Wasserstein Gradient Flows for Machine Learning

Anna Korba

CREST, ENSAE, Institut Polytechnique de Paris

Kantorovitch Initiative

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: Transport an initial probability distribution μ_0 on $\mathbb{R}^d$ to a target probability distribution π on $\mathbb{R}^d$.

Problem: Transport an initial probability distribution μ_0 on $\mathbb{R}^d$ to a target probability distribution π on $\mathbb{R}^d$.

Assume that $\pi \in \mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d), \int \|x\|^2 d\mu(x) < \infty\}$. This problem can be written as an optimization problem on $\mathcal{P}_2(\mathbb{R}^d)$, e.g.

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathrm{D}(\mu|\pi)$$

where D is a dissimilarity functional, seen as a loss, between probability distributions.

Problem: Transport an initial probability distribution μ_0 on $\mathbb{R}^d$ to a target probability distribution π on $\mathbb{R}^d$.

Assume that $\pi \in \mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d), \int \|x\|^2 d\mu(x) < \infty\}$. This problem can be written as an optimization problem on $\mathcal{P}_2(\mathbb{R}^d)$, e.g.

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathrm{D}(\mu|\pi)$$

where D is a dissimilarity functional, seen as a loss, between probability distributions.

Wasserstein Gradient Flows find *continuous* paths on $\mathcal{P}_2(\mathbb{R}^d)$ (equipped with the Wasserstein-2 geometry) that decrease this loss.

Problem: Transport an initial probability distribution μ_0 on $\mathbb{R}^d$ to a target probability distribution π on $\mathbb{R}^d$.

Assume that $\pi \in \mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d), \int \|x\|^2 d\mu(x) < \infty\}$. This problem can be written as an optimization problem on $\mathcal{P}_2(\mathbb{R}^d)$, e.g.

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathrm{D}(\mu|\pi)$$

where D is a dissimilarity functional, seen as a loss, between probability distributions.

Wasserstein Gradient Flows find *continuous* paths on $\mathcal{P}_2(\mathbb{R}^d)$ (equipped with the Wasserstein-2 geometry) that decrease this loss.

Different algorithms result from (1) the choice of *D* and (2) different time-space discretizations.

Many problems in ML can be formulated as the minimization of a functional on probability distributions :

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{G}(\mu), \quad \text{where } \mathcal{G}(\mu) = D(\mu|\pi)$$

▶ sampling (ex: π posterior distribution in Bayesian inference)

▶ optimizing Neural Networks (ex: π distribution over parameters of a big Neural Network)

▶ many others : generative modelling, reinforcement learning... [Chu et al., 2019]

One can design new schemes and/or study existing ones as discretizations of Wasserstein gradient flows.

Outline

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.

Def (pushforward) : Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $T : \mathbb{R}^d \to \mathbb{R}^d$. The pushforward measure $T_{\#}\mu$ is characterized by:

- $\forall$ B meas. set, $T_{\#}\mu(B) = \mu(T^{-1}(B))$
- $x \sim \mu$, $T(x) \sim T_{\#}\mu$

Def (pushforward) : Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $T : \mathbb{R}^d \to \mathbb{R}^d$. The pushforward measure $T_{\#}\mu$ is characterized by:

- ▶ $\forall$ B meas. set, $T_{\#}\mu(B) = \mu(T^{-1}(B))$
- ▶ $x \sim \mu$, $T(x) \sim T_{\#}\mu$

Brenier's theorem : Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ s.t. $\mu \ll$ *Leb*. Then,

- ▶ Then $\exists!$ $T_\mu^\nu : \mathbb{R}^d \to \mathbb{R}^d$ s.t. $T_{\mu\#}^\nu\mu = \nu$, and a convex function g s.t. $T_\mu^\nu = \nabla g$ μ-a.e.
- ▶ $W_2^2(\mu, \nu) = \|I - T_\mu^\nu\|_{L_2(\mu)}^2 = \inf_{T \in L_2(\mu)} \int (x - T(x))^2 d\mu(x)$ where $L^2(\mu) = \{f : \mathbb{R}^d \to \mathbb{R}^d, \int \|f(x)\|^2 d\mu(x) < \infty\}$

Def (pushforward) : Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $T : \mathbb{R}^d \to \mathbb{R}^d$. The pushforward measure $T_{\#}\mu$ is characterized by:

▶ $\forall$ B meas. set, $T_{\#}\mu(B) = \mu(T^{-1}(B))$

▶ $x \sim \mu$, $T(x) \sim T_{\#}\mu$

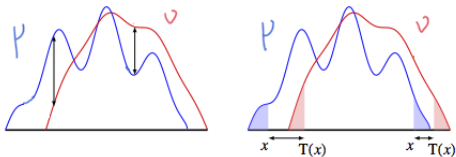**Brenier's theorem :** Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ s.t. $\mu \ll$ *Leb*. Then,

▶ Then $\exists! \ T_{\mu}^{\nu} : \mathbb{R}^d \to \mathbb{R}^d$ s.t. $T_{\mu\#}^{\nu}\mu = \nu$, and a convex function g s.t. $T_{\mu}^{\nu} = \nabla g$ μ-a.e.

