Accurate Quantization of Measures via Interacting Particle-based Optimization

Anna Korba
ENSAE/CREST

Quantization, Location, Sampling and Matching Workshop

Joint work with **Lantian Xu** (CMU) and **Dejan Slepčev** (CMU).

Outline

Quantization problem

Problem : approximate a target distribution $\pi \in \mathcal{P}(\mathbb{R}^d)$ by a finite set of n points $x_1, \dots, x_n$, e.g. to compute functionals $\int_{\mathbb{R}^d} f(x) d\pi(x)$.

The quality of the set can be measured by the integral approximation error:

$$err(x_1, \dots, x_n) = \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int_{\mathbb{R}^d} f(x) d\pi(x) \right|.$$

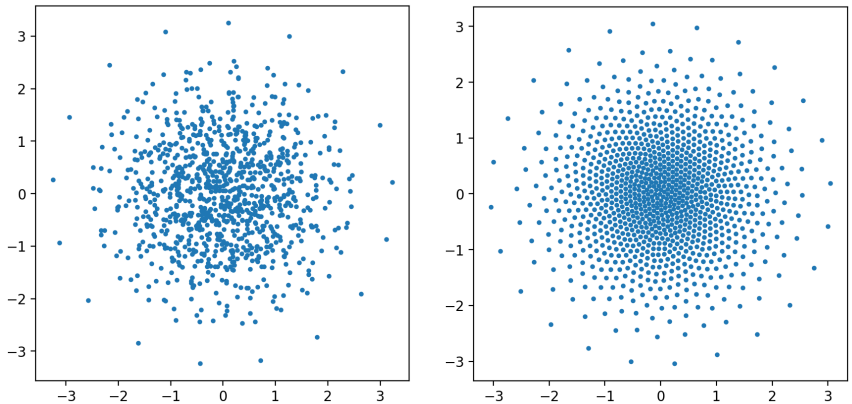
Several approaches, among which :

- MCMC methods : generate a Markov chain whose law converges to π, $err(x_1, \dots, x_n) = \mathcal{O}(n^{-1/2})$

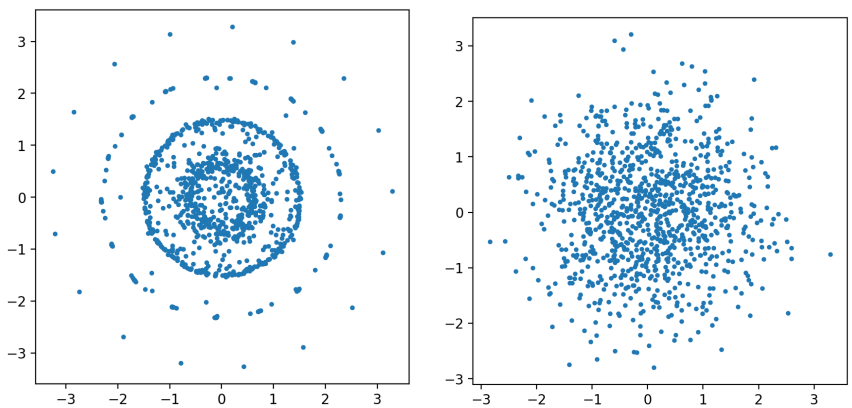
  [Łatuszyński et al., 2013]

- **deterministic particle systems**, $err(x_1, \dots, x_n)$?

Motivation



(a) SVGD Gaussian (b) NSVGD Laplace



(c) MMD-lbfgs (d) i.i.d.

Figure: (a)-(c) Final states of the algorithms for 1024 particles, after 1e4 iterations. Ring structures tend to appear with the Gaussian kernel. The kernel bandwidth for all algorithm is set to 1.

Sampling as optimization over distributions

3 algorithms/particle systems at study:

- Stein Variational Gradient Descent [Liu and Wang, 2016]
- Maximum Mean Discrepancy Descent [Arbel et al., 2019]
- Kernel Stein Discrepancy Descent [Korba et al., 2021]

These particle systems are designed to minimize a loss.

Sampling as optimization over distributions

3 algorithms/particle systems at study:

- Stein Variational Gradient Descent [Liu and Wang, 2016]
- Maximum Mean Discrepancy Descent [Arbel et al., 2019]
- Kernel Stein Discrepancy Descent [Korba et al., 2021]

These particle systems are designed to minimize a loss.

Assume that $\pi \in \mathcal{P}_2(\mathbb{R}^d) = \left\{\mu \in \mathcal{P}(\mathbb{R}^d), \int \|x\|^2 d\mu(x) < \infty\right\}$.
The sampling task can be recast as an optimization problem:

$$\pi = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \ \mathrm{D}(\mu|\pi) := \mathcal{F}(\mu),$$

where D is a **dissimilarity functional** and $\mathcal{F}$ **"a loss"**.

Starting from an initial distribution $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, one can then consider the **Wasserstein gradient flow** of $\mathcal{F}$ over $\mathcal{P}_2(\mathbb{R}^d)$ to transport μ_0 to π.

Wasserstein gradient flows (WGF) [Ambrosio et al., 2008]

The differential of $\mu \mapsto \mathcal{F}(\mu)$ evaluated at $\mu \in \mathcal{P}(\mathbb{R}^d)$ is the unique function $\frac{\partial \mathcal{F}(\mu)}{\partial \mu} : \mathbb{R}^d \to \mathbb{R}$ s. t. for any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $\nu - \mu \in \mathcal{P}(\mathbb{R}^d)$:

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon}(\mathcal{F}(\mu + \epsilon(\nu - \mu)) - \mathcal{F}(\mu)) = \int_{\mathbb{R}^d} \frac{\partial \mathcal{F}(\mu)}{\partial \mu}(x)(d\nu - d\mu)(x).$$

Wasserstein gradient flows (WGF) [Ambrosio et al., 2008]

The differential of $\mu \mapsto \mathcal{F}(\mu)$ evaluated at $\mu \in \mathcal{P}(\mathbb{R}^d)$ is the unique function $\frac{\partial \mathcal{F}(\mu)}{\partial \mu} : \mathbb{R}^d \to \mathbb{R}$ s. t. for any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $\nu - \mu \in \mathcal{P}(\mathbb{R}^d)$:

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon}(\mathcal{F}(\mu + \epsilon(\nu - \mu)) - \mathcal{F}(\mu)) = \int_{\mathbb{R}^d} \frac{\partial \mathcal{F}(\mu)}{\partial \mu}(x)(d\nu - d\mu)(x).$$

Then $\mu : [0, \infty] \to \mathcal{P}_2(\mathbb{R}^d), t \mapsto \mu_t$ satisfies a Wasserstein gradient flow of $\mathcal{F}$ if distributionnally:

$$\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \nabla_{W_2} \mathcal{F}(\mu_t)\right),$$

where $\nabla_{W_2} \mathcal{F}(\mu) := \nabla \frac{\partial \mathcal{F}(\mu)}{\partial \mu} \in L^2(\mu)$ denotes the Wasserstein gradient of $\mathcal{F}$.

