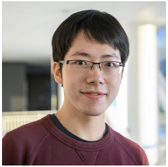
Accurate Quantization of Measures via Interacting Particle-based Optimization

Anna Korba
ENSAE, CREST, IP Paris

Ellis Theory workshop



Joint work with Lantian Xu, Dejan Slepčev (Carnegie Mellon University).

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:

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

Several approaches, among which :

- MCMC methods : generate a Markov chain whose law converges to π, $\mathrm{err}(x_1, \dots, x_n) = \mathcal{O}(n^{-1/2})$
  [Łatuszyński et al., 2013]
- **deterministic particle systems**, $\mathrm{err}(x_1, \dots, x_n)$?

Bayesian inference

Let $\mathcal{D} = (w_i, y_i)_{i=1}^{m}$ a dataset of labelled examples $(w_i, y_i) \overset{i.i.d.}{\sim} P_{\text{data}}$.
Assume an underlying model parametrized by w, e.g. :

$$y = g(w, x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathrm{I}).$$

Goal: learn the best distribution over parameter x to fit the data.

Bayesian inference

Let $\mathcal{D} = (w_i, y_i)_{i=1}^m$ a dataset of labelled examples $(w_i, y_i) \overset{i.i.d.}{\sim} P_{\text{data}}$.
Assume an underlying model parametrized by w, e.g. :

$$y = g(w, x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathrm{I}).$$

Goal: learn the best distribution over parameter x to fit the data.

1. Compute the Likelihood:

$$p(\mathcal{D}|x) = \prod_{i=1}^{m} p(y_i|x, w_i) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{m} \|y_i - g(w_i, x)\|^2\right).$$

Bayesian inference

Let $\mathcal{D} = (w_i, y_i)_{i=1}^m$ a dataset of labelled examples $(w_i, y_i) \overset{i.i.d.}{\sim} P_{\text{data}}$.
Assume an underlying model parametrized by w, e.g. :

$$y = g(w, x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathrm{I}).$$

Goal: learn the best distribution over parameter x to fit the data.

1. Compute the Likelihood:

$$p(\mathcal{D}|x) = \prod_{i=1}^m p(y_i|x, w_i) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^m \|y_i - g(w_i, x)\|^2\right).$$

2. Choose a prior distribution on the parameter:

$$x \sim p, \quad \text{e.g. } p(x) \propto \exp\left(-\frac{\|x\|^2}{2}\right).$$

Bayesian inference

Let $\mathcal{D} = (w_i, y_i)_{i=1}^m$ a dataset of labelled examples $(w_i, y_i) \overset{i.i.d.}{\sim} P_{\text{data}}$.
Assume an underlying model parametrized by w, e.g. :

$$y = g(w, x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathrm{I}).$$

Goal: learn the best distribution over parameter x to fit the data.

1. Compute the Likelihood:

$$p(\mathcal{D}|x) = \prod_{i=1}^m p(y_i|x, w_i) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^m \|y_i - g(w_i, x)\|^2\right).$$

2. Choose a prior distribution on the parameter:

$$x \sim p, \quad \text{e.g. } p(x) \propto \exp\left(-\frac{\|x\|^2}{2}\right).$$

3. Bayes' rule yields:

$$\pi(x) := p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p(x)}{Z} \quad Z = \int_{\mathbb{R}^d} p(\mathcal{D}|x)p(x)dx$$

i.e. $\pi(x) \propto \exp\left(-V(x)\right), \quad V(x) = \frac{1}{2}\sum_{i=1}^m \|y_i - g(w_i, x)\|^2 + \frac{\|x\|^2}{2}.$

π is needed both for

- prediction for a new input w:

$$y_{pred} = \int_{\mathbb{R}^d} g(w, x) d\pi(x)$$

"Bayesian model averaging"

- measure uncertainty on the prediction.

π is needed both for

- prediction for a new input w:

$$y_{pred} = \int_{\mathbb{R}^d} g(w, x) d\pi(x)$$

  "Bayesian model averaging"

- measure uncertainty on the prediction.