▶ $W_2^2(\mu, \nu) = \|I - T_{\mu}^{\nu}\|_{L_2(\mu)}^2 = \inf_{T \in L_2(\mu)} \int (x - T(x))^2 d\mu(x)$
   where $L^2(\mu) = \{f : \mathbb{R}^d \to \mathbb{R}^d, \ \int \|f(x)\|^2 d\mu(x) < \infty\}$

W_2 geodesics?

$\rho(0) = \mu, \rho(1) = \nu$.

$\rho(t) = ((1-t)I + tT_{\mu}^{\nu})_{\#}\mu$

$\neq \rho(t) = \underbrace{(1-t)\mu + t\nu}_{\text{mixture}}$

Continuity equations

Let $T > 0$. Consider a family $\mu : [0, T] \to \mathcal{P}_2(\mathbb{R}^d), t \mapsto \mu_t$. It satisfies a continuity equation if there exists $(V_t)_{t \in [0,T]}$ such that $V_t \in L^2(\mu_t)$ and distributionnally:

$$\frac{\partial \mu_t}{\partial t} + \boldsymbol{\nabla} \cdot (\mu_t V_t) = 0.$$

Continuity equations

Let $T > 0$. Consider a family $\mu : [0, T] \to \mathcal{P}_2(\mathbb{R}^d), t \mapsto \mu_t$. It satisfies a continuity equation if there exists $(V_t)_{t \in [0,T]}$ such that $V_t \in L^2(\mu_t)$ and distributionnally:

$$\frac{\partial \mu_t}{\partial t} + \boldsymbol{\nabla} \cdot (\mu_t V_t) = 0.$$

Rules density μ_t of particles $x_t \in \mathbb{R}^d$ driven by a vector field V_t:

$$\frac{dx_t}{dt} = V_t(x_t)$$

Wasserstein Gradient Flows (WGF) [Ambrosio et al., 2008]

Let $\mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}, \mu \mapsto \mathcal{G}(\mu)$ a regular functional.

The first variation of $\mathcal{G}$ evaluated at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is the unique function $\frac{\partial \mathcal{G}(\mu)}{\partial \mu} \colon \mathbb{R}^d \to \mathbb{R}$ s. t. for any $\mu, \mu' \in \mathcal{P}_2(\mathbb{R}^d)$, s.t. $\mu' - \mu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[\mathcal{G}(\mu + \epsilon(\mu' - \mu)) - \mathcal{G}(\mu) \right] = \int_{\mathbb{R}^d} \frac{\partial \mathcal{G}(\mu)}{\partial \mu}(x) \, (d\mu' - d\mu)(x).$$

Wasserstein Gradient Flows (WGF) [Ambrosio et al., 2008]

Let $\mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}, \mu \mapsto \mathcal{G}(\mu)$ a regular functional.

The first variation of $\mathcal{G}$ evaluated at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is the unique function $\frac{\partial \mathcal{G}(\mu)}{\partial \mu} \colon \mathbb{R}^d \to \mathbb{R}$ s. t. for any $\mu, \mu' \in \mathcal{P}_2(\mathbb{R}^d)$, s.t. $\mu' - \mu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[\mathcal{G}(\mu + \epsilon(\mu' - \mu)) - \mathcal{G}(\mu) \right] = \int_{\mathbb{R}^d} \frac{\partial \mathcal{G}(\mu)}{\partial \mu}(x) \, (d\mu' - d\mu)(x).$$

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

$$\frac{\partial \mu_t}{\partial t} - \boldsymbol{\nabla} \cdot \left(\mu_t \nabla \frac{\partial \mathcal{G}(\mu_t)}{\partial \mu_t} \right) = 0, \text{ i.e. } V_t = -\nabla_{W_2} \mathcal{G}(\mu_t)$$

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

WGF of Free energies

In particular, if the functional $\mathcal{G}$ is a free energy:

$$\mathcal{G}(\mu) = \underbrace{\int H(\mu(x))dx}_{\text{internal energy } \mathcal{H}(\mu)} + \underbrace{\int V(x)d\mu(x)}_{\text{potential energy } \mathcal{E}_V(\mu)} + \underbrace{\int W(x,y)d\mu(x)d\mu(y)}_{\text{interaction energy } \mathcal{W}(\mu)}$$

$$\text{Then}: \ \frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \underbrace{\nabla(H'(\mu_t) + V + W * \mu_t)}_{\nabla_{W_2}\mathcal{G}(\mu)} \right). \quad (1)$$

For instance, if $H = 0$ then (1) rules the density μ_t of particles $x_t \in \mathbb{R}^d$ driven by :

$$\frac{dx_t}{dt} = -\nabla V(x_t) - \int_{\mathbb{R}^d} \nabla W(x, x_t)d\mu_t(x)$$

$\mu_t = Law(x_t)$.

