KSD and MMD gradient descent

Anna Korba
ENSAE/CREST

OT Seminar - Orsay

Joint work with **Adil Salim** (Simons Institute), **Giulia Luise** (UCL), **Michael Arbel** (INRIA Grenoble), **Arthur Gretton** (UCL), **Pierre-Cyril Aubin-Frankowski** (INRIA Paris), **Szymon Majewski** (Polytechnique), **Pierre Ablin** (CNRS), **Lantian Xu** (CMU), **Dejan Slepčev** (CMU).

Outline

Problem and Motivation

Background on MMD/KSD Descent

Theoretical study

MMD and KSD Quantization

Experiments

Quantization problem

Problem : approximate a target distribution $\pi \in \mathcal{P}(\mathbb{R}^d)$ by a finite set of n points $x_1, \ldots, x_n$, 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:

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

Several approaches, among which :

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

Example 1 : Bayesian statistics

- Let $\mathcal{D} = (x_i, y_i)_{i=1,\dots,m}$ a labelled dataset.
- Assume an underlying model parametrized by $z \in \mathbb{R}^d$, e.g.
  $y \sim f(x,z) + \epsilon \quad (p(y|x,z)$ gaussian)

  $\implies$ Compute the likelihood: $p(\mathcal{D}|z) = \prod_{i=1}^{m} p(y_i|x_i, z)$.
- Assume a prior distribution on the parameter $z \sim p$.

Example 1 : Bayesian statistics

- Let $\mathcal{D} = (x_i, y_i)_{i=1,\dots,m}$ a labelled dataset.
- Assume an underlying model parametrized by $z \in \mathbb{R}^d$, e.g.
  $y \sim f(x,z) + \epsilon \quad (p(y|x,z)$ gaussian)
  $\Longrightarrow$ Compute the likelihood: $p(\mathcal{D}|z) = \prod_{i=1}^m p(y_i|x_i, z)$.
- Assume a prior distribution on the parameter $z \sim p$.

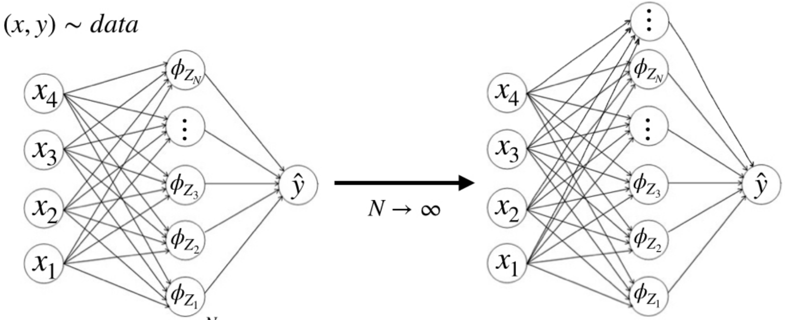
$$\text{Bayes' rule : } \pi(z) := p(z|\mathcal{D}) = \frac{p(\mathcal{D}|z)p(z)}{C}\ , \ C = \int_{\mathbb{R}^d} p(\mathcal{D}|z)p(z)dz.$$

π **is known up to a constant** since C is intractable.

How to sample from π then? e.g. to compute:

$$p(y|x,\mathcal{D}) = \int_{\mathbb{R}^d} p(y|x,z)d\pi(z)$$

Example 2 : Regression with infinite width NN



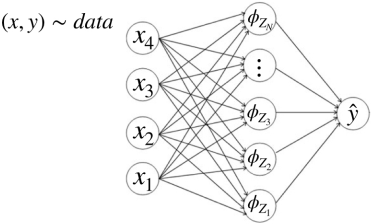
$$\min_{Z_1,\dots,Z_N} \mathbb{E}_{data}[\|y - \frac{1}{N}\sum_{i=1}^{N} \phi_{Z_i}(x)\|^2] \xrightarrow[N \to \infty]{} \min_{\nu \in \mathscr{P}} \mathbb{E}_{data}[\|y - \mathbb{E}_{Z\sim\nu}[\phi_Z(x)]\|^2]$$

[Chizat and Bach, 2018, Rotskoff and Vanden-Eijnden, 2018, Mei et al., 2018]

Illustration : Student-Teacher network

The output of the Teacher network is deterministic and given by $y = \int \phi_Z(x) d\pi(Z)$ where $\pi = \frac{1}{M} \sum_{m=1}^{M} \delta_{U^m}$.