Particle system approximating the WGF

Euler time-discretization : Starting from μ_0,

$$\mu_{l+1} = \left(I - \gamma \nabla_{W_2} \mathcal{F}(\mu_l)\right)_{\#} \mu_l$$

which corresponds in $\mathbb{R}^d$ to:

$$X_{l+1} = X_l - \gamma \nabla_{W_2} \mathcal{F}(\mu_l)(X_l) \sim \mu_{l+1}, \quad X_0 \sim \mu_0.$$

Space discretization/particle system : Since μ_l is unknown, introduce a particle system $X^1, \dots, X^n$ where μ_l is replaced by $\hat{\mu}_l = \frac{1}{n} \sum_{i=1}^n \delta_{X_l^i}$:

$$X_{l+1}^i = X_l^i - \gamma \nabla_{W_2} \mathcal{F}(\hat{\mu}_l)(X_l^i) \quad \text{for } i = 1, \dots, n,$$
$$X_0^1, \dots, X_0^n \sim \mu_0.$$

Question : how close is $\hat{\mu}_l$ to π at stationarity?

Outline

Background on kernels and RKHS [Steinwart and Christmann, 2008]

- Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a positive, semi-definite kernel, e.g.
  - the Gaussian kernel $k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{h}\right)$
  - the Laplace kernel $k(x, x') = \exp\left(-\frac{\|x-x'\|}{h}\right)$
  - the inverse multiquadratic kernel $k(x, x') = (c + \|x - x'\|)^{-\beta}$ with $\beta \in]0, 1[$
- $\mathcal{H}_k$ its corresponding RKHS (Reproducing Kernel Hilbert Space):

$$\mathcal{H}_k = \overline{\left\{ \sum_{i=1}^{m} \alpha_i k(\cdot, x_i);\ m \in \mathbb{N};\ \alpha_1, \ldots, \alpha_m \in \mathbb{R};\ x_1, \ldots, x_m \in \mathbb{R}^d \right\}}$$

- $\mathcal{H}_k$ is a Hilbert space with inner product $\langle ., . \rangle_{\mathcal{H}_k}$ and norm $\|.\|_{\mathcal{H}_k}$.
- It satisfies the reproducing property:

$$\forall \quad f \in \mathcal{H}_k,\ x \in \mathbb{R}^d, \quad f(x) = \langle f, k(x, .) \rangle_{\mathcal{H}_k}.$$

Maximum Mean Discrepancy [Gretton et al., 2012]

Let $\mu \in \mathcal{P}(\mathbb{R}^d)$. If $\int \sqrt{k(x,x)} d\mu(x) < \infty$, then the **kernel mean embedding** $m_\mu = \int k(x,.) d\mu(x) \in \mathcal{H}_k$.

If the map $m : \mathcal{P}(\mathbb{R}^d) \to \mathcal{H}_k, \mu \mapsto m_\mu$ is injective, it defines a distance on $\mathcal{P}(\mathbb{R}^d)$ called the Maximum Mean Discrepancy (MMD):

$$\begin{aligned}
\mathrm{MMD}^2(\mu,\pi) &= \sup_{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \leq 1} \left| \int f d\mu - \int f d\pi \right|^2 \\
&= \|m_\mu - m_\pi\|^2_{\mathcal{H}_k} \\
&= \iint_{\mathbb{R}^d} k(x,y) d\mu(x) d\mu(y) + \iint_{\mathbb{R}^d} k(x,y) d\pi(x) d\pi(y) \\
&- 2 \iint_{\mathbb{R}^d} k(x,y) d\mu(x) d\pi(y),
\end{aligned}$$

by the reproducing property $\langle f, k(x,.) \rangle_{\mathcal{H}_k} = f(x)$ for $f \in \mathcal{H}_k$.

Maximum Mean Discrepancy - remarks

- The MMD writes as a sum of integrals, hence it can be estimated as soon as one has access to samples of μ and π,
- It enables to bound the integral approximation error for functions in the RKHS, since by the reproducing property and Cauchy-Schwartz:

$$\left| \int_{\mathbb{R}^d} f(x) d\pi(x) - \int_{\mathbb{R}^d} f(x) d\mu(x) \right| \leq \|f\|_{\mathcal{H}_k} \, \mathrm{MMD}(\mu, \pi).$$

Kernel Stein Discrepancy [Chwialkowski et al., 2016, Liu et al., 2016]

If one does not have access to samples of π but only to its score, it is still possible to compute the KSD. For $\mu, \pi \in \mathcal{P}(\mathbb{R}^d)$, the KSD of μ relative to π is defined by

$$\mathrm{KSD}^2(\mu|\pi) = \iint k_\pi(x,y) d\mu(x) d\mu(y),$$

where $k_\pi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the **Stein kernel**, defined through

- the score function $s(x) = \nabla \log \pi(x)$,
- a p.s.d. kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $k \in C^2(\mathbb{R}^d)$[1]

For $x, y \in \mathbb{R}^d$,

$$\begin{aligned} k_\pi(x,y) &= s(x)^T s(y) k(x,y) + s(x)^T \nabla_2 k(x,y) \\ &\quad + \nabla_1 k(x,y)^T s(y) + \boldsymbol{\nabla} \cdot_1 \nabla_2 k(x,y) \\ &= \sum_{i=1}^d \frac{\partial \log \pi(x)}{\partial x_i} \cdot \frac{\partial \log \pi(y)}{\partial y_i} . k(x,y) + \frac{\partial \log \pi(x)}{\partial x_i} \cdot \frac{\partial k(x,y)}{\partial y_i} \\ &\quad + \frac{\partial \log \pi(y)}{\partial y_i} \cdot \frac{\partial k(x,y)}{\partial x_i} + \frac{\partial^2 k(x,y)}{\partial x_i \partial y_i} \in \mathbb{R}. \end{aligned}$$

KSD vs MMD

Under mild assumptions on k and π, the Stein kernel k_π is p.s.d. and satisfies a **Stein identity** [Oates et al., 2017]

$$\int_{\mathbb{R}^d} k_\pi(x, .) d\pi(x) = 0.$$

Consequently, **KSD is an MMD** with kernel k_π, since:

$$\begin{aligned}
\mathrm{MMD}^2(\mu|\pi) &= \int k_\pi(x,y) d\mu(x) d\mu(y) + \int k_\pi(x,y) d\pi(x) d\pi(y) \\
&\quad - 2 \int k_\pi(x,y) d\mu(x) d\pi(y) \\
&= \int k_\pi(x,y) d\mu(x) d\mu(y) \\
&= \mathrm{KSD}^2(\mu|\pi)
\end{aligned}$$

