(\mu_0 \| \pi\right).$$

- Larger r means more regularity of the trajectory and thus smaller bias.
- Smaller Poincaré ($C_P = 1/\alpha$) means faster convergence.
- For continuous time DrMMD flow, we would want $\lambda \to 0$ for ‘convexity’, however, that is not the case for discrete time flow.

DrMMD gradient descent

- DrMMD gradient flow (continuity equation)

$$\partial_t \mu_t + \nabla \cdot (\mu_t (1+\lambda) \nabla h_{\mu_t,\pi}) = 0$$

- DrMMD gradient descent: for step size $\gamma > 0$,

$$\mu_{n+1} = (\mathrm{Id} + \gamma(1+\lambda)\nabla h_{\mu_n,\pi})_{\#}\mu_n.$$

- Recall the Wasserstein Hessian of $\mathcal{F}_{\mathrm{DrMMD}}$

$$\left| \left\langle \mathrm{Hess}\, \mathcal{F}_{\mathrm{DrMMD}|\mu} T, T \right\rangle_{L^2(\mu)} \right| \leq 2(1+\lambda) \frac{2\sqrt{KK_{2d}} + K_{1d}}{\lambda} \| T \|^2_{L^2(\mu)}, \quad \forall T \in \mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d).$$

- Taking $\lambda \to 0$ breaks the smoothness of $\mathcal{F}_{\mathrm{DrMMD}}$.

Convergence of DrMMD gradient descent

Proposition 7 (Descent lemma of DrMMD gradient descent)

Let $(\mu_n)_{n\in\mathbb{N}}$ be the Wasserstein gradient descent of $\mathcal{F}_{\mathrm{DrMMD}}$. Suppose $\pi \propto \exp(-V)$ with $\mathbf{H}V \preceq \beta \mathrm{I}$. Suppose all assumptions in the proposition of the PL condition on $\mathcal{F}_{\mathrm{DrMMD}}$ hold. Suppose the step size γ is small enough.

$$\mathrm{KL}\left(\mu_{n+1}\|\pi\right) - \mathrm{KL}\left(\mu_n\|\pi\right) \leq -\frac{2}{C_P}\chi^2\left(\mu_n\|\pi\right)\gamma + \underbrace{\gamma\lambda^r Q(\mathcal{J}+\mathcal{I})}_{\text{Approximation error}} + \underbrace{\gamma^2\beta\chi^2\left(\mu_n\|\pi\right)\frac{K_{1d}+K_{2d}}{\lambda}}_{\text{Discretization error}}$$

- A trade-off between the approximation error and the time-discretization error.
- Optimal choice of adaptive λ_n at each iterate n:

$$\lambda_n = \left(2\gamma\chi^2\left(\mu_n\|\pi\right)\frac{\beta(K_{1d}+K_{2d})}{Q(\mathcal{J}+\mathcal{I})}\right)^{\frac{1}{r+1}} \propto \chi^2\left(\mu_n\|\pi\right)^{\frac{1}{r+1}}$$

- At the start, we want a larger λ to have more smoothness; when closer to the convergence, we want a smaller λ to operate in the χ^2 regime to better catch the difference of the distributions.

Convergence of DrMMD gradient descent

Theorem 2 (Convergence of DrMMD gradient descent)

Let $(\mu_n)_{n\in\mathbb{N}}$ be the Wasserstein gradient descent of $\mathcal{F}_{\mathrm{DrMMD}}$. Suppose all conditions from the descent lemma hold. Then, for any $n_{\max} \in \mathbb{N}$,

$$\mathrm{KL}\left(\mu_{n_{\max}} \| \pi\right) \leq \exp\left(-\frac{2 n_{\max}\gamma}{C_P}\right) \mathrm{KL}\left(\mu_0 \| \pi\right) + \gamma^{\frac{r}{r+1}} C_P Q^{\frac{2r+1}{r+1}} \left((K_{1d} + K_{2d})\beta\right)^{\frac{r}{r+1}} (\mathcal{J} + \mathcal{I})^{\frac{1}{r+1}}$$