Time discretizations

Time discretizations

For a step-size $\gamma > 0$:

1. Forward :

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

where $exp_\mu : L^2(\mu) \to \mathcal{P}, \phi \mapsto (I + \phi)_{\#}\mu$,

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

$$X_{l+1} = X_l - \gamma \nabla_{W_2}\mathcal{G}(\mu_l)(X_l) \sim \mu_{l+1}, \text{ if } X_l \sim \mu_l.$$

Time discretizations

For a step-size $\gamma > 0$:

1. Forward :

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

where $exp_\mu : L^2(\mu) \to \mathcal{P}, \phi \mapsto (I + \phi)_{\#}\mu$,

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

$$X_{l+1} = X_l - \gamma \nabla_{W_2} \mathcal{G}(\mu_l)(X_l) \sim \mu_{l+1}, \text{ if } X_l \sim \mu_l.$$

2. Backward :

$$\mu_{l+1} = \mathsf{JKO}_{\gamma\mathcal{G}}(\mu_l)$$

$$\text{where } \mathsf{JKO}_{\gamma\mathcal{G}}(\mu_l) = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \left\{ \mathcal{G}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \mu_l) \right\}.$$

Time discretizations

For a step-size $\gamma > 0$:

1. Forward :

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

where $exp_\mu : L^2(\mu) \to \mathcal{P}, \phi \mapsto (I + \phi)_{\#}\mu$,
and which corresponds in $\mathbb{R}^d$ to:

$$X_{l+1} = X_l - \gamma \nabla_{W_2}\mathcal{G}(\mu_l)(X_l) \sim \mu_{l+1}, \text{ if } X_l \sim \mu_l.$$

2. Backward :

$$\mu_{l+1} = \text{JKO}_{\gamma\mathcal{G}}(\mu_l)$$

where $\text{JKO}_{\gamma\mathcal{G}}(\mu_l) = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\text{argmin}} \left\{ \mathcal{G}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \mu_l) \right\}.$

3. Splitting schemes : if $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2$, e.g. Forward/Backward:

$$\nu_{l+1} = (I - \gamma \nabla_{W_2}\mathcal{G}_1(\mu_l))_{\#}\mu_l$$
$$\mu_{l+1} = \text{JKO}_{\gamma\mathcal{G}_2}(\nu_{l+1})$$

Space discretization - Interacting particle system

If the vector field depends on the density of the particles at time l, replace μ_l by the empirical measure of a system of n interacting particles:

$$X_0^1, \ldots, X_0^n \sim \mu_0$$

and for $j = 1, \ldots, n$:

$$
\begin{aligned}
X_{l+1}^j &= X_l^j - \gamma \nabla_{W_2} \mathcal{G}(\hat{\mu}_l)(X_l^j) \\
&= X_l^j - \frac{1}{\gamma} \left[\nabla V(X_l^j) + \frac{1}{n} \sum_{i=1}^n \nabla W(X_l^j, X_l^i) \right]
\end{aligned}
$$

where $\hat{\mu}_l = \frac{1}{n} \sum_{i=1}^n \delta_{X_l^j}$.

Outline

The relative entropy/Kullback-Leibler divergence

For any $\mu, \pi \in \mathcal{P}_2(\mathbb{R}^d)$, the Kullback-Leibler divergence of μ w.r.t. π is defined by

$$\mathrm{KL}(\mu|\pi) = \int_{\mathbb{R}^d} \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) \text{ if } \mu \ll \pi$$

and is $+\infty$ otherwise.

We consider the functional $\mathrm{KL}(\cdot|\pi) : \mathcal{P}_2(\mathbb{R}^d) \to [0, +\infty]$.

The relative entropy/Kullback-Leibler divergence

For any $\mu, \pi \in \mathcal{P}_2(\mathbb{R}^d)$, the Kullback-Leibler divergence of μ w.r.t. π is defined by

$$\mathrm{KL}(\mu|\pi) = \int_{\mathbb{R}^d} \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) \text{ if } \mu \ll \pi$$

and is $+\infty$ otherwise.

We consider the functional $\mathrm{KL}(\cdot|\pi) : \mathcal{P}_2(\mathbb{R}^d) \to [0, +\infty]$.