KSD as kernelized Fisher Divergence

Fisher Divergence:

$$\mathrm{FD}^2(\mu|\pi) = \left\|\nabla \log\left(\frac{\mu}{\pi}\right)\right\|^2_{L^2(\mu)} = \int \|\nabla \log\left(\frac{\mu}{\pi}(x)\right)\|^2 d\mu(x)$$

"Kernelized" with k:

$$\begin{aligned}\mathrm{KSD}^2(\mu|\pi) &= \left\|S_{\mu,k}\nabla \log\left(\frac{\mu}{\pi}\right)\right\|^2_{\mathcal{H}_k} \\ &= \int \nabla \log\left(\frac{\mu}{\pi}\right)(x) k(x,y) \nabla \log\left(\frac{\mu}{\pi}\right)(y) d\mu(x) d\mu(y)\end{aligned}$$

where $S_{\mu,k} : L^2(\mu) \to \mathcal{H}_k, f \mapsto \int k(x,.)f(x)d\mu(x)$.

$\Longrightarrow$ minimizing the KSD is close in spirit to score-matching [Hyvärinen and Dayan, 2005].

MMD and KSD Descent [Arbel et al., 2019, Korba et al., 2021]

Recall that we want to study particle systems

$$X_{l+1}^i = X_l^i - \gamma \nabla_{W_2} \mathcal{F}(\hat{\mu}_l)(X_l^i) \quad \text{for } i = 1, \dots, n,$$

where $\hat{\mu}_l = 1/n \sum_{i=1}^n \delta_{X_l^i}$ and $\mathcal{F}(\mu) = D(\mu|\pi)$.

MMD and KSD Descent [Arbel et al., 2019, Korba et al., 2021]

Recall that we want to study particle systems

$$X_{l+1}^i = X_l^i - \gamma \nabla_{W_2} \mathcal{F}(\hat{\mu}_l)(X_l^i) \quad \text{for } i = 1, \dots, n,$$

where $\hat{\mu}_l = 1/n \sum_{i=1}^n \delta_{X_l^i}$ and $\mathcal{F}(\mu) = D(\mu|\pi)$.

For discrete measures $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{X^i}$, the MMD/KSD are well defined, hence we let $F(X^1, \dots, X^n) := \mathcal{F}(\mu)$.

MMD and KSD Descent [Arbel et al., 2019, Korba et al., 2021]

Recall that we want to study particle systems

$$X^i_{l+1} = X^i_l - \gamma \nabla_{W_2} \mathcal{F}(\hat{\mu}_l)(X^i_l) \quad \text{for } i = 1, \dots, n,$$

where $\hat{\mu}_l = 1/n \sum_{i=1}^n \delta_{X^i_l}$ and $\mathcal{F}(\mu) = D(\mu|\pi)$.

For discrete measures $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{X^i}$, the MMD/KSD are well defined, hence we let $F(X^1, \dots, X^n) := \mathcal{F}(\mu)$.

- If D is the MMD, the gradient of F is readily obtained as

$$\nabla_{x^i} F(X^1, \dots, X^N) = \frac{1}{N} \sum_{j=1}^N \nabla_2 k(X^i, X^j) - \int \nabla_2 k(X^i, x) d\pi(x).$$

- In contrast, if D is the KSD,

$$\nabla_{x^i} F(X^1, \dots, X^n) = \frac{1}{n} \sum_{j=1}^n \nabla_2 k_\pi(X^i, X^j).$$

MMD and KSD Descent [Arbel et al., 2019, Korba et al., 2021]

Recall that we want to study particle systems

$$X_{l+1}^i = X_l^i - \gamma \nabla_{W_2} \mathcal{F}(\hat{\mu}_l)(X_l^i) \quad \text{for } i = 1, \dots, n,$$

where $\hat{\mu}_l = 1/n \sum_{i=1}^n \delta_{X_l^i}$ and $\mathcal{F}(\mu) = D(\mu|\pi)$.

For discrete measures $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{X^i}$, the MMD/KSD are well defined, hence we let $F(X^1, \dots, X^n) := \mathcal{F}(\mu)$.

- If D is the MMD, the gradient of F is readily obtained as

$$\nabla_{x^i} F(X^1, \dots, X^N) = \frac{1}{N} \sum_{j=1}^N \nabla_2 k(X^i, X^j) - \int \nabla_2 k(X^i, x) d\pi(x).$$

- In contrast, if D is the KSD,

$$\nabla_{x^i} F(X^1, \dots, X^n) = \frac{1}{n} \sum_{j=1}^n \nabla_2 k_\pi(X^i, X^j).$$

MMD/KSD Descent: at each time $l \geq 0$, for any $i = 1, \dots, n$:

$$X_{l+1}^i = X_l^i - \gamma \nabla_{x^i} F(X_l^1, \dots, X_l^n).$$

Stein Variational Gradient Descent [Liu and Wang, 2016]

Stein Variational Gradient Descent (SVGD) performs gradient descent in $\mathcal{P}(\mathbb{R}^d)$ of the Kullback-Leibler (KL) divergence :

$$\mathrm{KL}(\mu|\pi) = \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) & \text{if } \mu \ll \pi \\ +\infty & \text{otherwise.} \end{cases}$$

where the (W_2) gradient is smoothed through the kernel integral operator.

It corresponds to an Euler discretization of the gradient flow of the KL under a metric depending on k [Duncan et al., 2019]:

$$W_k^2(\mu_0, \mu_1) = \inf_{\mu, v} \left\{ \int_0^1 \|v_t(x)\|^2_{\mathcal{H}_k^d} dt(x) : \frac{\partial \mu_t}{\partial t} = \nabla \cdot (\mu_t v_t) \right\}.$$

Stein Variational Gradient Descent [Liu and Wang, 2016]

Fix a reproducing kernel k. In continuous time, SVGD flow is defined by the continuity equation

$$\frac{\partial \mu_t}{\partial t} + \boldsymbol{\nabla} \cdot (\mu_t v_{\mu_t}) = 0, \; v_{\mu_t} = k \star (\mu_t \nabla \log \pi) - \nabla k \star \mu_t,$$

i.e. $v_{\mu_t} = S_{\mu_t,k} \nabla \log\left(\frac{\mu}{\pi}\right)$ where

- $\nabla \log\left(\frac{\mu}{\pi}\right) = \nabla_{W_2} \mathrm{KL}(\mu|\pi)$,
- $S_{\mu,k} : L^2(\mu) \to \mathcal{H}_k, f \mapsto \int k(x,.) f(x) d\mu(x)$.

