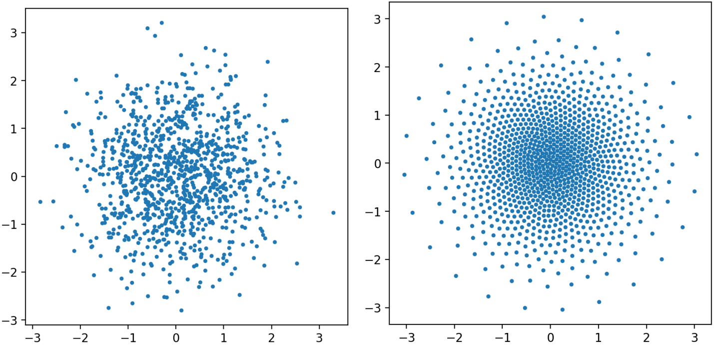
Gradient flows for sampling and their deterministic interacting particle approximations

Dejan Slepčev
Carnegie Mellon University

Centre International de Rencontres Mathématiques
Aggregation-Diffusion Equations & Collective Behavior: Analysis, Numerics and Applications
11.April 2024.

From Xu, Korba, S. *Accurate Quantization of Measures via Interacting Particle-based Optimization*, ICML 2022.



(a) i.i.d. sample (b) deterministic arrangement

Figure: Quantizing a Gaussian using 1024 particles.

Measuring Quantization error

- $\mathbf{d}$ – a metric or general dissimilarity measure on $\mathcal{P}(\mathbb{R}^d)$ or its subset [Wasserstein metric, MMD, KSD, $*$-discrepancy, etc.]
- $\mu \in \mathcal{P}(\mathbb{R}^d)$

Random quantization error

$$\mathcal{Q}_R(n, \mathbf{d}) = E[\mathbf{d}(\mu, \mu_n)]$$

where $\mu_n = \frac{1}{n}\sum_i \delta_{x_i}$ and $x_i \sim \mu$ are i.i.d samples of μ.

Optimal quantization error

$$\mathcal{Q}_O(n, \mathbf{d})) = \inf_{\{x_1,\ldots,x_n\}} \mathbf{d}(\mu, \mu_n)$$

Quantization error of optimal transport

Given $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, **transport plans,** π are probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with first marginal μ and second marginal ν:

$$\Pi(\mu,\nu) = \{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \; : \; \pi(A \times \mathbb{R}^d) = \mu(A), \, \pi(\mathbb{R}^d \times A) = \nu(A)\}.$$

p-OT distance

$$d_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^p \, d\pi(x,y) \right)^{\frac{1}{p}}$$

For μ with bounded support on a connected domain, with density bounded from below (Ajtai, Komlos, Tusnady 1984, Talagrand and Yukic 1993)

$$\mathcal{Q}_R(n, d_p) \lesssim \begin{cases} n^{-1/2} & \text{if } d = 1 \\ n^{-1/2}(\log n)^{\frac{1}{2}} & \text{if } d = 2 \\ n^{-1/d} & \text{if } d \geqslant 3. \end{cases}$$

and

$$\mathcal{Q}_O(n, d_p) \sim n^{-1/d}$$

Reproducing Kernel Hilbert Space (RKHS)

Definition. Hilbert space H is an RKHS if pointwise evaluation $f \mapsto f(x)$ is a continuous operator.

Example: Sobolev space H^s for $s > d/2$ is an RKHS.

- For all x there exists $\phi_x \in H$ such that $\langle \phi_x, f \rangle_H = f(x)$.
- The associated kernel is $K(x, y) = \langle \phi_x, \phi_y \rangle_H$.
- For $f = \sum_{i=1}^n a_i \phi_{x_i}$, $\langle f, f \rangle = \sum_{i,j} a_i a_j K(x_i, x_j) \geqslant 0$. So K is positive definite.
- If the Hilbert space is translation invariant, $K(x, y) = K(x - y)$
- Conversely, any positive definite continuous kernel $K(x - y)$ defines am RKHS, H_K, functions $f = K * \theta \in H_K$ for θ finite measure and

$$\|f\|_{H_K}^2 = \iint K(x-y) d\theta(x) d\theta(y) = \int \frac{1}{\widehat{K}(\xi)} |\widehat{f}(\xi)|^2 d\xi.$$

Examples: $K(x) = \exp(-|x|^2)$ -Gaussian, $K(x) = \exp(-|x|)$ - Laplace.

Maximum Mean Discrepancy (MMD)

Let H_K be RKHS corresponding to a kernel K.

$$\mathrm{MMD}_{H_K}(\rho,\pi) = \sup_{\|\phi\|_{H_K}\leqslant 1} \int \phi\, d\rho - \int \phi\, d\pi$$

It is known that

$$\mathrm{MMD}^2_{H_K}(\rho,\pi) = \iint K(x,y)\, d(\rho-\pi)(x)\, d(\rho-\pi)(y)$$

If $K(x,y) = K(x-y)$ then

$$\mathrm{MMD}^2_{H_K}(\rho,\pi) = \int K*\rho\, d\rho - 2\int K*\rho\, d\pi + \int K*\pi\, d\pi$$

For kernels K which decay at infinity and are strictly integrally positive definite, MMD_{H_K} metrizes narrow convergence of measures. (see *Sriperumbudur* 2016)

Quantization in MMD

For a broad set of Kernels and $\rho \in \mathcal{P}(\mathbb{R}^d)$ (see *Sriperumbudur* 2016)

$$\mathcal{Q}_R(\mathcal{L}, MMD) \lesssim \frac{1}{\sqrt{n}}$$

Theorem [Xu, Korba, S.]