The differential of $\mathrm{KL}(\cdot|\pi)$ evaluated at μ, $\frac{\partial \mathrm{KL}(\mu|\pi)}{\partial \mu} : \mathbb{R}^d \to \mathbb{R}$ is the function

$$\log\left(\frac{\mu}{\pi}\right)(.) + 1 : \mathbb{R}^d \to \mathbb{R}.$$

Hence, for μ regular enough, $\nabla_{W_2} \mathrm{KL}(\mu|\pi)$ is:

$$\nabla \log\left(\frac{\mu}{\pi}\right)(.) : \mathbb{R}^d \to \mathbb{R}.$$

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.
$$y \sim f(x, z) + \epsilon \quad (p(y|x, z) \text{ gaussian})$$

$\implies$ Compute the likelihood: $p(\mathcal{D}|z) = \prod_{i=1}^m p(y_i|x_i, z)$.

▶ Assume a prior distribution on the parameter $z \sim p$.

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

π **is known up to a constant** since C is untractable.
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 1 : Bayesian statistics

▶ Let $\mathcal{D} = (x_i, y_i)_{i=1,\ldots,m}$ a labelled dataset.

▶ Assume an underlying model parametrized by $z \in \mathbb{R}^d$, e.g.
$$y \sim f(x, z) + \epsilon \quad (p(y|x, z) \text{ gaussian})$$
$\implies$ Compute the likelihood: $p(\mathcal{D}|z) = \prod_{i=1}^{m} p(y_i|x_i, z)$.

▶ Assume a prior distribution on the parameter $z \sim p$.

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

π **is known up to a constant** since C is untractable.
How to sample from π then? e.g. to compute:

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

1. **MCMC methods** (Markov Chain Monte Carlo)
2. **Sampling as optimization of the KL** [Wibisono, 2018]

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

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 $k(z, z') = \exp\left(-\frac{\|z-z'\|^2}{h}\right)$
  ▶ the Laplace kernel $k(z, z') = \exp\left(-\frac{\|z-z'\|}{h}\right)$
  ▶ the inverse multiquadratic kernel
    $k(z, z') = (c + \|z - z'\|)^{-\beta}$ with $\beta \in]0, 1[$

▶ $\mathcal{H}_k$ its corresponding RKHS (Reproducing Kernel Hilbert Space):

$$\mathcal{H}_k = \overline{\left\{ \sum_{i=1}^m \alpha_i k(\cdot, z_i); \ m \in \mathbb{N}; \ \alpha_1, \ldots, \alpha_m \in \mathbb{R}; \ z_1, \ldots, z_m \in \mathbb{R}^d \right\}}$$

▶ $\mathcal{H}_k$ is a Hilbert space with inner product $\langle ., . \rangle_{\mathcal{H}_k}$ and norm $\|.\|_{\mathcal{H}_k}$.

▶ It satisfies the reproducing property:

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

Maximum Mean Discrepancy [Gretton et al., 2012]

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

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

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

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

Maximum Mean Discrepancy [Gretton et al., 2012]

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

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

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

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

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

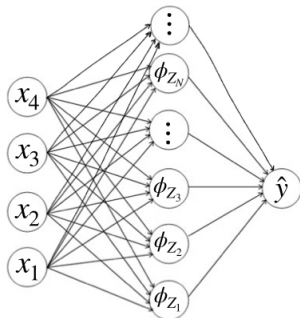Hence, for k regular enough, $\nabla_{W_2} \frac{1}{2} \mathrm{MMD}^2(\mu, \pi)$ is:

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

Example 2 : Regression with infinite width NN

$(x, y) \sim data$
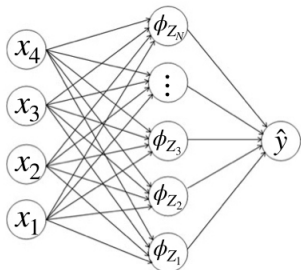
$$\min_{Z_1,...,Z_N} \mathbb{E}_{data}[\|y - \frac{1}{N} \sum_{i=1}^N \phi_{Z_i}(x)\|^2]$$

$N \to \infty$

$$\min_{\nu \in \mathscr{P}} \mathbb{E}_{data}[\|y - \mathbb{E}_{Z \sim \nu}[\phi_Z(x)]\|^2]$$

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

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

$$\Updownarrow$$

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

$$\Updownarrow$$

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

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

$$\Updownarrow$$

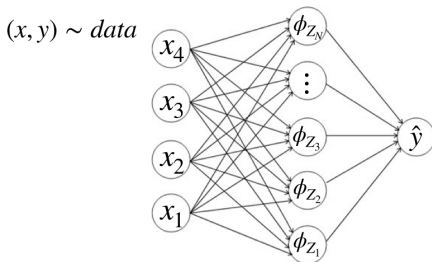$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{2} \text{MMD}^2(\mu, \pi)$$

Illustration : Student-Teacher network

Satisfies the "well-specified" assumption ! ($\exists \pi$, $\mathbb{E}[y|X = x] = \mathbb{E}_{Z \sim \pi}[\phi_Z(x)]$)

- ▶ 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 parametrized by $\mu_0 = \frac{1}{N} \sum_{j=1}^{N} \delta_{Z_0^j}$ tries to learn the

  mapping $x \mapsto \int \phi_Z(x) d\pi(Z)$.