Stein Variational Gradient Descent [Liu and Wang, 2016]

Fix a reproducing kernel k. In continuous time, SVGD flow is defined by the continuity equation

$$\frac{\partial \mu_t}{\partial t} + \boldsymbol{\nabla} \cdot (\mu_t v_{\mu_t}) = 0, \; v_{\mu_t} = k \star (\mu_t \nabla \log \pi) - \nabla k \star \mu_t,$$

i.e. $v_{\mu_t} = S_{\mu_t,k} \nabla \log\left(\frac{\mu}{\pi}\right)$ where

- $\nabla \log\left(\frac{\mu}{\pi}\right) = \nabla_{W_2} \mathrm{KL}(\mu|\pi)$,
- $S_{\mu,k} : L^2(\mu) \to \mathcal{H}_k, f \mapsto \int k(x,.) f(x) d\mu(x)$.

Let $\gamma > 0$ be a fixed step-size. Starting from $x_0^1, \ldots, x_0^n \sim \mu_0$, SVGD algorithm updates the n particles as follows at each iteration :

$$x_{l+1}^i = x_l^i - \frac{\gamma}{n} \sum_{j=1}^{n} \left[-\nabla \log \pi(x_l^j) k(x_l^i, x_l^j) + \nabla_{x_l^j} k(x_l^i, x_l^j) \right],$$

for any $i = 1, \ldots, n$.

Outline

Kernel Herding (KH) and Stein Points (SP)

They attempt to solve MMD or KSD quantization in a greedy manner, i.e. by sequentially constructing μ_ns, adding one new particle at each iteration to minimize MMD/KSD.

Kernel Herding (KH) for the MMD [Chen et al., 2012]:

$$x^{n+1} = \underset{x \in \mathbb{R}^d}{\operatorname{argmax}} \langle w_n, k(x, .) \rangle_{\mathcal{H}_k}$$

$$w_{n+1} = w_n + m_\pi - k(x_{n+1}, .)$$

citebach2012equivalence obtain a linear rate of convergence $\mathcal{O}(e^{-bn})$

- if the mean embedding $m_\pi = \mathbb{E}_{x \sim \pi}[k(x, .)]$ lies in the relative interior of the marginal polytope $convexhull(\{k(x, .), x \in \mathbb{R}^d\})$ with distance b away from the boundary
- however for infinite-dimensional kernels $b = 0$ and the rate does not hold.

Stein Points for the KSD [Chen et al., 2018] greedily minimizes the KSD similarly. The authors establish a $\mathcal{O}((\log(n)/n)^{\frac{1}{2}})$ rate, which seem slower than their empirical observations.

Contributions of our work

- We investigate the quantization properties of particle systems derived from WGF, *assuming the particles have attained a minimizer of their discrepancy objective*.
- Furthermore, as these algorithms might be difficult to tune to guarantee particles convergence, we also discuss practical improvements.

Outline

Normalized SVGD

Problem: v_{μ_t} is small where μ_t is small. This creates convergence issues, especially if the initial distribution, μ_0, is spread in space.

We introduce a normalized SVGD (NSVGD) reweighing the kernel by a kernel density estimate of μ.

Consider a translation-invariant kernel parametrized by a bandwidth $\tau > 0$: $\eta_\tau(x-y) = \eta(\frac{x-y}{\tau})$ with $\eta \in C^1(\mathbb{R}^d \setminus \{0\})$, and μ a (potentially discrete) distribution.

We now introduce a density-dependent kernel:

$$K_\mu(x,y) = \eta_\tau(x-y)\mu_h(x)^{-\frac{1}{2}}\mu_h(y)^{-\frac{1}{2}} \tag{1}$$

where μ_h denotes the smoothed density $\mu \star \eta_h$.

Normalised SVGD (NSVGD)

In the discrete setting where $\mu = {}^1/_n \sum_{i=1}^n \delta_{x_i}$, we can write the NSVGD vector field ruling the particle system as

$$v_\mu(x) = -\frac{1}{n} \sum_{j=1}^{n} \left(\mu_h(x)\mu_h(x^j)\right)^{-\frac{1}{2}} w^j(x), \tag{2}$$

where $\mu_h(x) = \frac{1}{n} \sum_{i=1}^{n} \eta_h(x - x_i)$, and

$$w^j(x) = \nabla\eta_\tau(x - x^j) - \eta_\tau(x - x^j)\nabla \log \pi(x^j) \tag{3}$$

$$+ \frac{\eta_\tau(x - x^j)}{2\mu_h(x^j)} \frac{1}{n} \sum_{m=1}^{n} \nabla\eta_h(x^j - x^m). \tag{4}$$

- the term $\mu_h(.)\mu_h(x^j)$ acts as a preconditioner,
- (3) is the vector field of the original SVGD algorithm
- (4) can be understood as a weighted repulsive term

The preconditioner accelerates or slows down the dynamic depending on the density regions and makes NSVGD less sensitive to the choice of the step-size than original SVGD.

Outline

We are interested in establishing bounds on the quantization error

$$Q_n = \inf_{x_n = x_1, \ldots, x_n} \mathrm{D}(\pi, \mu_n), \qquad \text{for } \mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i},$$

where D is the MMD or KSD.

Remark: For $x_1, \ldots, x_n \sim \pi$ i.i.d., the rate is known to be $\mathcal{O}(n^{-1/2})$ [Gretton et al., 2006, Tolstikhin et al., 2017, Lu and Lu, 2020].

We are interested in establishing bounds on the quantization error

$$Q_n = \inf_{X_n = x_1, \ldots, x_n} \mathrm{D}(\pi, \mu_n), \quad \text{for } \mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i},$$

where D is the MMD or KSD.

Remark: For $x_1, \ldots, x_n \sim \pi$ i.i.d., the rate is known to be $\mathcal{O}(n^{-1/2})$ [Gretton et al., 2006, Tolstikhin et al., 2017, Lu and Lu, 2020].

We first consider the following assumption on the Fourier transform of kernel k.

Assumption A1: Let $k(x, y) = \eta(x - y)$ a translation invariant kernel on $\mathbb{R}^d$. Assume that $\eta \in C(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$, and that its Fourier transform verifies : $\exists C_{1,d} \geq 0$ such that $(1 + |\xi|^2)^{d/2} \leq C_{1,d} |\hat{\eta}(\xi)|^{-1}$ for any $\xi \in \mathbb{R}^d$.

(Satisfied for the Gaussian and Laplace kernel.)