Given a discrete approximation $\mu_n = \frac{1}{n} \sum_{j=1}^{n} \delta_{x_j}$ of π:

$$y_{pred} \approx \frac{1}{n} \sum_{j=1}^{n} g(w, x_j).$$

Question: how can we approximate π?

Outline

Sampling as optimization over distributions

3 algorithms/particle systems at study:

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

These particle systems are designed to minimize a loss.

Sampling as optimization over distributions

3 algorithms/particle systems at study:

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

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:

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

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.

Euclidean gradient flow and continuity equation

Let $V : \mathbb{R}^d \to \mathbb{R}$ and consider minimizing V. The gradient flow of V can be written

$$\boxed{\frac{dx_t}{dt} = -\nabla V(x_t)}$$

and assume x_0 random with density μ_0.

Euclidean gradient flow and continuity equation

Let $V : \mathbb{R}^d \to \mathbb{R}$ and consider minimizing V. The gradient flow of V can be written

$$\boxed{\frac{dx_t}{dt} = -\nabla V(x_t)}$$

and assume x_0 random with density μ_0.

What are the dynamics of the density μ_t of x_t ? Let $\phi \in C_c\infty(\mathbb{R}^d)$.

$$\frac{d}{dt}\mathbb{E}(\phi(x_t)) = \int \phi(x)\frac{\partial \mu_t(x)}{\partial t}dx,$$

and applying the chain rule and using I.P.P.,

$$\frac{d}{dt}\mathbb{E}(\phi(x_t)) = -\int \langle \nabla\phi(x), \nabla V(x)\rangle \mu_t(x)dx = \int \phi(x)\boldsymbol{\nabla}\cdot(\mu_t(x)\nabla V(x))dx.$$

Therefore,

$$\boxed{\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla}\cdot(\mu_t \nabla V).}$$

Setting - The Wasserstein space

Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the space of probability measures on $\mathbb{R}^d$ with finite second moments, i.e.

$$\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d),\ \int \|x\|^2 d\mu(x) < \infty\}$$

Setting - The Wasserstein space

Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the space of probability measures on $\mathbb{R}^d$ with finite second moments, i.e.

$$\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d),\ \int \|x\|^2 d\mu(x) < \infty\}$$

Setting - The Wasserstein space

Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the space of probability measures on $\mathbb{R}^d$ with finite second moments, i.e.

$$\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d),\ \int \|x\|^2 d\mu(x) < \infty\}$$

$\mathcal{P}_2(\mathbb{R}^d)$ is endowed with the Wasserstein-2 distance from **Optimal transport** :

$$W_2^2(\nu, \mu) = \inf_{s \in \Gamma(\nu,\mu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, ds(x, y) \qquad \forall \nu, \mu \in \mathcal{P}_2(\mathbb{R}^d)$$

where $\Gamma(\nu, \mu)$ is the set of possible couplings between ν and μ (joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with first marginals ν and μ).

Setting - The Wasserstein space

Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the space of probability measures on $\mathbb{R}^d$ with finite second moments, i.e.

$$\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d),\ \int \|x\|^2 d\mu(x) < \infty\}$$

$\mathcal{P}_2(\mathbb{R}^d)$ is endowed with the Wasserstein-2 distance from **Optimal transport** :

$$W_2^2(\nu,\mu) = \inf_{s \in \Gamma(\nu,\mu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, ds(x,y) \qquad \forall \nu, \mu \in \mathcal{P}_2(\mathbb{R}^d)$$

where $\Gamma(\nu,\mu)$ is the set of possible couplings between ν and μ (joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with first marginals ν and μ).