Assume $K(x, y) = K(x - y)$ and $\widehat{K}(\xi) \lesssim (1 + |\xi|^2)^{-d/2}$, which holds for Gaussian, a range of Matérn kernels and others.

- **Lebesgue measure on** $[0, 1]^d$**.**

$$\mathcal{Q}_O(\mathcal{L}, MMD) \lesssim \frac{(\ln n)^{d-1}}{n}.$$

- **Light-tailed probability measure on** $\mathbb{R}^d$**.**

$$\mathcal{Q}_O(\pi, MMD) \lesssim \frac{(\ln n)^{(5d+1)/2}}{n}.$$

Open: Optimal rate on n. Dependance of constants on d.

MMD gradient flows in Wasserstein Metric

Arbel, Korba, Salim, Gretton, '19
For fixed μ consider $\mathrm{MMD}(\rho, \pi)$ as a functional of ρ. More precisely let

$$E(\rho) = \frac{1}{2} \int K * \rho \, d\rho - \int K * \pi \, d\rho$$

Note: total energy = interaction energy $+$ potential energy.

Gradient flow in Wasserstein metric

$$\partial_t \rho + \nabla \cdot \left(\rho \nabla K * (\pi - \rho) \right) = 0.$$

MMD gradient flows: Discrete measures

For fixed $\rho_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$

$$E(\rho_n) = \frac{1}{2n^2}\sum_i\sum_j K(x_i - x_j) - \frac{1}{n}\sum_i K * \pi(x_i)$$

Gradient flow in Wasserstein metric

$$\partial_t \rho + \nabla \cdot (\rho \nabla K * (\pi - \rho)) = 0.$$

Gradient flow for discrete measures: $\rho_n(t) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$

$$\dot{x}_i = \nabla K * \pi(x_i) - \frac{1}{n}\sum_{i=1}^n \nabla K(x_i - x_j)$$

Note: We need to know π which is not available in sampling problems.

Open Problems

- Does $\mathrm{MMD}(\rho(t), \mu) \to 0$ as $n \to \infty$ if ρ is absolutely continuous wrt Lebesgue measure? At what rate?
- What is the limit of $\mathrm{MMD}(\rho_n(t), \mu)$ as $t \to \infty$?

$$\partial_t \rho + \nabla \cdot (\rho \nabla K * (\pi - \rho)) = 0.$$

Gradient flow for discrete measures: $\rho_n(t) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$

$$\dot{x}_i = \nabla K * \pi(x_i) - \frac{1}{n}\sum_{i=1}^n \nabla K(x_i - x_j)$$

Open Problems

- Does $\text{MMD}(\rho(t), \mu) \to 0$ as $n \to \infty$ if ρ is absolutely continuous wrt Lebesgue measure? At what rate?
  Boufadene, Vialard show that for $K(x, y) = |x - y|^{-d+2}$ for $d \geqslant 3$, C^1 positive solutions on compact manifolds satisfy
  $$W(\mu_t, \pi) \lesssim e^{-\lambda t}.$$
- What is the limit of $\text{MMD}(\rho_n(t), \mu)$ as $t \to \infty$?
- Approaches for $\pi \sim e^{-U}$.

Fokker–Planck equation

Consider **Kullback-Leibler divergence**, that is the relative entropy

$$\mathrm{KL}(\rho) = \int \ln\left(\frac{\rho}{\pi}\right) \rho \, dx.$$

Wasserstein gradient flow is given by $\partial_t \rho = -\nabla \cdot (\rho v)$, where the vector field v minimizes the Rayleigh functional

$$R(v) = \frac{1}{2} g_\rho(v, v) + \frac{\delta \, \mathrm{KL}}{\delta \rho}[v] \quad = \frac{1}{2}\int |v|^2 \rho(x) dx - \int (\ln \rho + U) \nabla \cdot (\rho v) dx$$
$$= \frac{1}{2} \int |v|^2 \rho(x) dx + \nabla \rho \cdot v + \nabla U \cdot v \rho dx$$

where $\pi = C \exp(-U)$. Minimizing over v gives $v = -\left(\frac{\nabla \rho}{\rho} + \nabla U\right)$. Thus Wasserstein gradient flow is the Fokker-Planck equation

$$\partial_t \rho = \nabla \cdot (\nabla \rho + \rho \nabla U).$$

Q: Is there a related model where the velocity makes sense for particles?

Blob model

KL-divergence $$\mathrm{KL}(\rho) = \int \ln\left(\frac{\rho}{\pi}\right) \rho \, dx$$

Fokker-Planck equation $$\partial_t \rho = \nabla \cdot (\rho \nabla (\ln \rho + U))$$

Q: Is there a related model where the velocity makes sense for particles?
A1: Blob model by *Carrillo, Craig, and Patacchini*, 2019: Regularize ρ in the KL divergence, using a mollifier η_ε.

$$E_\varepsilon(\rho) = \int \ln\left(\frac{\rho * \eta_\varepsilon}{\pi}\right) \rho \, dx.$$

Wasserstein gradient flow

$$\partial_t \rho = \nabla \cdot (\rho \nabla (\ln(\rho * \eta_\varepsilon) + U)).$$

- Particle ODE give a true solution of the equation.

Blob model (cont.)