First result for the MMD

Theorem: Suppose A1 holds. Assume that (i) π is the Lebesgue measure or (ii) a non-negative normalized Borel measure on $[0,1]^d$. Then, there exists a constant C_d, such that for all $n \geq 2$,

- if (i): there exist points $x_1, \ldots, x_n$ such that

$$\mathrm{MMD}(\pi, \mu_n) \leq C_d \frac{(\log n)^{d-1}}{n}.$$

- if (ii): there exist points $x_1, \ldots, x_n$ such that

$$\mathrm{MMD}(\pi, \mu_n) \leq C_d \frac{(\log n)^{\frac{3d+1}{2}}}{n}.$$

Proof: Denote by $\mathcal{H}_k$ the RKHS of k, we have:

$$\mathcal{H}_k = \left\{ f \in C(\mathbb{R}^d) \cap L^2(\mathbb{R}^d), \|f\|^2_{\mathcal{H}_k} := \frac{1}{(2\pi)^{d/2}} \int |\hat{\eta}(\xi)|^{-1} |\hat{f}(\xi)|^2 d\xi < \infty \right\}.$$

We also have that the $H^d = W^{d,2}(\mathbb{R}^d)$ Sobolev norm of f is

$$\|f\|^2_{H^d} = \int (1 + |\xi|^2)^{d/2} |\hat{f}(\xi)|^2 d\xi.$$

Moreover, A1 $\Longrightarrow \exists C_{1,d}$ s.t. $\forall \xi$, $(1 + |\xi|^2)^{d/2} \leq C_{1,d} |\hat{\eta}(\xi)|^{-1}$. Hence, $\mathcal{H}_k$ continuously embeds into H^d and for any $f \in \mathcal{H}_k$, $\|f\|_{H^d} \leq \|f\|_{\mathcal{H}_k}$.

We then use a Koksma-Hlawka inequality [Aistleitner and Dick, 2015](Th1):

$$\left| \int_{[0,1]^d} f(x) d\pi(x) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right| \leq \mathcal{D}(X_n, \pi) V(f),$$

- $\mathcal{D}(X_n, \pi) = 2^d \sup_{I = \Pi_{i=1}^n [a_i, b_i]} |\pi(I) - \mu_n(I)|$ is the discrepancy of the point set X_n, can be bounded by [Aistleitner and Dick, 2015](Cor 2)
- $V(f) = \sum_{\alpha \,:\, |\alpha| \leq d} 2^{d - |\alpha|} \|\partial^\alpha f\|_{L^1(\pi)}$ is the Hardy & Krause variation of f which can be bounded by $4^d \|f\|_{H^d}$.

By the definition of MMD , we have that $\text{MMD}(\mu_n, \pi) \leq 4^d \mathcal{D}(X_n, \pi)$.

Result for non compactly supported distributions π

Proposition: Suppose A1 holds and that k is bounded. Assume π is a light-tailed distribution on $\mathbb{R}^d$ (i.e. which has a thinner tail than an exponential distribution). Then, for $n \geq 2$ there exist points $x_1, ..., x_n$ such that

$$\mathrm{MMD}(\pi, \mu_n) \leq C_d \frac{(\log n)^{\frac{5d+1}{2}}}{n}.$$

Result for non compactly supported distributions π

Proposition: Suppose A1 holds and that k is bounded. Assume π is a light-tailed distribution on $\mathbb{R}^d$ (i.e. which has a thinner tail than an exponential distribution). Then, for $n \geq 2$ there exist points $x_1, ..., x_n$ such that

$$\mathrm{MMD}(\pi, \mu_n) \leq C_d \frac{(\log n)^{\frac{5d+1}{2}}}{n}.$$

Proof: Decompose :

$$\mathrm{MMD}(\pi, \mu_n) \leq \mathrm{MMD}(\pi, \mu) + \mathrm{MMD}(\mu, \mu_n),$$

and choose μ compactly supported on $A_n = [-\log n, \log n]^d$.

As π is light-tailed, μ is close to π in L^1 distance, and we first get $\mathrm{MMD}(\pi, \mu) \leq C/n$.

Then, we can take a discrete μ_n supported on A_n and bound $\mathrm{MMD}(\mu, \mu_n)$ using similar arguments as the previous Theorem.

Result for the KSD

Theorem: Assume that k is a Gaussian kernel and that $\pi \propto \exp(-U)$ where $U \in C^\infty(\mathbb{R}^d)$ is such that $U(x) > c_1|x|$ for large enough x, there exists polynomial f with degree m such that $\|\partial^\alpha U(x)\| \leq f(x)$ for all $1 \leq |\alpha| \leq d$. Then there exist points $x_1, ..., x_n$ such that

$$\mathrm{KSD}(\mu_n|\pi) \leq C_d \frac{(\log n)^{\frac{6d+2m+1}{2}}}{n}.$$

We note that for Gaussian mixtures π, U satisfies the conditions of the theorem.

Result for the KSD

Theorem: Assume that k is a Gaussian kernel and that $\pi \propto \exp(-U)$ where $U \in C^\infty(\mathbb{R}^d)$ is such that $U(x) > c_1|x|$ for large enough x, there exists polynomial f with degree m such that $\|\partial^\alpha U(x)\| \leq f(x)$ for all $1 \leq |\alpha| \leq d$. Then there exist points $x_1, ..., x_n$ such that

$$\mathrm{KSD}(\mu_n|\pi) \leq C_d \frac{(\log n)^{\frac{6d+2m+1}{2}}}{n}.$$

We note that for Gaussian mixtures π, U satisfies the conditions of the theorem.

Proof: The proof relies on bounding the first and last term of the $\mathrm{KSD}(\mu_n, \pi)$ as the cross terms can be upper bounded by the former ones by a simple computation.

Then, the two remaining terms in the $\mathrm{KSD}(\mu_n, \pi)$ are treated independently as two $\mathrm{MMD}(\mu_n, \pi)$, with $k_1(x, y) = s(x)^T s(y) k(x, y)$ and $k_2(x, y) = \nabla \cdot_x \nabla_y k(x, y)$.

The second one is controlled by our Proposition on MMD's for bounded kernels. The first one relies on controlling $\nabla \log \pi$ Sobolev norms and our Proposition for MMD.