Student network by $\mu_0 = \frac{1}{N} \sum_{j=1}^{N} \delta_{Z_0^j}$ tries to learn the mapping $x \mapsto \int \phi_Z(x) d\pi(Z)$.



$$\min_{Z_1,\ldots,Z_N} \mathbb{E}_{data}[\|\frac{1}{M}\sum_{m}^{M}\phi_{U^m}(x) - \frac{1}{N}\sum_{n=1}^{N}\phi_{Z^n}(x)\|^2]$$

Can be written as minimizing an MMD(μ, π).

Outline

Sampling as optimization over distributions

2 algorithms/particle systems at study:

- 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

2 algorithms/particle systems at study:

- 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:

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

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 first variationl 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 variationl 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 distributionnally:

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

where $\nabla_{W_2} \mathcal{F}(\mu) := \nabla \frac{\partial \mathcal{F}(\mu)}{\partial \mu} \in L^2(\mu)$ 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 $X^1, \ldots, X^n$ where μ_l is replaced by $\hat{\mu}_l = \frac{1}{n} \sum_{i=1}^n \delta_{X_l^i}$:

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

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 $x_1, \ldots, x_n \in \mathbb{R}^d)$
- examples:
  - the Gaussian kernel $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{h}\right)$
  - the Laplace kernel $k(x, y) = \exp\left(-\frac{\|x-y\|}{h}\right)$
  - the inverse multiquadratic kernel
    $k(x, y) = (c + \|x - y\|)^{-\beta}$ 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]

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. : $k(x,y) = \exp(-\|x-y\|^2/h)$

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:

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

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:

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

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

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

MMD and KSD Descent

Recall that we want to study particle systems

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

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

MMD and KSD Descent

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 $\mathcal{F}(\mu) = \mathrm{D}(\mu|\pi)$.

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

Recall that we want to study particle systems

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

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

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

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

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

- In contrast, if D is the KSD,

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

MMD and KSD Descent

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 $\mathcal{F}(\mu) = \mathrm{D}(\mu|\pi)$.

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 $i = 1, \dots, n$:

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

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 $\pi \implies$ use L-BFGS to automatically select the best step-size
- depending on the information on π, choose the KSD (unnormalized density) or MMD (samples)

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 $\mathcal{H}_{k_\pi}$.

Outline

Functional inequalities

How fast $\mathcal{F}(\mu_t)$ decreases along its WGF ?

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

$$\begin{aligned}
\frac{d\mathcal{F}(\mu_t)}{dt} &= \left\langle V_t, \nabla_{W_2} \mathcal{F}(\mu_t) \right\rangle_{L^2(\mu_t)} \\
&= -\left\| \nabla_{W_2} \mathcal{F}(\mu_t) \right\|^2_{L^2(\mu_t)} \\
&= -\left\| \mathbb{E}_{x\sim\mu_t}[\nabla_2 k(x,y)] - \mathbb{E}_{x\sim\pi}[\nabla_2 k(x,y)] \right\|^2_{L^2(\mu_t)} \\
&= -\underbrace{\|\nabla f_{\mu_t,\pi}\|^2_{L_2(\mu_t)}}_{\|f_{\mu_t,\pi}\|_{\dot{H}^{-1}(\mu_t)}}
\end{aligned}$$

where $f_{\mu_t,\pi} = \mathbb{E}_{x\sim\mu_t}[k(x,.)] - \mathbb{E}_{x\sim\pi}[k(x,.)]$.

Functional inequalities

How fast $\mathcal{F}(\mu_t)$ decreases along its WGF ?