$(x, y) \sim data$

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

Gradient descent on each parameter $j \in \{1, \dots, N\}$:

$$z_{l+1}^j = z_l^j - \gamma \mathbb{E}_{x \sim data} \left[\left(\frac{1}{N} \sum_{i=1}^N \phi_{z_l^i}(x) - \frac{1}{M} \sum_{m=1}^M \phi_{u^m}(x) \right) \nabla_{z_l^j} \phi_{z_l^j}(x) \right],$$

Re-arranging terms and recalling that
$k(z, z') = \mathbb{E}_{x \sim data}[\phi_z(x)^T \phi_{z'}(x)]$, the update becomes:

$$z_{l+1}^j = z_l^j - \gamma \underbrace{\left(\frac{1}{N} \sum_{i=1}^N \nabla_2 k(z_l^i, z_l^j) - \frac{1}{M} \sum_{m=1}^M \nabla_2 k(u^m, z_l^j) \right)}_{\nabla_{W_2} \frac{1}{2} \mathrm{MMD}_{\pi, \hat{\mu}_t}^2(z_l^j)}$$

The above equation is a time-discretized version of the gradient flow of the MMD.

KL and MMD are free energies

The **relative entropy** $\mathcal{G}(\mu) = \mathrm{KL}(\mu|\pi) = \int \log(\mu/\pi) d\mu$, $\pi \propto e^{-V}$ can be written:

$$\mathcal{G}(\mu) = \underbrace{\int H(\mu(x)) dx}_{\mathcal{H}(\mu)} + \underbrace{\int V(x)\mu(x) dx}_{\mathcal{E}_V(\mu)} - C,$$

$H(s) = s \log(s)$, $V(x) = -\log(\pi(x))$, $C = \mathcal{H}(\pi) + \mathcal{E}_V(\pi)$.

KL and MMD are free energies

The **relative entropy** $\mathcal{G}(\mu) = \mathrm{KL}(\mu|\pi) = \int \log(\mu/\pi)d\mu$, $\pi \propto e^{-V}$ can be written:

$$\mathcal{G}(\mu) = \underbrace{\int H(\mu(x))dx}_{\mathcal{H}(\mu)} + \underbrace{\int V(x)\mu(x)dx}_{\mathcal{E}_V(\mu)} - C,$$

$H(s) = s\log(s), \ V(x) = -\log(\pi(x)), \ C = \mathcal{H}(\pi) + \mathcal{E}_V(\pi).$

The **Maximum Mean Discrepancy** $\mathcal{G}(\mu) = \frac{1}{2}\mathrm{MMD}^2(\mu, \pi)$ also:

$$\mathcal{G}(\mu) = \underbrace{\int V(x)d\mu(x)}_{\mathcal{E}_V(\mu)} + \underbrace{\frac{1}{2}\int W(x, y)d\mu(x)d\mu(y)}_{\mathcal{W}(\mu)} + C,$$

$V(x) = -\int k(x, x')d\pi(x'), \ W(x, x') = k(x, x'), \ C = \mathcal{W}(\pi).$

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

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

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

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

KSD vs MMD

Under mild assumptions on k and π, the Stein kernel k_π is p.s.d. and satisfies a **Stein identity**

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

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

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

KSD as kernelized Fisher Divergence

Fisher Divergence:

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

"Kernelized" with k:

$$\text{KSD}^2(\mu|\pi) = \left\| S_{\mu,k} \nabla \log\left(\frac{\mu}{\pi}\right) \right\|_{\mathcal{H}_k}^2$$

$$= \int \nabla \log\left(\frac{\mu}{\pi}\right)(x) k(x,y) \nabla \log\left(\frac{\mu}{\pi}\right)(y) d\mu(x) d\mu(y)$$

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

$\implies$ minimizing the KSD is close in spirit to score-matching

[Hyvärinen and Dayan, 2005].

Convergence of WGF - Geodesic convexity

Convergence of the WGF $\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \nabla_{W_2} \mathcal{G}(\mu_t) \right)$ starting from μ_0?

Convergence of WGF - Geodesic convexity

Convergence of the WGF $\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \nabla_{W_2} \mathcal{G}(\mu_t) \right)$ starting from μ_0?

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

$$\mathcal{G}(\rho(t)) \leq (1-t)\mathcal{G}(\rho(0)) + t\mathcal{G}(\rho(1)) - t(1-t)\frac{\lambda}{2}W_2^2(\rho(0), \rho(1))^2$$

where $\rho(t) = ((1-t)I + tT_{\rho(0)}^{\rho(1)})_{\#}\rho(0)$

Convergence of WGF - Geodesic convexity

Convergence of the WGF $\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \nabla_{W_2} \mathcal{G}(\mu_t) \right)$ starting from μ_0?