Blob model by *Carrillo, Craig, and Patacchini*, 2019:

$$E_\varepsilon(\rho_\varepsilon) = \int \ln\left(\frac{\rho_\varepsilon * \eta_\varepsilon}{\pi}\right) \rho_\varepsilon \, dx.$$

Wasserstein gradient flow

$$\partial_t \rho_\varepsilon = \nabla \cdot (\rho_\varepsilon \nabla(\ln(\rho_\varepsilon * \eta_\varepsilon) + U)).$$

- Particle ODE give a true solution of the equation.
- Model introduces a bias. Let π_ε be a minimizer.
  Lu, S., Wang, 2023 show $d_2(\pi, \pi_\varepsilon) \lesssim \varepsilon$.
- Convergence of $\rho_\varepsilon(t) \to \rho(t)$ as $\varepsilon \to 0$. [*Carrillo, Craig, and Patacchini*; *Craig, Jacobs, Topalova*]

Open problems/issues:

- Convergence of $\rho_\varepsilon(t)$ as $t \to \infty$.
- Convergence of $\rho_\varepsilon(\infty)$ as $\varepsilon \to 0$.
- Model is not viable in high dimensions.

Hellinger distance

$$d_H^2(\rho_0,\rho_1) = \inf_{(\rho_t,u_t)} \int_0^1 \int_{\mathbb{R}^d} u_t^2 \,\mathrm{d}\rho_t \,\mathrm{d}t,$$

where (ρ_t, u_t) satisfies the equation $\partial_t \rho_t = -\rho_t u_t$. If measures $\rho_0, \rho_1 \ll \lambda$ for some probability measure $\mathrm{d}\lambda(x)$, then

$$d_H^2(\rho_0,\rho_1) = 4\int_{\mathbb{R}^d} \left(\sqrt{\frac{\mathrm{d}\rho_1}{\mathrm{d}\lambda}} - \sqrt{\frac{\mathrm{d}\rho_0}{\mathrm{d}\lambda}}\right)^2 \mathrm{d}\lambda.$$

Restricted to probability measures

$$d_{SH}(\rho_0,\rho_1) = 4\arcsin\left(\frac{d_H(\rho_0,\rho_1)}{4}\right).$$

Pure birth-death dynamics is the gradient flow of KL divergence wrt d_{SH}.

$$\partial_t \rho_t = -\rho_t \log\frac{\rho_t}{\pi} + \rho_t \int_{\mathbb{R}^d} \rho_t \log\frac{\rho_t}{\pi}\,\mathrm{d}x.$$

Birth-death dynamics - convergence as $t \to \infty$.

Pure birth-death dynamics is the gradient flow of KL divergence wrt d_{SH}.

$$\partial_t \rho_t = -\rho_t \log \frac{\rho_t}{\pi} + \rho_t \int_{\mathbb{R}^d} \rho_t \log \frac{\rho_t}{\pi} \, \mathrm{d}x.$$

Lu, Lu, Nolen and *Lu, S., Wang* establish

Theorem. If $\inf_{x \in \Omega} \frac{\rho_0(x)}{\pi(x)} \geqslant e^{-M}$ then

$$\mathrm{KL}(\rho_t|\pi) \leqslant e^{-(2-3\delta)(t-t_*)} \, \mathrm{KL}(\rho_0|\pi)$$

for every $\delta \in (0, 1/4)$ and all $t \geqslant t_* := \log(M/\delta^3)$.

Regularization and particle based approximations:

$$\mathcal{F}_\varepsilon(\rho) = \int \rho \log(K_\varepsilon * \rho) - \int \rho \log \pi = \int \rho \log(K_\varepsilon * \rho) + \int \rho V.$$

$$\partial_t \rho^\varepsilon = -\rho^\varepsilon \left[\log\left(\frac{K_\varepsilon * \rho^\varepsilon}{\pi} \right) + K_\varepsilon * \left(\frac{\rho^\varepsilon}{K_\varepsilon * \rho^\varepsilon} \right) - \int \log\left(\frac{K_\varepsilon * \rho^\varepsilon}{\pi} \right) \rho^\varepsilon - 1 \right].$$

Birth-death dynamics - convergence as $\varepsilon \to 0$.

Results for dynamics of positive measures on torus:

(i) Regularized flow is well posed up to time $T_\varepsilon \to \infty$ as $\varepsilon \to 0$.

(ii) Solutions $\rho^\varepsilon \to \rho$ on finite time intervals.

(iii) If $\tau_\varepsilon < T_\varepsilon$ and $\tau_\varepsilon \to \infty$ then $\rho^\varepsilon(\tau_\varepsilon) \to \pi$.

Open problems:

(i) Long time existence of L^1 solutions

(ii) Well posedness of measure-valued solutions

(iii) Convergence of particle-based schemes.

(iv) Adding diffusion

Stein Variational Gradient Descent

Consider **Kullback-Leibler divergence**, that is the relative entropy

$$\mathrm{KL}(\rho) = \int \ln\left(\frac{\rho}{\mu}\right) \rho \, dx.$$

Wasserstein gradient flow is given by $\partial_t \rho = -\nabla \cdot (\rho v)$, where the vector field v minimizes the Rayleigh functional

$$R(v) = \frac{1}{2} g_\rho(v, v) + \frac{\delta \mathrm{KL}}{\delta \rho}[v] \;\; = \frac{1}{2}\int |v|^2 \rho(x) dx - \int (\ln \rho + U) \nabla \cdot (\rho v) dx$$

where $\mu = C \exp(-U)$. Minimizing over v identifies the Wasserstein gradient flow as the Fokker-Planck equation

$$\partial_t \rho = \nabla \cdot (\nabla \rho + \rho \nabla U).$$

Liu, Wang (2016) introduced Stein Variational Gradient Descent: $g(v, v) = \|v\|^2_{H_K}$, that is gradient descent vector minimizes

$$R(v) = \frac{1}{2}\|v\|^2_{H_K} + \frac{\delta \mathrm{KL}}{\delta \rho}[v].$$

Stein Variational Gradient Descent

Consider **Kullback-Leibler divergence**, that is the relative entropy

$$\mathrm{KL}(\rho) = \int \ln\left(\frac{\rho}{\mu}\right)\rho dx.$$

For Stein Variational Gradient Descent $g(v, v) = \|v\|^2_{H_K}$ that is gradient descent vector minimizes

$$R(v) = \frac{1}{2}\|v\|^2_{H_K} + \frac{\delta\,\mathrm{KL}}{\delta\rho}[v]$$

which, for $\mu \sim e^{-U}$ leads to

(SVGD) $$\partial_t \rho = \nabla \cdot ((\nabla K * \rho + K * (\rho \nabla U))\rho)$$