Can also be written (Benamou-Brenier formula):

$$W_2^2(\nu,\mu) = \inf_{(\rho_t, v_t)_{t \in [0,1]}} \left\{ \int_0^1 \|v_t\|^2_{L^2(\rho_t)} dt(x) : \frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t v_t), \rho_0 = \nu, \rho_1 = \mu \right\}.$$

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

The first variation 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 first variation 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 $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $\nu - \mu \in \mathcal{P}(\mathbb{R}^d)$:

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

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

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

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

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

The first variation 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 $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $\nu - \mu \in \mathcal{P}(\mathbb{R}^d)$:

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

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

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

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

It can be implemented by the deterministic process:

$$\boxed{\frac{dx_t}{dt} = -\nabla_{W_2} \mathcal{F}(\mu_t)(x_t), \quad \text{where } x_t \sim \mu_t}$$

Particle system approximating the WGF

Euler time-discretization : in $\mathbb{R}^d$, move particles as:

$$x_{l+1} = Xx_l - \gamma \nabla_{W_2} \mathcal{F}(\mu_l)(x_l) \sim \mu_{l+1}, \quad x_0 \sim \mu_0.$$

But μ_l is unknown.

Space discretization/particle system : Introduce a particle system $x_0^1, \ldots, x_0^n \sim \mu_0$, and at each step:

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

$$\text{where } \hat{\mu}_l = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_l^i}$$

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
  $((k(x_i, x_j)_{i=1}^n)$ is a p.s.d. matrix for all $x_1, \dots, x_n \in \mathbb{R}^d)$
- examples:
  - the Gaussian kernel $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{h}\right)$
  - the Laplace kernel $k(x, y) = \exp\left(-\frac{\|x-y\|}{h}\right)$
- $\mathcal{H}_k$ its corresponding RKHS (Reproducing Kernel Hilbert Space):

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

- $\mathcal{H}_k$ 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]

Assume $\mu \mapsto \int k(x,.)d\mu(x)$ injective.

Maximum Mean Discrepancy defines a distance on $\mathcal{P}_2(\mathbb{R}^d)$:

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

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

Maximum Mean Discrepancy [Gretton et al., 2012]

Assume $\mu \mapsto \int k(x,.)d\mu(x)$ injective.

Maximum Mean Discrepancy defines a distance on $\mathcal{P}_2(\mathbb{R}^d)$:

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

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

The differential of $\mu \mapsto \frac{1}{2}\mathrm{MMD}^2(.,\pi)$ evaluated at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is:

$$\int k(x,.)d\mu(x) - \int k(x,.)d\pi(x) : \mathbb{R}^d \to \mathbb{R}.$$

Hence, for k regular enough, $\nabla_{W_2}\frac{1}{2}\mathrm{MMD}^2(\mu,\pi)$ is:

$$\int \nabla_2 k(x,.)d\mu(x) - \int \nabla_2 k(x,.)d\pi(x) : \mathbb{R}^d \to \mathbb{R}.$$

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:

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

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

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

[1] e.g. : $k(x,y) = \exp(-\|x-y\|^2/h)$, $\quad \pi(x) \propto e^{-\|x\|^2}$, $s(x) = -x$

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]

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

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

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

MMD and KSD Descent

Let $\mathcal{F}(\mu) = \mathrm{D}(\mu|\pi)$ where D is the MMD or KSD.

For discrete measures $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{X^i}$, let $F(X^1, \dots, X^n) := \mathcal{F}(\mu)$. Then, for $i = 1, \dots, n$,

$$x_{l+1}^i = x_l^i - \gamma \nabla_{W_2}\mathcal{F}(\hat{\mu}_l)(x_l^i), \quad \hat{\mu}_l = \frac{1}{n}\sum_{i=1}^n \delta_{x_l^i}$$

$$\Updownarrow$$

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

MMD and KSD Descent

Let $\mathcal{F}(\mu) = \mathrm{D}(\mu|\pi)$ where D is the MMD or KSD.

For discrete measures $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{X^i}$, let $F(X^1, \ldots, X^n) := \mathcal{F}(\mu)$.
Then, for $i = 1, \ldots, n$,

$$x_{l+1}^i = x_l^i - \gamma \nabla_{W_2} \mathcal{F}(\hat{\mu}_l)(x_l^i), \quad \hat{\mu}_l = \frac{1}{n}\sum_{i=1}^n \delta_{x_l^i}$$

$$\Updownarrow$$

$$\boxed{x_{l+1}^i = x_l^i - \gamma \nabla_{x^i} F(x_l^1, \ldots, x_l^n).}$$

- If D is the MMD, the gradient of F is:

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

- In contrast, if D is the KSD, it is:

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

Remarks

- The MMD/KSD/their W_2 gradient write as sums of integrals of μ and π

Remarks

- The MMD/KSD/their W_2 gradient write as sums of integrals of μ and π
- Hence they can be evaluated in closed form for discrete μ and $\pi \implies$ use L-BFGS to automatically select the best step-size

Remarks

- The MMD/KSD/their W_2 gradient write as sums of integrals of μ and π
- Hence they can be evaluated in closed form for discrete μ and $\pi \Longrightarrow$ use L-BFGS to automatically select the best step-size
- depending on the information on π, choose the KSD (unnormalized density) or MMD (samples)

Remarks

- The MMD/KSD/their W_2 gradient write as sums of integrals of μ and π
- Hence they can be evaluated in closed form for discrete μ and $\pi \implies$ use L-BFGS to automatically select the best step-size
- depending on the information on π, choose the KSD (unnormalized density) or MMD (samples)
- The MMD upper bounds the integral approximation error for functions in the RKHS, since by the reproducing property and Cauchy-Schwartz:

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

Similarly for the KSD with $\mathcal{H}_{k_\pi}$.

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 :

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

with respect to a "kernelized Wasserstein distance" depending on **a kernel** k [Liu, 2017, Duncan et al., 2019]:

$$W_k^2(\mu_0, \mu_1) = \inf_{(\mu_t, v_t)_{t\in[0,1]}} \left\{ \int_0^1 \|v_t\|^2_{\mathcal{H}_k^d} dt(x) : \frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot (\mu_t v_t) \right\}.$$

Stein Variational Gradient Descent [Liu and Wang, 2016]

In continuous time, SVGD flow is defined by the continuity equation

$$\frac{\partial \mu_t}{\partial t} + \boldsymbol{\nabla} \cdot (\mu_t v_{\mu_t}) = 0,\ v_{\mu_t} = S_{\mu_t,k} \nabla \log \left(\frac{\mu_t}{\pi}\right)$$

where

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

and one can write $v_{\mu_t} = k \star (\mu_t \nabla \log \pi) - \nabla k \star \mu_t$.

Stein Variational Gradient Descent [Liu and Wang, 2016]

In continuous time, SVGD flow is defined by the continuity equation

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

where

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

and one can write $v_{\mu_t} = k \star (\mu_t \nabla \log \pi) - \nabla k \star \mu_t$.

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 :

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

Remarks

- for discrete measures, the KL is not defined
- SVGD does not minimize a well-defined functional for discrete measures, it is only a discrete approximation of the KL flow
- cannot be used with L-BFGS (or not straightforwardly)
- how to measure the quantization, i.e. the quality of the particles obtained?

Outline

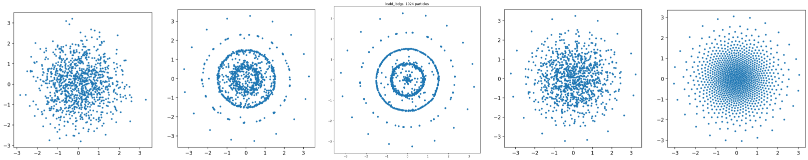
Problem and Motivation

Background on Interacting Particle Systems

MMD and KSD Quantization

Experiments

Motivation - Final states for a Gaussian target



(a) i.i.d. (b) MMD k_G (c) KSD k_G (d) SVGD k_G (e) SVGD k_L

Figure: Final states of the algorithms for 1000 particles, kernel bandwidth = 1. k_G is the Gaussian kernel and k_L is the Laplace kernel.

We run MMD/KSD descent with Gaussian kernel only, since

$$(1)\ \nabla_{x^i} \mathrm{MMD}^2(\mu_n, \pi) = \frac{1}{n} \sum_{j=1,\dots,n} \nabla_2 k(x^i, x^j) - \int \nabla_2 k(x^i, x) d\pi(x),$$

$$(2)\ \nabla_{x^i} \mathrm{KSD}^2(\mu_n, \pi) = \frac{1}{n} \sum_{j=1,\dots n} \nabla_2 k_\pi(x^i, x^j),$$

$$(3)\ \nabla_{x^i} \mathrm{SVGD} = \frac{1}{n} \sum_{j=1,\dots,n} \nabla \log \pi(x^i) k(x^i, x) + \nabla_{x^i} k(x^i, x)$$

(1) available in closed form for π and k Gaussian, (2) involves high order derivatives of the kernel, (3) can be run with any kernel including k_L

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

Assumption A1: Assume that the kernel is d-times continuously differentiable. Assume also that any mixed partial derivative of the kernel of order smaller than d has a RKHS norm bounded by a constant $C_{k,d} \geq 0$.