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

$$\mathcal{G}(\rho(t)) \leq (1-t)\mathcal{G}(\rho(0)) + t\mathcal{G}(\rho(1)) - t(1-t)\frac{\lambda}{2}W_2^2(\rho(0), \rho(1))^2$$

where $\rho(t) = ((1-t)I + tT_{\rho(0)}^{\rho(1)})_{\#}\rho(0)$

Convergence of WGF - Geodesic convexity

Convergence of the WGF $\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \nabla_{W_2} \mathcal{G}(\mu_t) \right)$ starting from μ_0?

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

$$\mathcal{G}(\rho(t)) \leq (1-t)\mathcal{G}(\rho(0)) + t\mathcal{G}(\rho(1)) - t(1-t)\frac{\lambda}{2}W_2^2(\rho(0), \rho(1))^2$$

where $\rho(t) = ((1-t)I + tT_{\rho(0)}^{\rho(1)})_{\#}\rho(0)$

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

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

Convergence of WGF - Functional inequalities

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

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

Apply the chain rule in the Wasserstein space:

$$\begin{aligned}
\frac{d\mathcal{G}(\mu_t)}{dt} &= \left\langle V_t, \nabla_{W_2} \mathcal{G}(\mu_t) \right\rangle_{L^2(\mu_t)} \\
&= -\left\langle \nabla_{W_2} \mathcal{G}(\mu_t), \nabla_{W_2} \mathcal{G}(\mu_t) \right\rangle_{L^2(\mu_t)} \\
&= -\left\| \nabla_{W_2} \mathcal{G}(\mu_t) \right\|^2_{L^2(\mu_t)} \\
&\leq 0.
\end{aligned}$$

Assume a functional inequality of the form :

$$\left\| \nabla_{W_2} \mathcal{G}(\mu_t) \right\|^2_{L^2(\mu_t)} \geq \frac{1}{\lambda} \mathcal{G}(\mu_t).$$

Then, by Gronwall's lemma,

$$\mathcal{G}(\mu_t) \leq e^{-\lambda t} \mathcal{G}(\mu_0).$$

Choice of the functional?

Choose the functional depending on :

- ▶ the available information on the target π (samples, unnormalized density)
  - ▶ $\mathrm{KL}(.|\pi)$, $\mathrm{KSD}(.|\pi)$ require unnormalized density
  - ▶ $\mathrm{MMD}(.|\pi)$ requires samples of π
- ▶ its properties
  - ▶ the $\mathrm{KL}(.|\pi)$ is λ-geo convex if π is strongly logconcave ($\pi \propto \exp(-V)$ with V λ-convex), the $\mathrm{MMD}(.|\pi)$ and $\mathrm{KSD}(.|\pi)$ are not in general [Arbel et al., 2019, Korba et al., 2021]
  - ▶ the $\mathrm{KL}(.|\pi)$ satisfies a functional inequality for small perturbations of strongly log-concave distributions; $\mathrm{MMD}(.|\pi)$ and $\mathrm{KSD}(.|\pi)$ satisfy weaker functional inequalities [Arbel et al., 2019, Korba et al., 2021]
- ▶ practical optimization ($\mathrm{KSD}(.|\pi)$, $\mathrm{MMD}(.|\pi)$ can be used with L-BFGS algorithm, not $\mathrm{KL}(.|\pi)$))

Outline

The KL as a composite functional

$$\mathrm{KL}(\mu|\pi) = \int \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) \text{ if } \mu \ll \pi, +\infty \text{ else.}$$

The KL as a composite functional

$$\mathrm{KL}(\mu|\pi) = \int \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) \text{ if } \mu \ll \pi, +\infty \text{ else.}$$

It is written as **a composite functional** $(\pi \propto \exp(-V))$:

$$\mathrm{KL}(\mu|\pi) = \underbrace{\int V(x)d\mu(x)}_{\mathcal{E}_V(\mu) \text{ external potential}} + \underbrace{\int \log(\mu(x))d\mu(x)}_{\mathcal{H}(\mu) \text{ negative entropy}} + cte$$

The KL as a composite functional

$$\mathrm{KL}(\mu|\pi) = \int \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) \text{ if } \mu \ll \pi, +\infty \text{ else.}$$

It is written as **a composite functional** $(\pi \propto \exp(-V))$:

$$\mathrm{KL}(\mu|\pi) = \underbrace{\int V(x)d\mu(x)}_{\mathcal{E}_V(\mu) \text{ external potential}} + \underbrace{\int \log(\mu(x))d\mu(x)}_{\mathcal{H}(\mu) \text{ negative entropy}} + cte$$