Note that the equation makes sense for discrete measures $\rho = \mu_n$

(SVGD$_n$) $$\dot{x}_i = -\frac{1}{n}\sum_{j=1}^{n} \nabla K(x_i - x_j) - \frac{1}{n}\sum_{j=1}^{n} K(x_i - x_j)\nabla U(x_j).$$

Stein Variational Gradient Descent II

Lu, Lu, and Nolen

Theorem 1. For K smooth and ρ_0 smooth with $\mathrm{KL}(\rho_0) < \infty$ the solution of (SVGD) satisfies

$$\rho(t) \to \pi \text{ weakly} \quad \text{as} t \to \infty.$$

There is no rate known. Linearization (*Duncan, Nüsken, Szpruch*) indicates that the convergence is not exponential.

Theorem 2. For K smooth if $\rho_n(0) \to \rho(0)$ in d_2 then for all $t > 0$

$$\rho_n(t) \to \rho(t) \quad \text{as } n \to \infty.$$

Numerical Quantization Errors

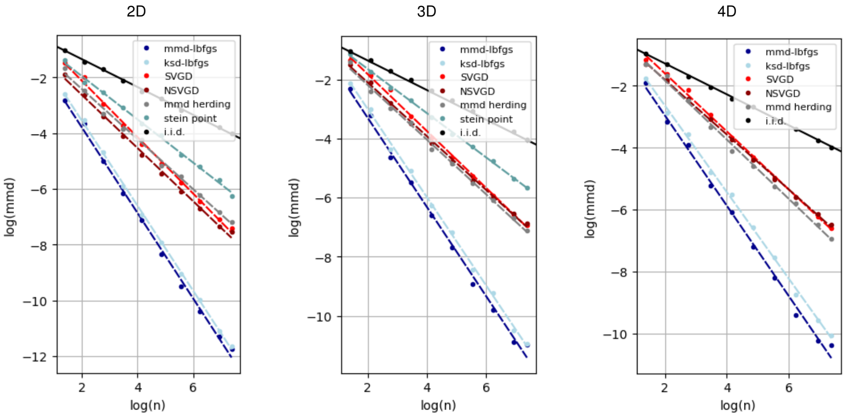


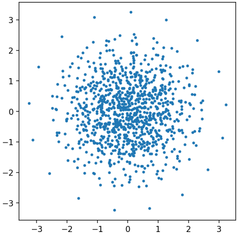
Figure: Quantization rates of the algorithms at study when $\pi = \mathcal{N}(0, \frac{1}{d} I_d)$. MMD/KSD Descent use bandwidth 1; SVGD use Laplace kernel; NSVGD use Laplace kernel with adaptive choice of bandwidth.

More Numerical quantization Errors

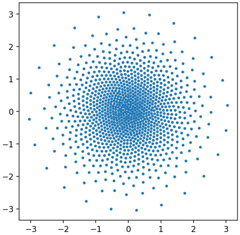
d	**Eval.**	**SVGD**	**NSVGD**	**MMD-lbfgs**	**KSD-lbfgs**	**KH**	**SP**
2	**KSD**	-0.98	-0.94	-1.48	-1.46	-0.84	-0.77
	MMD	-1.04	-1.00	-1.60	-1.54	-0.93	-0.77
3	**KSD**	-0.91	-0.81	-1.38	-1.44	-0.84	-0.78
	MMD	-0.96	-0.91	-1.51	-1.49	-0.92	-0.75
4	**KSD**	-0.91	-0.81	-1.35	-1.39	-0.89	–
	MMD	-0.94	-0.89	-1.46	-1.40	-0.95	–
8	**KSD**	-0.84	-0.80	-1.14	-1.16	–	–
	MMD	-0.77	-0.90	-1.25	-1.13	–	–

Table: Slopes for the quantization measured in KSD/MMD, for the different algorithms at study and several dimensions *d*.

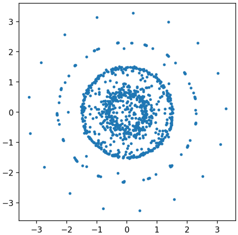
Final States



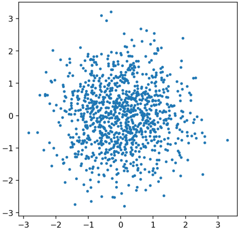
(a) SVGD Gaussian



(b) NSVGD Laplace



(c) MMD-lbfgs



(d) i.i.d.