First result for the MMD

Theorem: Suppose A1 holds. Assume that (i) π is the Lebesgue measure or (ii) a probability 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: We use the well-known Koksma-Hlawka inequality [Aistleitner and Dick, 2015](Th1):

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

- $\mathcal{D}(X_n, \pi) = \sup_{I=\Pi_{i=1}^{n}[a_i,b_i]} |\pi(I) - \mu_n(I)|$ is the discrepancy of the point set X_n, can be bounded by $C_{\pi,d} g(n)$ [Aistleitner and Dick, 2015]
- The variation of a function $f : [0,1]^d \to \mathbb{R}$ with continuous mixed partial derivatives is defined as

$$V(f) = \sum_{\alpha \subseteq \{1,\ldots,d\}} \int_{[0,1]^{|\alpha|}} \left| \frac{\partial^{|\alpha|} f(x_\alpha, 1)}{\partial x_\alpha} \right| dx_\alpha.$$

Then, use the reproducing property on partial derivatives with Cauchy-Schwarz inequality, and **A1**:

$$\left| \frac{\partial^{|\alpha|} f(x_\alpha, 1)}{\partial x_\alpha} \right| \leq \left\| \frac{\partial^{|\alpha|} k((x_\alpha, 1), \cdot)}{\partial^{|\alpha|} x_\alpha} \right\|_{\mathcal{H}_k} \|f\|_{\mathcal{H}_k} \leq C_{k,d}.$$

Result for non compactly supported distributions π

Proposition 1: 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 1: 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 $\mathrm{MMD}(\pi, \mu_n) \leq \mathrm{MMD}(\pi, \mu) + \mathrm{MMD}(\mu, \mu_n)$, choosing μ compactly supported on $A_n = [-\log n, \log n]^d$.

As π is light-tailed, $\|\mu - \pi\|_{TV} \leq C_1/n$ distance, and we first get $\mathrm{MMD}(\pi, \mu) \leq C_k \|\mu - \pi\|_{TV} \leq C/n$.

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

Result for the KSD

Theorem: Assume that k is Gaussian and that $\pi \propto \exp(-U)$ with $U \in C^\infty(\mathbb{R}^d)$. Assume furthermore that $U(x) > c_1\|x\|$ for large enough x, and that there exists a real-valued polynomial V of degree $m \geq 0$, such that for any multi-index β, $\left|\frac{\partial^\beta U(x)}{\partial^{\beta_1}x_1 \ldots \partial^{\beta_j}x_j}\right| \leq V(x)$ for all $1 \leq |\beta| \leq d+1$. Then there exist points $x_1, \ldots, x_n$ such that

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

Satisfied for gaussian mixtures π.

Result for the KSD

Theorem: Assume that k is Gaussian and that $\pi \propto \exp(-U)$ with $U \in C^\infty(\mathbb{R}^d)$. Assume furthermore that $U(x) > c_1\|x\|$ for large enough x, and that there exists a real-valued polynomial V of degree $m \geq 0$, such that for any multi-index β, $\left|\frac{\partial^\beta U(x)}{\partial^{\beta_1}x_1 \dots \partial^{\beta_j}x_j}\right| \leq V(x)$ for all $1 \leq |\beta| \leq d+1$. Then there exist points $x_1, ..., x_n$ such that

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

Satisfied for gaussian mixtures π.