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

$$\begin{aligned}
\frac{d\mathcal{F}(\mu_t)}{dt} &= \left\langle V_t, \nabla_{W_2}\mathcal{F}(\mu_t)\right\rangle_{L^2(\mu_t)} \\
&= -\left\|\nabla_{W_2}\mathcal{F}(\mu_t)\right\|^2_{L^2(\mu_t)} \\
&= -\|\mathbb{E}_{x\sim\mu_t}[\nabla_2 k(x,y)] - \mathbb{E}_{x\sim\pi}[\nabla_2 k(x,y)]\|^2_{L^2(\mu_t)} \\
&= -\underbrace{\|\nabla f_{\mu_t,\pi}\|^2_{L_2(\mu_t)}}_{\|f_{\mu_t,\pi}\|_{\dot{H}^{-1}(\mu_t)}}
\end{aligned}$$

where $f_{\mu_t,\pi} = \mathbb{E}_{x\sim\mu_t}[k(x,.)] - \mathbb{E}_{x\sim\pi}[k(x,.)]$.

It can be shown that:

$$\|f_{\mu_t,\pi}\|^2_{\mathcal{H}_k} \leq \|f_{\mu_t,\pi}\|_{\dot{H}(\mu_t)} \quad \underbrace{\|\mu_t - \pi\|_{\dot{H}^{-1}(\mu_t)}}_{\sup_{\|g\|^2_{\dot{H}(\mu_t)}\leq 1} |\int g d\mu_t - \int g d\pi|}$$

Hence, if $\|\mu_t - \pi\|_{\dot{H}^{-1}(\mu_t)} \leq C$ for all $t \geq 0$, we have

$$\frac{d\mathcal{F}(\nu_t)}{dt} \leq -C\mathcal{F}(\nu_t)^2, \text{ hence}$$

$$\mathcal{F}(\mu_t) \leq \frac{1}{\mathcal{F}(\mu_0) + 4C^{-1}t}$$

where $\mathcal{F}(\mu_0) = \frac{1}{2}\,\mathrm{MMD}^2(\mu_t, \pi)$.

Problems:

- depends on the whole sequence $(\mu_t)_{t\geq 0}$ (not only π)
- hard to verify in practice
- we observed convergence issues in practice (more for the MMD than the KSD)

Geodesic convexity

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and ρ_t a W_2 geodesic between μ and ν.

A functional $\mathcal{F}$ is (λ)-geodesically convex if it is convex along W_2 geodesics, i.e. if for any $t \in [0, 1]$:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\mu) + t\mathcal{F}(\nu) - t(1-t)\frac{\lambda}{2}W_2^2(\mu, \nu)^2$$

where $\rho_t = ((1-t)I + tT_\mu^\nu)_\# \mu$.

If $\mathcal{G}$ is λ-convex with $\lambda > 0$:

$$W_2(\mu_t, \pi) \leq e^{-\lambda t} W_2(\mu_0, \pi)$$

Geodesic convexity

Let $\psi \in C_c^\infty(\mathbb{R}^d)$ and :

$$\mathrm{Hess}_\mu \mathcal{F}(\psi,\psi) = \langle H_{\mathcal{F},\mu}\nabla\psi, \nabla\psi\rangle_{L^2(\mu_t)} = \frac{d^2}{dt^2}\Big|_{t=0}\mathcal{F}(\rho_t)$$

if $\rho_t = (I + t\nabla\psi)_\#\mu$ is a geodesic starting at μ.

For $\psi \in C_c^\infty(\mathbb{R}^d)$, we have

$$\mathrm{Hess}_\mu \mathcal{F}(\psi,\psi) = \underbrace{\mathbb{E}_{x,y\sim\mu}\left[\nabla\psi(x)^T\nabla_1\nabla_2 k(x,y)\nabla\psi(y)\right]}_{\left\|\mathbb{E}_{x\sim\mu}[\nabla\psi(x)^T\nabla k(x,.)]\right\|^2_{\mathcal{H}_k}}$$

$$+\mathbb{E}_{x\sim\mu}\left[\nabla\psi(x)^T\left(\mathbb{E}_{x\sim\mu}[H_1 k(x,y)] - \mathbb{E}_{x\sim\pi}[H_1 k(x,y)]\right)\nabla\psi(x)\right].$$

- the first term is always positive but not the second one
- i.e. we don't have generally $\mathrm{Hess}_\mu \mathcal{F}(\psi,\psi) \geq 0$
- i.e. neither the MMD nor the KSD are convex w.r.t. W_2 geodesics

Third strategy : curvature near equilibrium?