Testing with different bandwidths

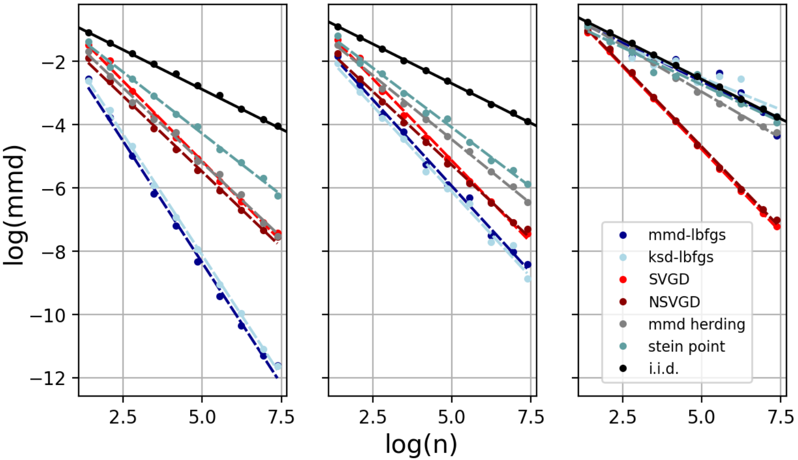


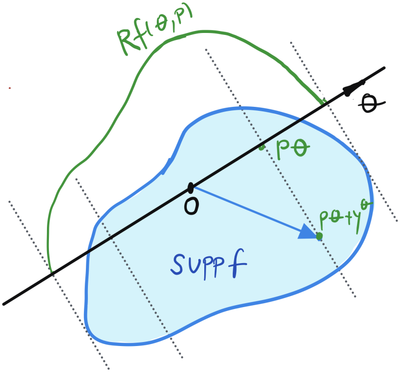
Figure: Changing the bandwidth in MMD evaluation metric when, in 2D. From Left to Right: (evaluation) MMD bandwidth = 1, 0.7, 0.3.

Radon transform

Radon Transform

For $\theta \in \mathbb{S}^{d-1}$ and $p \in \mathbb{R}$

$$Rf(\theta, p) = \widehat{f}(\theta, p) := \int_{\theta^\perp} f(p\theta + y^\theta)\, dy^\theta,$$



Sliced Wasserstein distance

Radon Transform

For $\theta \in \mathbb{S}^{d-1}$ and $p \in \mathbb{R}$

$$Rf(\theta, p) = \widehat{f}(\theta, p) := \int_{\theta^\perp} f(p\theta + y^\theta)\, dy^\theta,$$

Sliced Wasserstein distance

For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$

$$SW^2(\mu, \sigma) = \int_{\mathbb{S}^{d-1}} W^2(\mathrm{P}^\theta_\# \mu, \mathrm{P}^\theta_\# \sigma)\, d\theta = \int_{\mathbb{S}^{d-1}} W^2(\widehat{\mu}^\theta, \widehat{\sigma}^\theta)\, d\theta,$$

where $\mathrm{P}^\theta(x) = (x \cdot \theta)\theta$.

Bonnotte '13 (on bounded domains) and *Bayraktar and Guo '21* on $\mathbb{R}^d$ show that W and SW induce the same topology on $\mathcal{P}_2$.

Particle approximation error of W and SW

- **d** – a metric or general dissimilarity measure on $\mathcal{P}(\mathbb{R}^d)$ or its subset [Wasserstein metric, Sliced Wasserstein, MMD, etc.]
- $\mu \in \mathcal{P}_2(\mathbb{R}^d)$

Random quantization error

$$\mathcal{Q}_R(n, \mathbf{d}) = E[\mathbf{d}(\mu, \mu_n)]$$

where $\mu_n = \frac{1}{n}\sum_i \delta_{x_i}$ and $x_i \sim \mu$ are i.i.d samples of μ.

For μ with bounded support, with density bounded from below

$$\mathcal{Q}_R(n, W) \sim \begin{cases} n^{-\frac{1}{2}}(\log n)^{\frac{1}{2}} & \text{if } d = 2 \\ n^{-\frac{1}{d}} & \text{if } d \geqslant 3. \end{cases}$$

$$\mathcal{Q}_R(n, SW) \sim n^{-\frac{1}{2}} \quad \text{for all } d.$$

Background of Radon Transform

$$Rf(\theta, p) = \int_{\theta^\perp} f(p\theta + y^\theta)\, dy^\theta, \qquad R^* g(x) = \check{g}(x) := \int_{\mathbb{S}^{d-1}} g(\theta, x \cdot \theta)\, d\theta.$$

$$\langle Rf, g \rangle_{L^2(\mathbb{P}_d)} = \langle f, R^* g \rangle_{L^2(\mathbb{R}^d)}$$

Attenuated Sobolev Spaces

$$\|f\|^2_{H^s_t(\mathbb{R}^d)} = \int_{\mathbb{R}^d} |\xi|^{2t}(1 + |\xi|^2)^{s-t} |\mathcal{F}_d f(\xi)|^2\, dy$$

Theorem (Sharafutdinov)

For $s \in \mathbb{R}$ and $t > -\frac{d}{2}$ Radon transform is an isometry between $H^s_t(\mathbb{R}^d)$ and $H^{s+(d-1)/2}_{t+(d-1)/2}(\mathbb{P}_d)$:

$$\|f\|_{H^s_t(\mathbb{R}^d)} = \|Rf\|_{H^{s+(d-1)/2}_{t+(d-1)/2}(\mathbb{P}_d)}.$$

Consequently $\|f\|_{H^{-(d-1)/2}_{-(d-1)/2}(\mathbb{R}^d)} = \|Rf\|_{L^2(\mathbb{P}_d)}$.

Let Λ be given by

$$\Lambda = \begin{cases} (-i)^{d-1} \frac{\partial^{d-1}}{\partial p^{d-1}} & \text{when } d \text{ is odd} \\ (-i)^{d-1} \mathcal{H}_p \frac{\partial^{d-1}}{\partial p^{d-1}} & \text{when } d \text{ is even} \end{cases}$$

where $\mathcal{H}_p$ is the Hilbert transform in p variable.