Proof: The proof relies on bounding the first and last term of the

$$\begin{aligned}\mathrm{KSD}(\mu_n, \pi) = 2 \iint \nabla \log(\pi)(x)^T \nabla_y k(x, y) d\mu(x) d\mu(y) \\ + \underbrace{\iint \nabla \log(\pi)(x)^T \nabla \log(\pi)(y) k(x, y) d\mu(x) d\mu(y)}_{(1)} + \underbrace{\iint \boldsymbol{\nabla} \cdot_x \nabla_y k(x, y) d\mu(x) d\mu(y)}_{(2)},\end{aligned}$$

$\mu = \mu_n - \pi$, as the cross terms can be upper bounded by the former ones by CS and reproducing property.

(1) $\mathrm{MMD}(\mu_n, \pi)$, with $k_1(x, y) = s(x)^T s(y) k(x, y)$, bounded by controlling $\|\nabla \log \pi\|_{H^d}$

(2) $\mathrm{MMD}(\mu_n, \pi)$, with $k_2(x, y) = \boldsymbol{\nabla} \cdot_x \nabla_y k(x, y)$, bounded by Prop 1 for bounded kernels.

Outline

Algorithms

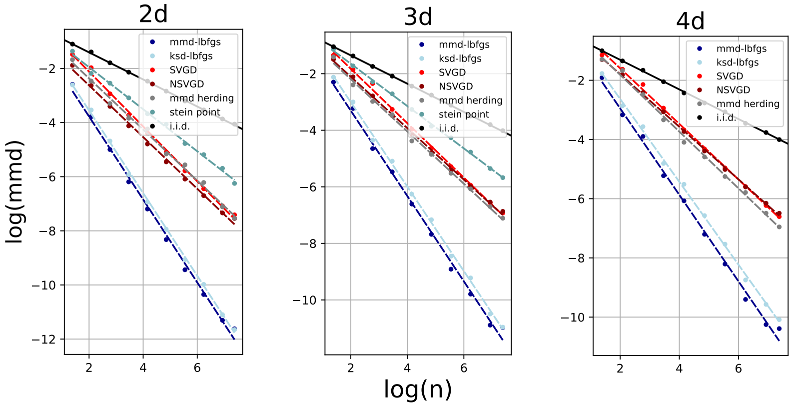
we investigate numerically the quantization properties of :

- SVGD
- MMD descent
- KSD Descent
- Kernel Herding (KH) : greedy minimization of the MMD
- Stein points (SP) : greedy minimization of the KSD

Hyperparameters:

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

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



Averaged over 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; the same bandwith is used for evaluation.

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

Some remarks:

- The slopes remain much steeper than the Monte Carlo rate (-0.5), even when the dimension increases
- The slopes are better than our theoretical upper bounds

Robustness to evaluation discrepancy

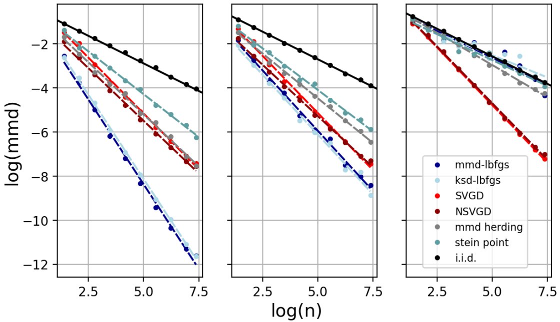


Figure: Importance of the choice of the bandwidth in the MMD evaluation metric when evaluating the final states, in 2D. From Left to Right: (evaluation) MMD bandwidth = 1, 0.7, 0.3.

- 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 : Samples of MMD and KSD with Gaussian kernel have internal structures which can affect the discrepancy at lower bandwidths.