The W_2 gradient flow of the KL is the Fokker-Planck equation:

$$\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \underbrace{\boldsymbol{\nabla} \log\left(\frac{\mu_t}{\pi}\right)}_{\nabla_{W_2} \mathrm{KL}(\mu_t|\pi)} \right) = \boldsymbol{\nabla} \cdot \left(\mu_t \underbrace{\boldsymbol{\nabla} V}_{\nabla_{W_2}\mathcal{E}_V(\mu)} \right) + \Delta(\mu_t).$$

The KL as a composite functional

$$\mathrm{KL}(\mu|\pi) = \int \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) \text{ if } \mu \ll \pi, +\infty \text{ else.}$$

It is written as **a composite functional** $(\pi \propto \exp(-V))$:

$$\mathrm{KL}(\mu|\pi) = \underbrace{\int V(x)d\mu(x)}_{\mathcal{E}_V(\mu) \text{ external potential}} + \underbrace{\int \log(\mu(x))d\mu(x)}_{\mathcal{H}(\mu) \text{ negative entropy}} + cte$$

The W_2 gradient flow of the KL is the Fokker-Planck equation:

$$\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \underbrace{\boldsymbol{\nabla} \log\left(\frac{\mu_t}{\pi}\right)}_{\nabla_{W_2} \mathrm{KL}(\mu_t|\pi)} \right) = \boldsymbol{\nabla} \cdot \left(\mu_t \underbrace{\boldsymbol{\nabla} V}_{\nabla_{W_2}\mathcal{E}_V(\mu)} \right) + \Delta(\mu_t).$$

It is the continuity equation $(X_t \sim \mu_t)$ of the Langevin diffusion :

$$dX_t = -\nabla V(X_t) + \sqrt{2}dB_t$$

where (B_t) is the brownian motion in $\mathbb{R}^d$.

Gradient flow of the entropy

The gradient flow of the negative entropy $\mathcal{H}(\mu)$ is the heat equation

$$\frac{\partial \mu_t}{\partial t} = \Delta \mu_t$$

This has an exact solution which is the heat flow $\mu_t = \mu_0 * \mathcal{N}(0, 2t I_d)$.

In space, this is implemented by adding Gaussian noise [2]

$$X_t = X_0 + \sqrt{2t} Z \qquad (2)$$

where $Z \sim \mathcal{N}(0, I_d)$ and Z independent of X_0.

---

[2] The true solution of the heat flow is the Brownian motion in space. However, at each time, the solution has the same distribution as (2)

Unadjusted Langevin Algorithm (ULA)

$$X_{l+1} = X_l - \gamma \nabla V(X_l) + \sqrt{2\gamma}\xi_l \ \text{ where } \xi_l \sim \mathcal{N}(0, I_d)$$

and $\gamma > 0$ is a step-size.

Unadjusted Langevin Algorithm (ULA)

$$X_{l+1} = X_l - \gamma \nabla V(X_l) + \sqrt{2\gamma}\xi_l \text{ where } \xi_l \sim \mathcal{N}(0, I_d)$$

and $\gamma > 0$ is a step-size.

Problem : ULA is biased (has stationary distribution $\pi_\gamma \neq \pi$).

Unadjusted Langevin Algorithm (ULA)

$$X_{l+1} = X_l - \gamma \nabla V(X_l) + \sqrt{2\gamma}\xi_l \text{ where } \xi_l \sim \mathcal{N}(0, I_d)$$

and $\gamma > 0$ is a step-size.

Problem : ULA is biased (has stationary distribution $\pi_\gamma \neq \pi$).

We can write ULA as the composition :

$Y_{l+1} = X_l - \gamma \nabla V(X_l)$ gradient descent/forward method for V

$X_{l+1} = Y_{l+1} + \sqrt{2\gamma}\xi_l$ exact solution for the heat flow

$\implies$ Forward-Flow discretization

Unadjusted Langevin Algorithm (ULA)

$$X_{l+1} = X_l - \gamma \nabla V(X_l) + \sqrt{2\gamma}\xi_l \text{ where } \xi_l \sim \mathcal{N}(0, I_d)$$

and $\gamma > 0$ is a step-size.

Problem : ULA is biased (has stationary distribution $\pi_\gamma \neq \pi$).

We can write ULA as the composition :

$Y_{l+1} = X_l - \gamma \nabla V(X_l)$ gradient descent/forward method for V

$X_{l+1} = Y_{l+1} + \sqrt{2\gamma}\xi_l$ exact solution for the heat flow

$\implies$ Forward-Flow discretization

In the space of measures $\mathcal{P}$:

$\nu_{l+1} = (I - \gamma \nabla V)_{\#}\mu_l$ gradient descent for $\mathcal{E}_V$

$\mu_{l+1} = \mathcal{N}(0, 2\gamma I) * \nu_{l+1}$ exact gradient flow for $\mathcal{U}$

This Forward-flow discretization is biased [Wibisono, 2018]**.**

Forward Backward discretization [Wibisono, 2018, Salim et al., 2020]

$$\mathcal{G}(\mu) = \mathcal{E}_V(\mu) + \mathcal{H}(\mu)$$

We analyzed :

$$\nu_{l+1} = (I - \gamma \nabla V)_{\#} \mu_l$$
$$\mu_{l+1} = \mathsf{JKO}_{\gamma \mathcal{H}}(\nu_{l+1})$$

where $\mathsf{JKO}_{\mathcal{H}}(\nu_{l+1}) = \mathrm{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{H}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \nu_{l+1})$.