- To reach error $\mathrm{KL}\left(\mu_{n_{\max}} \| \pi\right) \leq \delta$, it takes $\mathcal{O}((\frac{1}{\delta})^{\frac{r+1}{r}} \log \frac{1}{\delta})$ iterations.
- By comparison, for Langevin Monte Carlo, it takes $\mathcal{O}(\frac{1}{\delta} \log \frac{1}{\delta})$ [Chewi et al., 2024].
- DrMMD gradient descent takes more iterations due to the additional approximation error $\mathcal{O}(\lambda^r)$, but it operates without the knowledge of potential V and only requires $\Sigma_\pi = \int k(x,\cdot) \otimes k(x,\cdot) \, \mathrm{d}\pi$ and the embedding $\int k(x,\cdot) \, \mathrm{d}\pi$.

Particle DrMMD gradient descent

- To operate in the setting of generative models, we only have access to samples.
- We are given N samples from the initial distribution $\{x_i^{(0)}\}_{i=1}^N \sim \mu_0$ and M samples from the target distribution $\{y_i\}_{i=1}^M \sim \pi$.
- The DrMMD particle descent from time n to time $n+1$, is defined as

$$x_i^{(n+1)} = x_i^{(n)} - \gamma\,(1 + \lambda_n)\,\nabla h_{\hat{\mu}_n,\hat{\pi}}(x_i^{(n)}), \quad i = 1, \ldots, N.$$

- $h_{\hat{\mu}_n,\hat{\pi}}$ admits a closed-form expression with Gram matrices.
- λ_n is taken to be proportionate to $\mathrm{DrMMD}(\hat{\mu}_n \| \hat{\pi})^{\frac{1}{r+1}}$.
- r is selected from a set $\{0.1, 0.2, \ldots, 1.0\}$.

Conclusions

- We propose DrMMD gradient flow as an interpolation of MMD gradient flow and χ^2 gradient flow.
- DrMMD gradient flow / descent has global convergence results, compared to MMD flow, under an adaptive regularization parameter λ.
- This justifies the application of adaptive kernels in recent MMD flow (MMD GAN) papers that achieve SOTA empirical performances.
- More in the paper `https://arxiv.org/pdf/2409.14980`.
    - Empirical results on synthetic datasets.
    - An example of DrMMD flow with a Gaussian target distribution π which satisfies all conditions in the theorems.
    - Finite-particle convergence results with propagation of chaos bound.

F. Altekrüger, J. Hertrich, and G. Steidl. Neural wasserstein gradient flows for maximum mean discrepancies with riesz kernels. *arXiv preprint arXiv:2301.11624*, 2023.

M. Arbel, A. Korba, A. Salim, and A. Gretton. Maximum mean discrepancy gradient flow. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

D. Bakry, I. Gentil, M. Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

A. Belhadji, D. Sharp, and Y. Marzouk. Weighted quantization using mmd: From mean field to mean shift via gradient flows. *arXiv preprint arXiv:2502.10600*, 2025.

S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

S. Chewi, T. Le Gouic, C. Lu, T. Maunu, and P. Rigollet. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2098–2109. Curran Associates, Inc., 2020.

S. Chewi, M. A. Erdogdu, M. Li, R. Shen, and M. S. Zhang. Analysis of Langevin Monte Carlo from Poincare to log-Sobolev. *Foundations of Computational Mathematics*, pages 1–51, 2024.

F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*, volume 24. Cambridge University Press, 2007.

A. Galashov, V. D. Bortoli, and A. Gretton. Deep MMD gradient flow without adversarial training. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Pf85K2wtz8.

O. Hagrass, B. K. Sriperumbudur, and B. Li. Spectral regularized kernel two-sample tests. *The Annals of Statistics*, 52(3):1076–1101, 2024.

Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

J. Hertrich, C. Wald, F. Altekrüger, and P. Hagemann. Generative sliced MMD flows with riesz kernels. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VdkGRV1vcf.

R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468)*, pages 41–48, 1999.

S.-i. Ohta and A. Takatsu. Displacement convexity of generalized relative entropies. *Advances in Mathematics*, 228(3):1742–1787, 2011.

F. Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.

S. Särkkä and A. Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

S. Vempala and A. Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

C. Villani et al. *Optimal Transport: Old and New*, volume 338. Springer, 2009.