What happens near equilibrium π? the second term vanishes due to the Stein property of k_π and :

$$\mathrm{Hess}_\pi \mathcal{F}(\psi, \psi) = \| S_{\pi, k_\pi} \mathcal{L}_\pi \psi \|^2_{\mathcal{H}_{k_\pi}} \geq 0$$

where

$$\mathcal{L}_\pi : f \mapsto -\Delta f - \langle \nabla \log \pi, \nabla f \rangle_{\mathbb{R}^d}$$

$$S_{\mu, k_\pi} : f \mapsto \int k_\pi(x, .) f(x) d\mu(x) \in \mathcal{H}_{k_\pi}$$

Third strategy : curvature near equilibrium?

What happens near equilibrium π? the second term vanishes due to the Stein property of k_π and :

$$\mathrm{Hess}_\pi \mathcal{F}(\psi, \psi) = \|S_{\pi,k_\pi} \mathcal{L}_\pi \psi\|^2_{\mathcal{H}_{k_\pi}} \geq 0$$

where

$$\mathcal{L}_\pi : f \mapsto -\Delta f - \langle \nabla \log \pi, \nabla f \rangle_{\mathbb{R}^d}$$

$$S_{\mu,k_\pi} : f \mapsto \int k_\pi(x,.) f(x) d\mu(x) \in \mathcal{H}_{k_\pi}$$

Question: can we bound from below the Hessian at π by a quadratic form on the tangent space of $\mathcal{P}_2(\mathbb{R}^d)$ at π ($\subset L^2(\pi)$)?

$$\boxed{\mathrm{Hess}_\pi \mathcal{F}(\psi, \psi) \geq \lambda \|\nabla \psi\|^2_{L^2(\pi)} \; ?}$$

That would imply exponential decay of $\mathcal{F}$ near π.

Curvature near equilibrium - negative result

Theorem : Let $\pi \propto e^{-V}$. Assume that $V \in C^2(\mathbb{R}^d)$, ∇V is Lipschitz and $\mathcal{L}_\pi$ has discrete spectrum. Then exponential decay near equilibium does not hold.

Curvature near equilibrium - negative result

Theorem : Let $\pi \propto e^{-V}$. Assume that $V \in C^2(\mathbb{R}^d)$, ∇V is Lipschitz and $\mathcal{L}_\pi$ has discrete spectrum. Then exponential decay near equilibium does not hold.

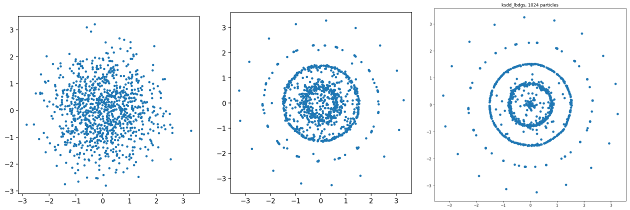
Proof: The previous inequality

$$\boxed{\|S_{\pi,k_\pi}\mathcal{L}_\pi\psi\|^2_{\mathcal{H}_{k_\pi}} \geq \lambda\|\nabla\psi\|^2_{L^2(\pi)}}$$

- can be seen as a kernelized version of the Poincaré inequality for π :
$$\|\mathcal{L}_\pi\psi\|^2_{L_2(\pi)} \geq \lambda_\pi\|\nabla\psi\|^2_{L_2(\pi)}.$$
- can be written:
$$\langle\psi, P_{\pi,k_\pi}\psi\rangle_{L_2(\pi)} \geq \lambda\langle\psi, \mathcal{L}_\pi^{-1}\psi\rangle_{L_2(\pi)},$$
where $P_{\pi,k_\pi} : L^2(\pi) \to L^2(\pi), f \mapsto \int k_\pi(x,.)f(x)d\pi(x).$
- compare decay of eigenvalues

Outline

Motivation - Final states for a Gaussian target



(a) i.i.d. (b) MMD Gaussian kernel (c) KSD Gaussian kernel

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.