Inversion formula for the Radon transform: on $\mathcal{S}(\mathbb{R}^d)$

$$f = c_d R^* \Lambda R f.$$

Local geometry of Sliced Wasserstein distance

- Metric derivative: Formally

$$\frac{SW^2(\mu_t,\mu_{t+h})}{h^2} = \int_{\mathbb{S}^{d-1}} \frac{W^2(\widehat{\mu}_t^\theta,\widehat{\mu}_{t+h}^\theta)}{h^2}\,d\theta \xrightarrow{h\to 0} \int_{\mathbb{S}^{d-1}} |\widehat{\mu}'(\theta,\cdot)|_W^2\,d\theta.$$

- If $\partial\mu + \mathrm{div}(\mu v) = 0$ then $\partial_t\widehat{\mu}^\theta + \mathrm{div}_p(\widehat{\mu}^\theta \Pi_\mu^\theta v) = 0$ and

$$|\mu'|_{SW}^2(t) = \int_{\mathbb{S}^{d-1}} |\widehat{\mu}'(\theta,\cdot)|_W^2(t)\,d\theta = \left\|\theta\cdot\frac{d\widehat{v_t\mu_t}}{d\widehat{\mu}_t}\right\|_{L^2(\widehat{\mu}_t)}^2.$$

- We define $B_{SW}(\mu,J) = \left\|\theta\cdot\frac{d\widehat{J}}{d\widehat{\mu}}\right\|_{L^2(\widehat{\mu};\mathbb{P}_d)}^2$.

Local geometry of Sliced Wasserstein distance

- Recall $|\mu'|^2_{SW}(t) = \fint_{\mathbb{S}^{d-1}} |\widehat{\mu}'(\theta,\cdot)|^2_W(t)\, d\theta$. The geodesic must satisfy $\widehat{\mu}^\theta_t = [\widehat{\mu}^\theta_0, \widehat{\mu}^\theta_1]_t$, where $[\,\cdot\,,\,\cdot\,]_t$ is displacement interpolation in 1D. However $\mu_t := R^{-1}([\widehat{\mu}^\theta_0, \widehat{\mu}^\theta_1]_t)$ may fail to be nonnegative!
- $(\mathcal{P}_2, SW)$ is not a geodesic length space. Let $\ell_{sw}(\mu_0, \mu_1)$ be the minimal length of curves connecting μ_0, μ_1.
- If $\mu_t := \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$ then

$$|\mu'|^2_{SW}(t) = \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{S}^{d-1}} |\theta\cdot x_i'(t)|^2\, d\theta = \frac{1}{d}|\mu'|^2_W(t).$$

So when restricted do discrete measures $\ell_{SW,discrete} = \frac{1}{d} d_W$.

Theorem

Let $\mu, \nu, \lambda \in \mathcal{P}ac, 2(\mathbb{R}^d)$. Assume $0 < a \leqslant b < \infty$ such that

$$a\widehat{\lambda}^\theta \leqslant \widehat{\mu}^\theta \leqslant b\widehat{\lambda}^\theta \text{ and } \widehat{\nu}^\theta \leqslant b\widehat{\lambda}^\theta \quad \text{for a.e. } \theta \in \mathbb{S}^{d-1}.$$

(i) If λ is log-concave then $\ell_{SW}(\mu, \nu) \leqslant 2\sqrt{\frac{b}{a}} SW(\mu, \nu)$.

(ii) Assume $\|\widehat{\lambda}^\theta\|_{L^\infty(\mathbb{P}_d)} \leqslant C_\lambda$. Then

$$\sqrt{\frac{1}{bC_\lambda}} \|\mu - \nu\|_{\dot{H}^{-(d+1)/2}(\mathbb{R}^d)} \leqslant SW(\mu, \nu).$$

If further $\mu = \nu$ on a δ strip of $\partial\Omega$ then

$$\sqrt{\frac{1}{bC_\lambda}} \|\mu - \nu\|_{\dot{H}^{-\frac{(d+1)}{2}}} \leqslant SW(\mu, \nu) \leqslant \ell_{SW}(\mu, \nu) \leqslant \frac{C}{\sqrt{a}} \|\mu - \nu\|_{\dot{H}^{-\frac{(d+1)}{2}}}.$$

Let f^θ and F^θ the density and the CDF of $\widehat{\mu}_\theta$,

$$SJ_2(\mu) = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \frac{F^\theta(r)(1 - F^\theta(r))}{f^\theta(r)} \, dr \, d\theta$$

Theorem

Assume $\widehat{\mu}_\theta \ll \mathcal{L}_1$ and $SJ_2(\mu) < \infty$. Let $\mu^n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ where X_i are i.i.d samples of μ. Then

$$\ell_{sw}(\mu^n, \mu) \leqslant c\sqrt{SJ_2(\mu)} \, \frac{\log n}{\sqrt{n}} \quad \text{with high probability.}$$

SW near particle measures; SW gradient flows

Lemma

Assume $\mu = \sum_{i=1}^n m_i \delta_{y_i}$. Let $L = \min_{i \neq j} |y_i - y_j|$, Then there exists $C > 0$ only dependent on d such that if $W_\infty(\mu, \nu) < \frac{L}{2}$ then

$$0 \leqslant \frac{1}{d} W_2^2(\mu, \nu) - SW_2^2(\mu, \nu) \leqslant C n W_\infty(\mu, \nu) SW_2^2(\mu, \nu).$$