Outline

Algorithms

we investigate numerically the quantization properties of :

- SVGD
- Normalized SVGD
- MMD descent
- KSD Descent
- Kernel Herding (KH)
- Stein points (SP)

Hyperparameters:

- kernel: Gaussian, Laplace...
- bandwith of the kernel
- step-size

Normalised SVGD

We found that

- Laplace kernel leads to more regular configurations than Gaussian kernel
- NSVGD reaches convergences much faster than SVGD

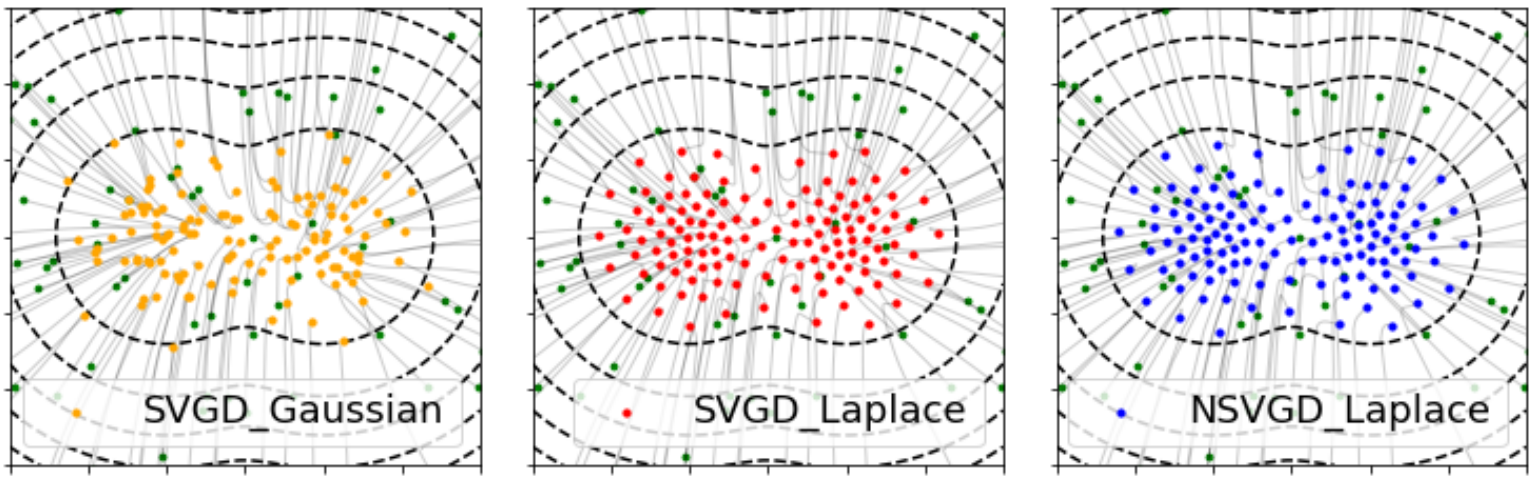


Figure: Example of a 2D Gaussian mixture. The configuration of 128 particles are plotted in green at initialization, and in different colors after convergence. The light grey curves correspond to their trajectories. From left to right: SVGD with Gaussian and Laplace kernel, γ=0.5, after 1000 iters; NSVGD with Laplace kernel and γ=0.1, after 30 iters.

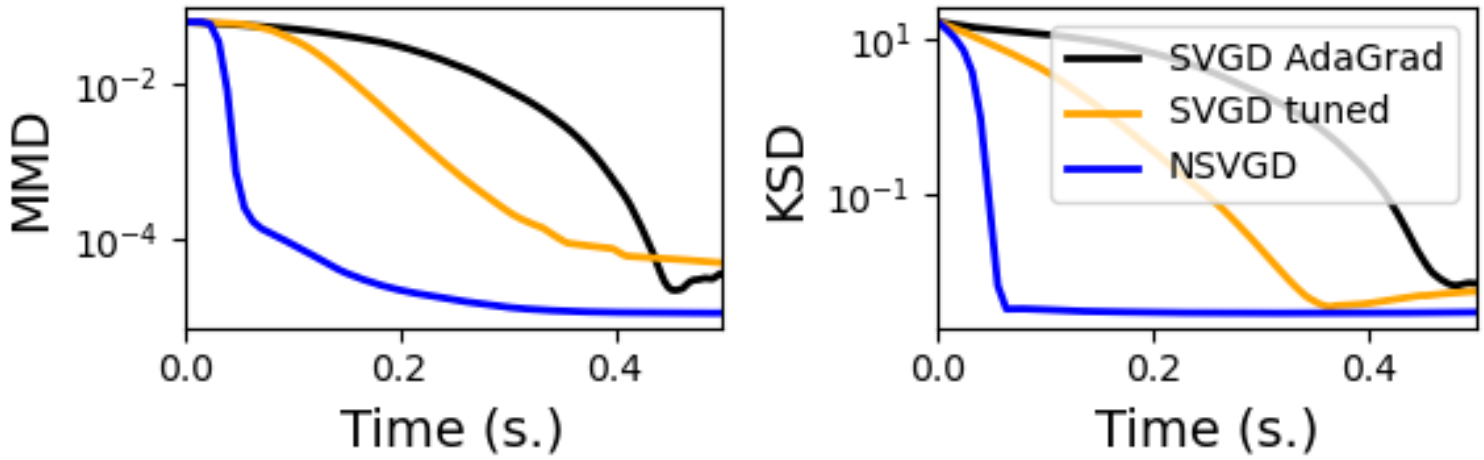
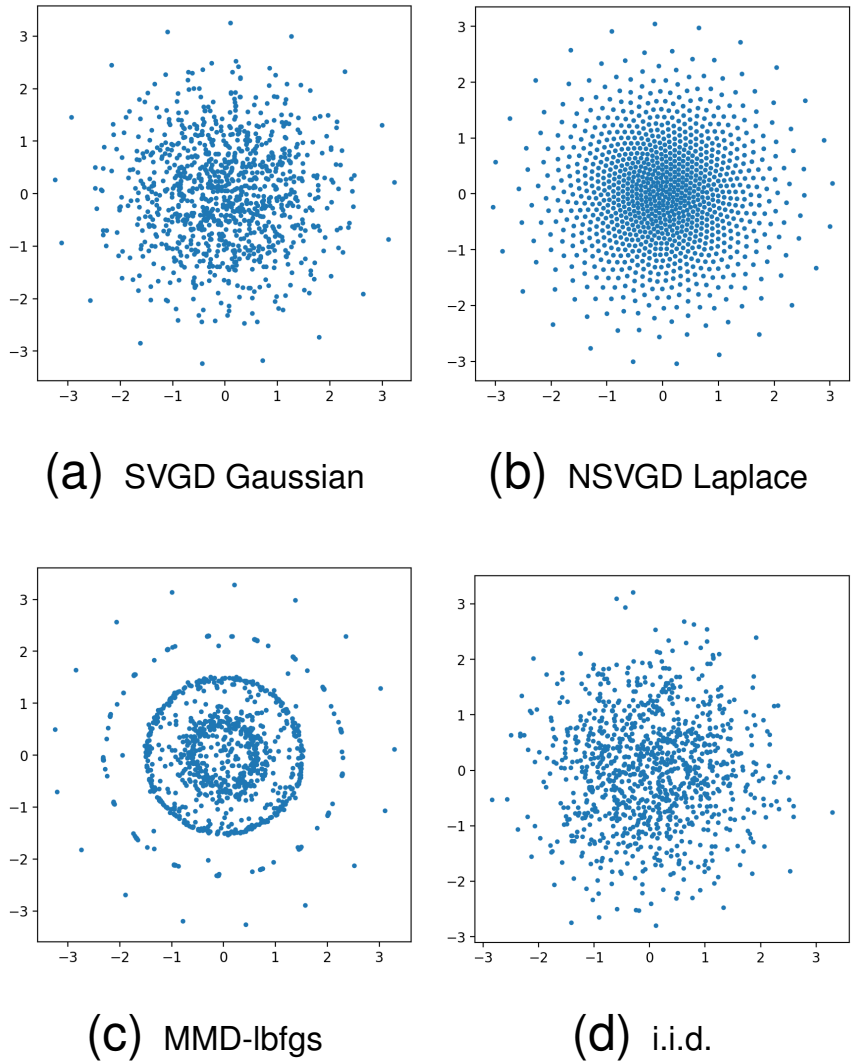


Figure: Convergence speed of SVGD (tuned time-step or AdaGrad) and Normalized SVGD (fixed time-step) on a 2D mixture of Gaussians, with 128 particles.