For $\nu, \mu \in \mathcal{P}_p(\mathbb{R}^d)$, the Sliced p-Wasserstein (SW) distance is defined as:

$$d_{sw,p}(\nu, \mu) = \int_{\mathbb{S}^{d-1}} W_p(P_{\theta\#}\nu, P_{\theta\#}\mu) d\theta, \quad P_\theta : x \mapsto x \cdot \theta.$$

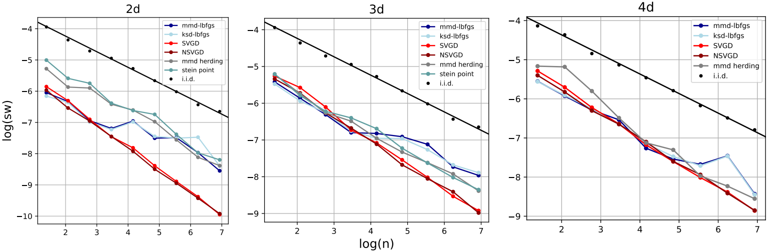


Figure: Quantization rates measured in SW distance of the algorithms $\pi = \mathcal{N}(0, {}^1\!/_d I_d)$. We use $p = 1$ and 50 random directions drawn uniformly on $\mathbb{S}^{d-1}$ to discretize the integration.

The rates for SVGD are approximately $n^{-0.72}, n^{-0.65}, n^{-0.63}$ for $d = 2, 3$, and 4. We note that these are quite close to the rate we theoretically predict for the distance between the measure on a grid in $[0, 1]^d$, and the Lebesgue measure: $d_{sw,1} \sim n^{-\frac{1}{2} - \frac{1}{2d}}$, which is $n^{-0.75}, n^{-0.67}, n^{-0.625}$ for $d = 2, 3$, and 4.

Conclusion

- MMD/ KSD descent, SVGD can create "super samples" that approximate π at fast rates

Open questions/future work:

- improve our quantization bounds for MMD/KSD (dependence in dimension, Laplace kernel?)
- obtain quantization bounds for SVGD

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*.

Carrillo, J. A., Craig, K., and Patacchini, F. S. (2019).
A blob method for diffusion.
Calculus of Variations and Partial Differential Equations, 58(2):1–53.

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

References III

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

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.

References IV

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. J. (2012).
A kernel two-sample test.
JMLR, 13.

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. (2017).
Stein variational gradient descent as gradient flow.
In *Advances in neural information processing systems*, pages 3115–3123.

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.

References VI

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.

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.

References VII

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.

Alternative assumption for the MMD bound:

A2. 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 : $\exists C_{k,d} \geq 0$ such that $(1 + |\xi|^2)^d \leq C_{k,d} |\hat{\eta}(\xi)|^{-1}$ for any $\xi \in \mathbb{R}^d$.

A2 includes kernels which are not smooth, such as Matern kernels that can be defined through their Fourier transform $\hat{\eta}(\xi) \propto \frac{1}{(1+\|\xi\|^2)^j}$, $j \geq d$ whose RKHS correspond to Sobolev spaces of order j, and which are not smooth at $z = 0$.

Laplace kernel $k(x, y) = \exp(-\|x - y\|)$ corresponds to $j =: fracd + 12$ and does not satisfy **A2.**

A1 is satisfied by the Gaussian kernel with $C_{k,d} = (2d)!$.

Proof. By the reproducing property, we have

$$\left\| \frac{\partial^{|\alpha|} k((x_\alpha, 1), \cdot)}{\partial^{|\alpha|} x_\alpha} \right\|_{\mathcal{H}_k} = \left(\frac{\partial^{|\alpha|,|\alpha|} k((x_\alpha, 1), (x_\alpha, 1))}{\partial^{|\alpha|} x_\alpha \partial^{|\alpha|} y_\alpha} \right)^{\frac{1}{2}}.$$

Consider the Gaussian kernel, i.e. for $x, y \in \mathbb{R}^d$, $k(x, y) = e^{-\|x-y\|^2/h}$. Hence, for any $x, y \in \mathbb{R}^d$, the $|\alpha|$-th partial derivative of the kernel in both variables is equal to

$$\frac{\partial^{|\alpha|,|\alpha|} k(x, y)}{\partial^{|\alpha|} x_\alpha \partial^{|\alpha|} y_\alpha} = (-1)^{|\alpha|} \frac{\partial^{2|\alpha|} e^{-t^2}}{\partial^{2|\alpha|} t} = (-1)^{|\alpha|} e^{-t^2} h_{2|\alpha|}(t)$$

where h_u, $u \geq 0$ denotes the u-th Hermite polynomial. In particular for $x = y$, i.e. $t = 0$, evaluations of Hermite polynomials at zero correspond to the well-known Hermite numbers $(-1)^{|\alpha|} 2^{|\alpha|} (2|\alpha| - 1)!!$ with $(2|\alpha| - 1)!! = 1 \times 3 \times \cdots \times (2|\alpha| - 1)$. We conclude using $|\alpha| \leq d$.