- "Gradient flows" **in** Sliced Wasserstein metric are high order integro-differential equations
- Gradient flows **of** Sliced Wasserstein metric with respect to Wasserstein metric are of interest in generative sampling (Bonnotte '13, Li, Moosmüller '23, Tanguy, Flamary, Delon, '23)

Sliced Wasserstein quantization

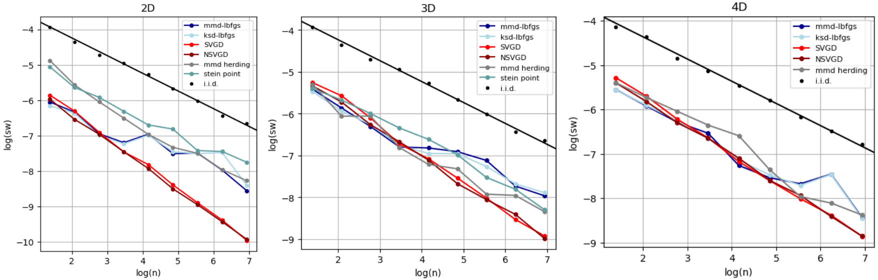


Figure: Quantization measured in Sliced Wasserstein distance for $\mu = \mathcal{N}(0, \frac{1}{d} I_d)$. States are same as before. In practice, we use 50 random directions drawn uniformly on $\mathbb{S}^{d-1}$. Slopes of red lines are -0.71, -0.64, and -0.61 in 2,3, and 4D, respectively.

Open Problem: Establish the optimal quantization rate in Sliced Wasserstein metric. Theoretical prediction for grids $n^{-\frac{1}{2}-\frac{1}{2d}}$.

Radon-Wasserstain metric

Given unit vector θ let

$$g_\theta(w, w) = \begin{cases} \int u(x\cdot\theta))^2 d\rho(x) & \text{if } w_\theta(x) = \theta u(\theta\cdot x) \\ \infty & \textit{otherwise} \end{cases}$$

Consider the *Radon-Wasserstein* metric $\overline{g}$ given by

$$\begin{aligned} \overline{g}(v, v) &= \inf\left\{ \int_{\mathbb{S}^{d-1}} g_\theta(w_\theta, w_\theta) d\theta \;:\; v(x) = \int_{\mathbb{S}^{d-1}} w_\theta dS(\theta) \right\} \\ &= \inf\left\{ \int_{\mathbb{S}^{d-1}} g_\theta(w_\theta, w_\theta) d\theta \;:\; v = \vec{R}^* w \right\} \end{aligned}$$

The resulting distance $\overline{d}$ satisfies $d_W \leqslant \overline{d} \lesssim d_W$. However the geodesics often do not exist.

Projected KL gradient flow

Consider $\pi \sim e^{-U}$. We want to determine the gradient flow of

$$E(\rho) = \int \log \frac{\rho}{\pi} d\rho$$

with respect to $\overline{g}$. Given unit vector θ, for $s \in \mathbb{R}$ Radon transform

$$R_\theta(s) = \int_{\theta^\perp} f(s\theta + y)dy$$

Gradient flow is given by

$$\begin{aligned}
&\partial_t \rho + \nabla \cdot (\rho v) = 0 \\
&v = -\int_{\mathbb{S}^{d-1}} \theta \left(\partial_s \ln(R_\theta \rho) + \frac{R_\theta(\rho \nabla U \cdot \theta)}{R_\theta \rho} \right) (x \cdot \theta) d\theta \\
&\quad = -R^* \nabla \ln(R\rho) + R^* \frac{R_\theta(\rho \nabla U \cdot \theta)}{R_\theta \rho}
\end{aligned}$$

Particle approximation of projected KL gradient flow

Gradient flow

$$\partial_t \rho + \nabla \cdot (\rho v) = 0$$

$$v = -\int_{\mathbb{S}^{d-1}} \theta \left(\partial_s \ln(R_\theta \rho) + \frac{R_\theta(\rho \nabla U \cdot \theta)}{R_\theta \rho} \right) (x \cdot \theta) d\theta$$

Consider $\rho_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$. From $R_\theta \rho_n = \frac{1}{n} \sum_i \delta_{x_i \cdot \theta}$ we approximate projected density $R_\theta \rho$ using a 1D KDE. Accuracy does not decay with d!

$$\dot{x}_i = -\int_{\mathbb{S}^{d-1}} \theta \left(\partial_s \ln(K * (R_\theta \rho_n)) + \frac{K * R_\theta(\rho_n \nabla U \cdot \theta)}{K * R_\theta \rho_n} \right) (x_i \cdot \theta) d\theta$$

We approximate the above by taking a random angle θ at each step

$$x_i(\Delta t) = x_i(0) + \Delta t \, \theta \, \frac{\sum_j K'((x_j - x_i) \cdot \theta) + K((x_j - x_i) \cdot \theta) D_\theta U(x_j)}{\sum_j K((x_j - x_i) \cdot \theta)}$$

For convergence, complexity of each step is O(n), up to logarithms!

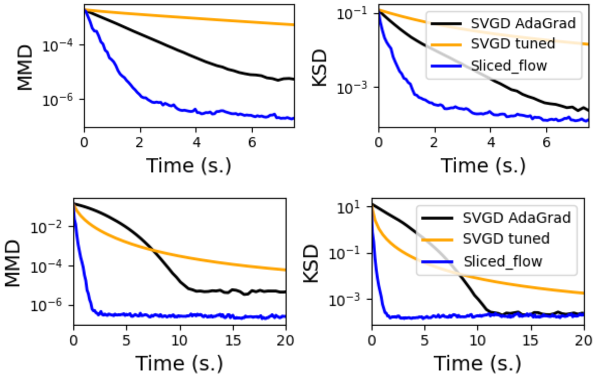


Figure: MMD and KSD convergence rates in 2D, using processor time. We run SVGD with bandwidth 0.7 and sliced flow with bandwidth 0.3. There are 1024 particles. Above: Initial distribution is the uniform distribution in a ball, target is Gaussian. Below: Initial distribution is a Gaussian centered at $(0, 2)$ while target distribution is a Gaussian mixture, centered at $(1, 0)$ and $(-1, 0)$.