Final states for a Gaussian target



(a) SVGD Gaussian (b) NSVGD Laplace

(c) MMD-lbfgs (d) i.i.d.

Figure: (a)-(c) Final states of the algorithms for 1024 particles, after 1e4 iterations. Ring structures tend to appear with the Gaussian kernel. The kernel bandwidth for all algorithm is set to 1.

Quantization rates of the algorithms, $\pi = \mathcal{N}(0, 1/d I_d)$

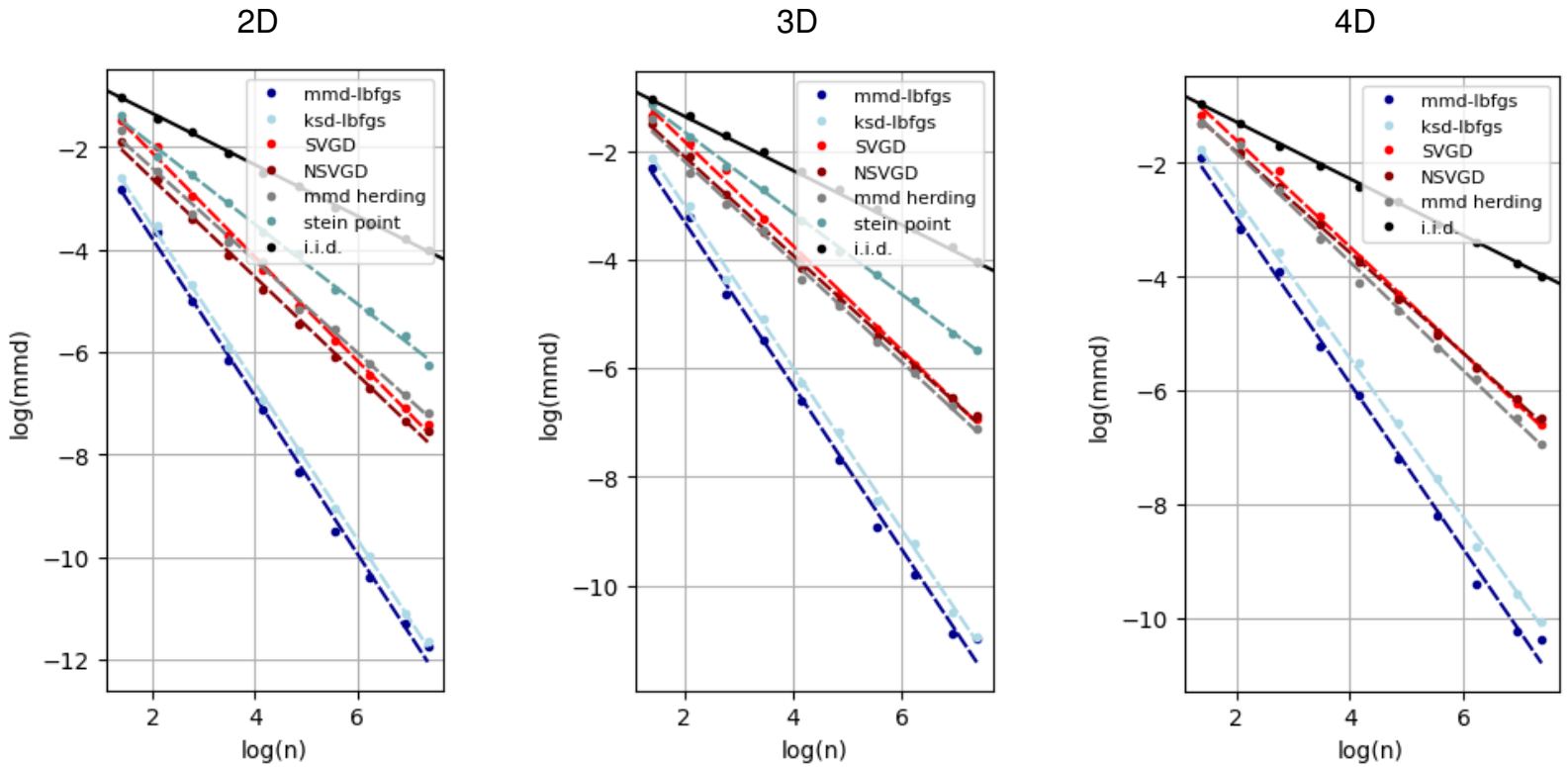


Figure: Each point is the result of averaging 3 runs of each algorithm run for 1e4 iterations, where the initial particles are i.i.d. samples of π. MMD/KSD Descent use bandwidth 1; SVGD use Laplace kernel with median trick; NSVGD use Laplace kernel with adaptive choice of bandwidth. Stein points use gridsize = 200 points in 2d, 50 in 3d; in 4d grid search was too slow.

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.

Some remarks:

- The slopes remain much steeper than the Monte Carlo rate, even when the dimension increases
- MMD/KSD Descent performs the best, but they are designed to minimize the MMD/KSD
- Their slopes are better than our theoretical upper bounds