MMD gradient is available in closed form for $\pi = \mathcal{N}(0_d, \theta I_d)$

$$\dot{x}_i = -\frac{1}{nh^2(\sqrt{2\pi h^2})^d}\sum_{j=1}^{n} e^{-\frac{|x_j - x_i|^2}{2h^2}}(x_j - x_i) - \frac{1}{(h^2+\theta^2)(\sqrt{2\pi(h^2+\theta^2)})^d} e^{-\frac{|x_i|^2}{2(h^2+\theta^2)}} x_i.$$

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 $\implies \exists C_{1,d}$ 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 $\mathrm{MMD}(\mu_n, \pi) \leq 4^d \mathcal{D}(X_n, \pi)$.

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 :

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

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

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

Satisfied for gaussian mixtures π.

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 $x_1, \dots, x_n$ 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

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

$\mu = \mu_n - \pi$, as the cross terms can be upper bounded by the former ones by a simple computation.

(1) $\mathrm{MMD}(\mu_n, \pi)$, with $k_1(x,y) = s(x)^T s(y) k(x,y)$, bounded by Prop 1

(2) $\mathrm{MMD}(\mu_n, \pi)$, with $k_2(x,y) = \boldsymbol{\nabla} \cdot_x \nabla_y k(x,y)$, bounded by controlling $\|\nabla \log \pi\|_{H^d}$

Outline

Algorithms

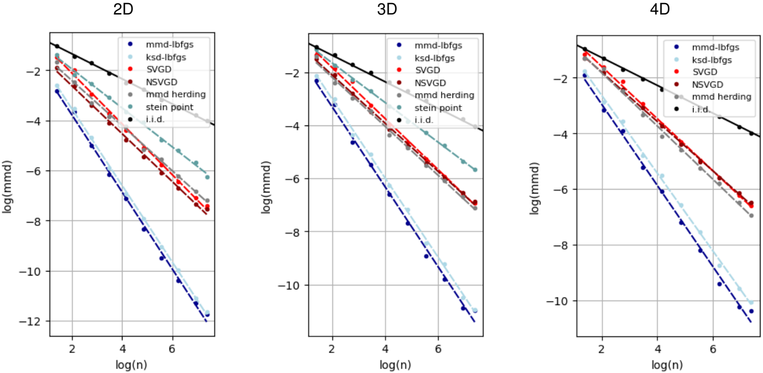
we investigate numerically the quantization properties of :

- 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; Stein points use gridsize = 200 points in 2d, 50 in 3d; in 4d grid search was too slow.

d	Eval.	MMD-lbfgs	KSD-lbfgs	KH	SP
2	**KSD**	-1.48	-1.46	-0.84	-0.77
	MMD	-1.60	-1.54	-0.93	-0.77
3	**KSD**	-1.38	-1.44	-0.84	-0.78
	MMD	-1.51	-1.49	-0.92	-0.75
4	**KSD**	-1.35	-1.39	-0.89	–
	MMD	-1.46	-1.40	-0.95	–
8	**KSD**	-1.14	-1.16	–	–
	MMD	-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
- Their slopes are better than our theoretical upper bounds

Robustness to evaluation discrepancy

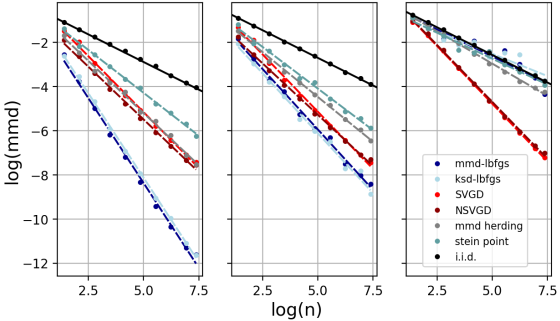


Figure: Importance of the choice of the bandwidth in the MMD evaluation metric when evaluating the final states, in 2D. From Left to Right: (evaluation) MMD bandwidth = 1, 0.7, 0.3.

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

Conclusion

- MMD and KSD descent convergence are not well grounded theoretically
- Still, they can create "super samples"

Open questions/future work:

- explain the convergence of KSD gradient flow
- improve our quantization 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

Chizat, L. and Bach, F. (2018).
On the global convergence of gradient descent for over-parameterized models using optimal transport.
Advances in neural information processing systems, 31.

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

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.

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.

References V

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

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.