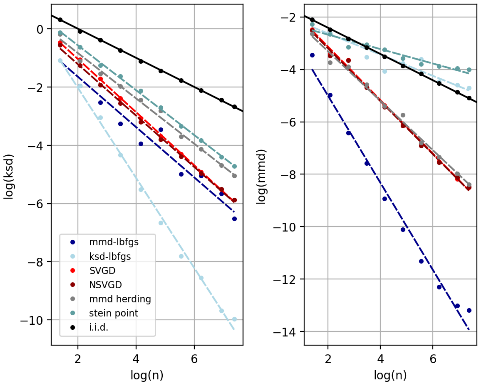


Figure: Quantization rates of the algorithms at study when the target distribution is a 2D-Gaussian mixture distribution with variance 0.3, centred at [1,0] and [-1,0]. We evaluate them using MMD and KSD with bandwidth 1. We run algorithms under the same setting as the 2-4D experiments on Figure 30.

L-BFGS

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

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

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:

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

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

Step2. (requires F and ∇F) A line-search is performed to find the best step-size in (1) :

$$\begin{aligned} F(x_l + \gamma_l d_l) &\leq F(x_l) + c_1 \gamma_l \nabla F(x_l)^T d_l \\ \nabla F(x_l + \gamma_l d_l)^T d_l &\geq c_2 \nabla F(x_l)^T d_l \end{aligned}$$

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_n, adding one new particle at each iteration to minimize MMD/KSD.

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

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

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

[Bach et al., 2012] 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.

Forward method for the KL

Problem: $\nabla_{W_2} \mathrm{KL}(\mu_n|\pi) = \nabla \log\left(\frac{\mu_n}{\pi}\right)$ where μ_n is unknown.

While $\nabla \log \pi$ is known, $\nabla \log \mu_n$ has to be estimated from N particles $X_n^1, \dots, X_n^N$, e.g. with[2] :

1. **Kernel Density Estimation (KDE)**:

$$\mu_n(.) \approx \frac{1}{N} \sum_{i=1}^{N} k(X_n^i - .)$$

Then,

$$-\nabla_{W_2} \mathrm{KL}(\mu_n|\pi)(.) \approx -\left(\nabla V(.) + \frac{\sum_{i=1}^N \nabla k(. - X_n^i)}{\sum_{i=1}^N k(. - X_n^i)}\right)$$

Remark: it is not the W_2 gradient of some functional (see the next slide)

[2] assume a symmetric, translation invariant kernel

2. **Blob Method** [Carrillo et al., 2019]:

Instead of

$$\mathcal{U}(\mu) = \int \log(\mu(x)) d\mu(x),$$

consider

$$\mathcal{U}_k(\mu) = \int \log(k \star \mu(x)) d\mu(x), \text{ where } k \star \mu(x) = \int k(x-y) d\mu(y).$$

Then,

$$\frac{\partial \mathcal{U}_k(\mu)}{\partial \mu}(.) = k \star \left(\frac{\mu}{k \star \mu} \right) + \log(k \star \mu)$$

$$\implies \nabla_{W_2} \mathcal{U}_k(\mu) == \nabla k \star \left(\frac{\mu}{k \star \mu} \right) + \underbrace{\nabla \log(k \star \mu)}_{\frac{\nabla k \star \mu}{k \star \mu}}$$

$$\implies \nabla_{W_2} \mathrm{KL}(\mu_n | \pi)(.) \approx -(\nabla V(.) + \sum_{i=1}^{N} \frac{\nabla k(. - X_n^i)}{\sum_{m=1}^{N} k(X_n^i - X_n^m)} + \frac{\sum_{i=1}^{N} \nabla k(. - X_n^i)}{\sum_{i=1}^{N} k(. - X_n^i)})$$