Robustness to evaluation discrepancy

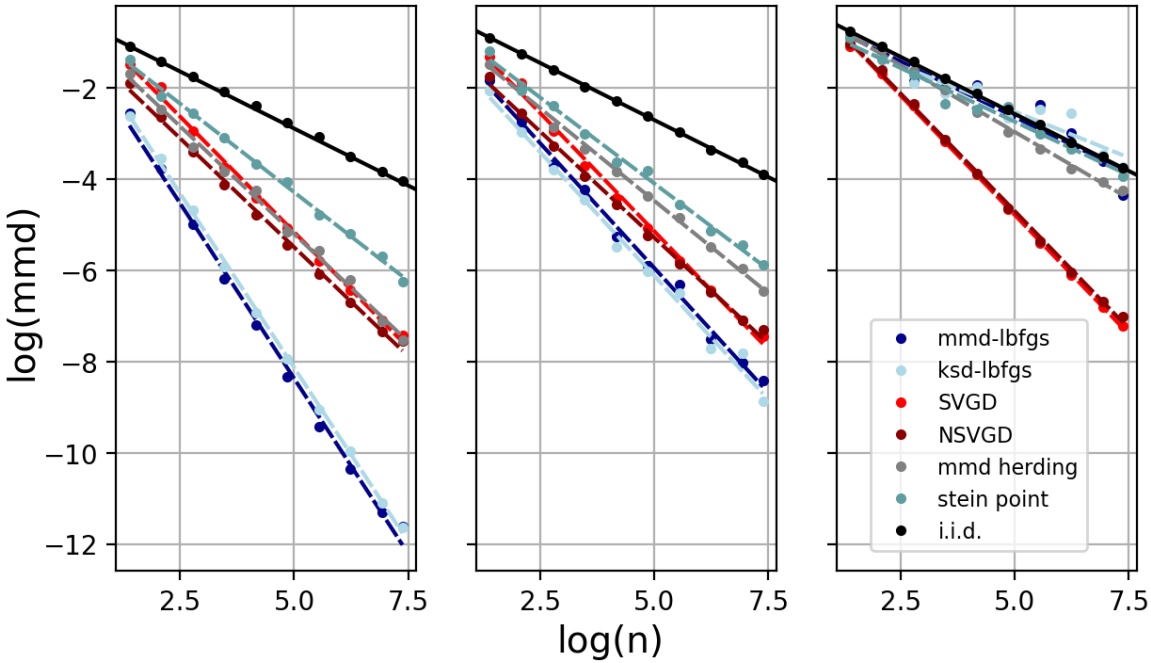


Figure: Importance of the choice of the bandwidth in the MMD evaluation metric when evaluating the final states, in 2D. From Left to Right: (evaluation) MMD bandwidth = 1, 0.7, 0.3.

- if we measure the discrepancy using a kernel with smaller bandwidth, MMD and KSD results deteriorate significantly and SVGD/NSVGD perform the best.
- likely reason : SVGD samples are more regular, while samples of MMD and KSD with Gaussian kernel have internal structures which can affect the discrepancy at lower bandwidths.

Conclusion

- We highlighted both theoretically and numerically that interacting particle systems derived from Wasserstein (and related) gradient flows, such as SVGD, MMD and KSD descent create "super samples"
- we proposed a normalized version of SVGD which accelerates the dynamics and observed that Laplace kernels produce more regular sample point distributions.

Open questions:

- proving quantization rates for SVGD
- improve our bounds

Thank you !

References I


Aistleitner, C. and Dick, J. (2015).
Functions of bounded variation, signed measures, and a general Koksma-Hlawka inequality.
Acta Arith., 167(2):143–171.


Ambrosio, L., Gigli, N., and Savaré, G. (2008).
Gradient flows: in metric spaces and in the space of probability measures.
Springer Science & Business Media.


Arbel, M., Korba, A., Salim, A., and Gretton, A. (2019).
Maximum mean discrepancy gradient flow.
In *Advances in Neural Information Processing Systems*, pages 6481–6491.

References II


Bach, F., Lacoste-Julien, S., and Obozinski, G. (2012).
On the equivalence between herding and conditional gradient algorithms.
In *ICML 2012 International Conference on Machine Learning*.


Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. J. (2018).
Stein points.
International Conference on Machine Learning (ICML).


Chen, Y., Welling, M., and Smola, A. (2012).
Super-samples from kernel herding.
arXiv preprint arXiv:1203.3472.

References III


Chwialkowski, K., Strathmann, H., and Gretton, A. (2016).
A kernel test of goodness of fit.
In *International conference on machine learning*.


Duncan, A., Nüsken, N., and Szpruch, L. (2019).
On the geometry of stein variational gradient descent.
arXiv preprint arXiv:1912.00894.


Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2006).
A kernel method for the two-sample-problem.
Advances in neural information processing systems, 19:513–520.


Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012).
A kernel two-sample test.
Journal of Machine Learning Research, 13:723–773.

References IV


Hyvärinen, A. and Dayan, P. (2005).
Estimation of non-normalized statistical models by score matching.
Journal of Machine Learning Research, 6(4).


Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. (2021).
Kernel Stein discrepancy descent.
International Conference of Machine Learning.


Łatuszyński, K., Miasojedow, B., and Niemiro, W. (2013).
Nonasymptotic bounds on the estimation error of mcmc algorithms.
Bernoulli, 19(5A):2033–2066.

References V


Liu, Q., Lee, J., and Jordan, M. (2016).
A kernelized stein discrepancy for goodness-of-fit tests.
In *International conference on machine learning*, pages 276–284.


Liu, Q. and Wang, D. (2016).
Stein variational gradient descent: A general purpose bayesian inference algorithm.
In *Advances in neural information processing systems*, pages 2378–2386.


Lu, Y. and Lu, J. (2020).
A universal approximation theorem of deep neural networks for expressing probability distributions.
Advances in Neural Information Processing Systems, 33.

References VI


Oates, C. J., Girolami, M., and Chopin, N. (2017).
Control functionals for monte carlo integration.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3):695–718.


Steinwart, I. and Christmann, A. (2008).
Support vector machines.
Springer Science & Business Media.


Tolstikhin, I., Sriperumbudur, B. K., and Muandet, K. (2017).
Minimax estimation of kernel mean embeddings.
The Journal of Machine Learning Research, 18(1):3002–3048.

L-BFGS

L-BFGS (Limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm) is a quasi-Newton method:

$$x_{l+1} = x_l - \gamma_n B_l^{-1} \nabla F(x_l) := x_l + \gamma_l d_l \tag{5}$$

where B_l^{-1} is a p.s.d. matrix approximating the inverse Hessian at x_l.

Step1. (requires ∇F) It computes a cheap version of d_l based on BFGS recursion:

$$B_{l+1}^{-1} = \left(I - \frac{\Delta x_l y_l^T}{y_l^T \Delta x_l} \right) B_l^{-1} \left(I - \frac{y_l \Delta x_l^T}{y_l^T \Delta x_l} \right) + \frac{\Delta x_l \Delta x_l^T}{y_l^T \Delta x_l}$$

$$\text{where} \quad \Delta x_l = x_{l+1} - x_l$$
$$y_l = \nabla F(x_{l+1}) - \nabla F(x_l)$$

Step2. (requires F and ∇F) A line-search is performed to find the best step-size in (5) :

$$F(x_l + \gamma_l d_l) \leq F(x_l) + c_1 \gamma_l \nabla F(x_l)^T d_l$$
$$\nabla F(x_l + \gamma_l d_l)^T d_l \geq c_2 \nabla F(x_l)^T d_l$$