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

Mei, S., Montanari, A., and Nguyen, P.-M. (2018).
A mean field view of the landscape of two-layer neural networks.
Proceedings of the National Academy of Sciences, 115(33):E7665–E7671.

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.

Rotskoff, G. M. and Vanden-Eijnden, E. (2018).
Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error.
stat, 1050:22.

References VII

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.

The well-specified case [Arbel et al., 2019]

We have $(x, y) \sim data$.

Assume $\exists \pi \in \mathcal{P}$ **,** $\mathbb{E}[y|X = x] = \mathbb{E}_{Z\sim\pi}[\phi_Z(x)]$.

Then :

$$\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathbb{E}[\|y - \mathbb{E}_{Z\sim\mu}[\phi_Z(x)]\|^2]$$

$$\Updownarrow$$

$$\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathbb{E}[\|\mathbb{E}_{Z\sim\pi}[\phi_Z(x)] - \mathbb{E}_{Z\sim\mu}[\phi_Z(x)]\|^2]$$

$$\Updownarrow$$

$$\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathbb{E}_{\substack{Z\sim\pi\\ Z'\sim\pi}}[k(Z,Z')] + \mathbb{E}_{\substack{Z\sim\mu\\ Z'\sim\mu}}[k(Z,Z')] - 2\mathbb{E}_{\substack{Z\sim\pi\\ Z'\sim\mu}}[k(Z,Z')]$$

with $k(Z, Z') = \mathbb{E}_{x\sim data}[\phi_Z(x)^T\phi_{Z'}(x)]$

$$\Updownarrow$$

$$\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \frac{1}{2}\mathrm{MMD}^2(\mu, \pi)$$

L-BFGS

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

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

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]:

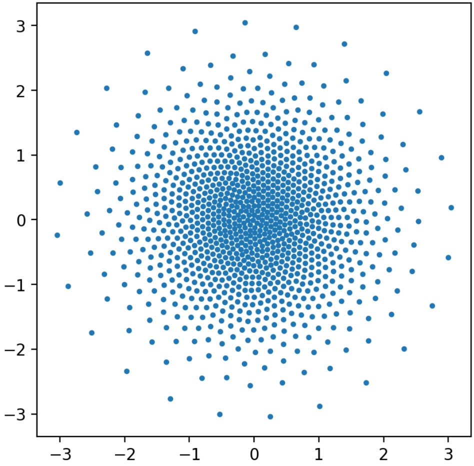
$$x^{n+1} = \underset{x \in \mathbb{R}^d}{\operatorname{argmax}} \langle w_n, k(x,.) \rangle_{\mathcal{H}_k}$$
$$w_{n+1} = w_n + m_\pi - k(x_{n+1},.)$$

[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.

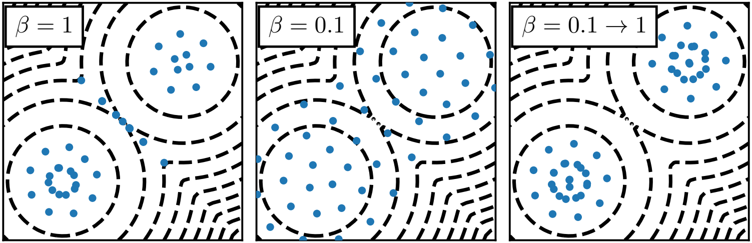
Stein Points for the KSD [Chen et al., 2018] greedily minimizes the KSD similarly. The authors establish a $\mathcal{O}((\log(n)/n)^{\frac{1}{2}})$ rate, which seem slower than their empirical observations.

SVGD with laplace kernel



Isolated Gaussian mixture - annealing

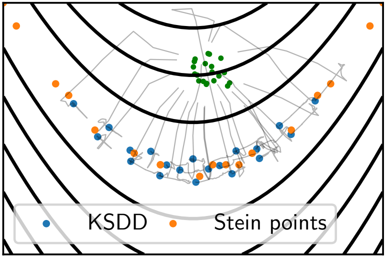
Add an inverse temperature variable β : $\pi^{\beta}(x) \propto \exp(-\beta V(x))$, with $0 < \beta \leq 1$ (i.e. multiply the score by β.)