Final States

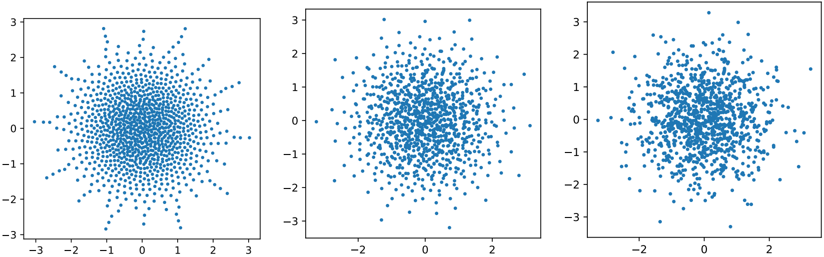


Figure: Sampling 2-dimension normal distribution, final states. From column 1 to 3: kernel bandwidth 0.1, 0.3 and 1. We ran algorithms for 50,000 steps, take timestep 0.03. The number of particles is 1024.

Approximation Error

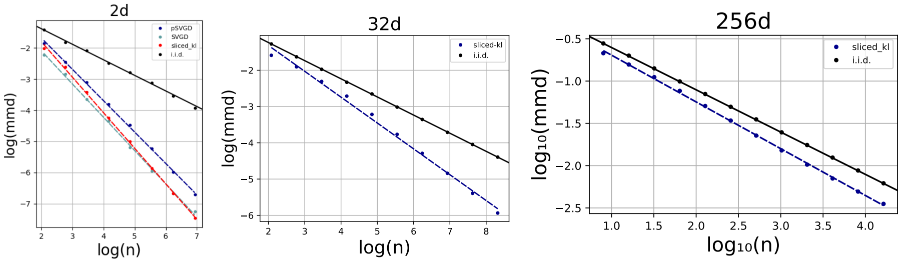


Figure: The target distribution is Gaussian. We note that sliced-KL flow does not result in variance collapse. [Variance is approximated very well.]

Projection based metric 2

Let H_K be an RKHS on $\mathbb{R}$. Given unit vector θ instead of

$$g_\theta(w,w) = \begin{cases} \int u(x\cdot\theta))^2 d\rho(x) & \text{if } w(x) = \theta u(\theta\cdot x) \\ \infty & \textit{otherwise} \end{cases}$$

consider

$$g_\theta(w,w) = \begin{cases} \|u\|_{H_K}^2 & \text{if } w(x) = \theta u(\theta\cdot x) \\ \infty & \textit{otherwise}. \end{cases}$$

Consider the *sliced* metric $\overline{g}$ given as

$$\overline{g}(v,v) = \inf\left\{ \int_{S^{d-1}} g_\theta(w_\theta, w_\theta)\, d\theta \;:\; v = \int_{S^{d-1}} w_\theta\, dS(\theta) \right\}$$

Radon-Stein gradient flow ($\sim$SVGD)

Full gradient flow is given by

(RSVGD)
$$\partial_t \rho + \nabla \cdot (\rho v) = 0$$
$$v = -\int_{\mathbb{S}^{d-1}} \theta \, \big(K' * (R_\theta \rho) + K * (R_\theta(\rho \nabla U \cdot \theta)) \big) \, d\theta$$

Consider $\rho_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$.

$$\dot{x}_i = -\int_{\mathbb{S}^{d-1}} \theta \frac{1}{n} \sum_j \big(K'((x_j - x_i) \cdot \theta) + D_\theta U(x_j) K((x_j - x_i) \cdot \theta) \big) \, d\theta$$

We approximate the above by taking a random angle θ at each step

$$x_i(\Delta t) = x_i(0) + \Delta t \, \theta \, \frac{1}{n} \sum_j \big(K'((x_j - x_i) \cdot \theta) + D_\theta U(x_j) K((x_j - x_i) \cdot \theta) \big)$$

Existence and convergence of Projected SVGD gradient flow

$\sim$ *Lu, Lu, Nolen*

Assumption 1. $K \in C^{\infty}(\mathbb{R}, \mathbb{R})$ is positive definite, integrable, even, and K, K', K'' are bounded.
Assumption 2. $U \in C^2(\mathbb{R}^d)$ is nonnegative, coercive, and satisfies $|\nabla U| \leqslant C(1 + U)$ and $|D^2 U| \leqslant C(1 + U)$.
Assumption 3. $\int (1 + U)\rho_0 dx < \infty$.

Theorem [S. and Xu]

Under assumptions above (RSVGD) has a unique solution $\rho \in C([0, \infty), \mathcal{P})$.

Theorem [S. and Xu]

Under some mild further assumptions $\rho(t) \to \pi$ weakly as $t \to \infty$.

Open Problems

- What is the optimal quantization error for MMD (for various kernels)?
- What is a robust way to measure quantization error? [Remove sensitivity to kernel width.]
- Convergence properties of SVGD, especially for nonsmooth kernels
- Well-posedness and convergence for Radon Wasserstein KL gradient flow
- Birth-death dynamics in high dimension
- Quantitative information on convergence.