This is a hard problem, even for Langevin diffusions, where tempering strategies also have been proposed.

Beyond Log-concavity: Provable Guarantees for Sampling Multi-modal Distributions using Simulated Tempering Langevin Monte Carlo. Rong Ge, Holden Lee, Andrej Risteski. 2017.

So.. when does it work?



Comparison of KSD Descent and Stein points on a “banana” distribution. Green points are the initial points for KSD Descent. Both methods work successfully here, **even though it is not a log-concave distribution.**

We posit that KSD Descent succeeds because **there is no saddle point in the potential.**

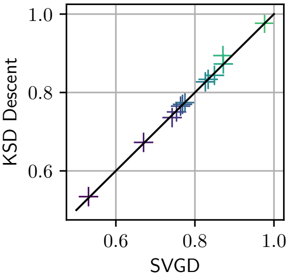
1 - Bayesian Logistic regression

Datapoints $d_1, \ldots, d_q \in \mathbb{R}^p$, and labels $y_1, \ldots, y_q \in \{\pm 1\}$.

Labels y_i are modelled as $p(y_i = 1 | d_i, w) = (1 + \exp(-w^\top d_i))^{-1}$ for some $w \in \mathbb{R}^p$.

The parameters w follow the law $p(w|\alpha) = \mathcal{N}(0, \alpha^{-1} I_p)$, and $\alpha > 0$ is drawn from an exponential law $p(\alpha) = \mathrm{Exp}(0.01)$.

The parameter vector is then $x = [w, \log(\alpha)] \in \mathbb{R}^{p+1}$, and we use KSD-LBFGS to obtain samples from $p(x| \left(d_i, y_i\right)_{i=1}^q)$ for 13 datasets, with $N = 10$ particles for each.



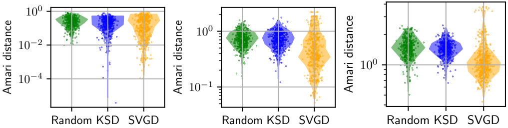
Accuracy of the KSD descent and SVGD on bayesian logistic regression for 13 datasets.
Both methods yield similar results.
KSD is better by 2% on one dataset.

2 - Bayesian Independent Component Analysis

ICA: $x = W^{-1}s$, where x is an observed sample in $\mathbb{R}^p$, $W \in \mathbb{R}^{p\times p}$ is the unknown square unmixing matrix, and $s \in \mathbb{R}^p$ are the independent sources.

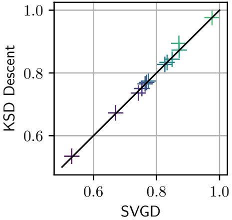
1)Assume that each component has the same density $s_i \sim p_s$.
2) The likelihood of the model is $p(x|W) = \log|W| + \sum_{i=1}^{p} p_s([Wx]_i)$.
3)Prior: W has i.i.d. entries, of law $\mathcal{N}(0, 1)$.

The posterior is $p(W|x) \propto p(x|W)p(W)$, and the score is given by $s(W) = W^{-\top} - \psi(Wx)x^\top - W$, where $\psi = -\frac{p_s'}{p_s}$. In practice, we choose p_s such that $\psi(\cdot) = \mathtt{tanh}(\cdot)$. We then use the presented algorithms to draw 10 particles $W \sim p(W|x)$ on 50 experiments.



Left: $p = 2$. Middle: $p = 4$. Right: $p = 8$.
Each dot = Amari distance between an estimated matrix and the true unmixing matrix.
KSD Descent is not better than random. Explanation: ICA likelihood is highly non-convex.

Real world experiments (10 particles)

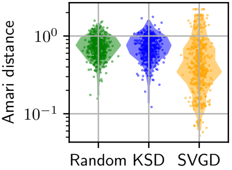


Bayesian logistic regression.

Accuracy of the KSD descent and SVGD for 13 datasets ($d \approx 50$).

Both methods yield similar results. KSD is better by 2% on one dataset.

Hint: convex likelihood.



Bayesian ICA.

Each dot is the Amari distance between an estimated matrix and the true unmixing matrix ($d \leq 8$).

KSD is not better than random.

Hint: highly non-convex likelihood.