Forward Backward discretization [Wibisono, 2018, Salim et al., 2020]

$$\mathcal{G}(\mu) = \mathcal{E}_V(\mu) + \mathcal{H}(\mu)$$

We analyzed :

$$\nu_{l+1} = (I - \gamma \nabla V)_{\#} \mu_l$$
$$\mu_{l+1} = \mathsf{JKO}_{\gamma \mathcal{H}}(\nu_{l+1})$$

where $\mathsf{JKO}_{\mathcal{H}}(\nu_{l+1}) = \mathrm{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{H}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \nu_{l+1})$.

We showed [Salim et al., 2020] that this scheme enjoys the same rates than proximal gradient in the euclidean setting, i.e.

Assume V is L-smooth, λ-strongly convex, and assume the step size $\gamma < 1/L$ and $\mu_0 \ll Leb$. Then for all $l \geq 0$:

1. $\mathcal{G}(\mu_l) - \mathcal{G}(\pi) \leq \frac{W_2^2(\mu_0, \pi)}{2\gamma l}$ in the convex case ($\lambda = 0$)

2. $W_2^2(\mu_l, \pi) \leq (1 - \gamma\lambda)^l W_2^2(\mu_0, \pi)$ when $\lambda > 0$

$\implies$ faster than ULA ($1/\sqrt{l}$ for $\lambda = 0$ and $1/l$ for $\lambda > 0$)

Implementation of the JKO of the negative entropy

▶ some subroutines exist to compute the JKO [Santambrogio, 2017], or the JKO w.r.t. the entropy-regularized W_2 [Peyré, 2015]

▶ It is possible to compute the JKO of negative entropy in closed form in the gaussian case (i.e. for π, μ_0 gaussians) [Wibisono, 2018].
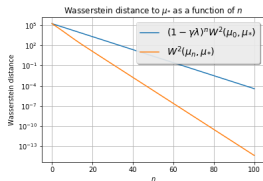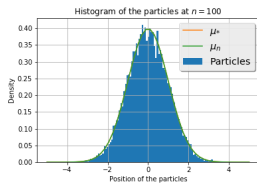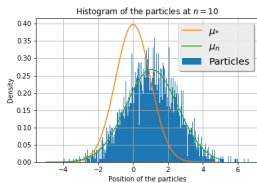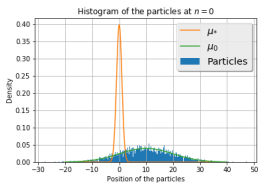
Figure: Convergence of μ_n to π (Top: d=1, Bottom: d=1000).

Forward discretization for the KL

Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Forward discretization (gradient descent on $(\mathcal{P}_2(\mathbb{R}^d), W_2)$) is written:

$$\mu_{l+1} = \left(I - \gamma \nabla \log\left(\frac{\mu_l}{\pi}\right)\right)_{\#} \mu_l \tag{3}$$

where $\gamma > 0$ is a step-size.

i.e. in $\mathbb{R}^d$, given $X_0 \sim \mu_0$,

$$X_{l+1} = X_l - \gamma \nabla \log\left(\frac{\mu_l}{\pi}\right)(X_l) \sim \mu_{l+1} \text{ if } X_l \sim \mu_l.$$

Forward discretization for the KL

Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Forward discretization (gradient descent on $(\mathcal{P}_2(\mathbb{R}^d), W_2)$) is written:

$$\mu_{l+1} = \left(I - \gamma \nabla \log \left(\frac{\mu_l}{\pi} \right) \right)_{\#} \mu_l \tag{3}$$

where $\gamma > 0$ is a step-size.

i.e. in $\mathbb{R}^d$, given $X_0 \sim \mu_0$,

$$X_{l+1} = X_l - \gamma \nabla \log \left(\frac{\mu_l}{\pi} \right) (X_l) \sim \mu_{l+1} \text{ if } X_l \sim \mu_l.$$

Problem: In practice, we do not know the density μ_l, we only have access to particles X_l.

We studied Stein Variational Gradient Descent [Liu and Wang, 2016], which proposes a particle scheme to implement (3).

Stein Variational Gradient Descent [Liu and Wang, 2016]

- Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a positive, semi-definite kernel

- example : $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{h}\right)$

- $\mathcal{H}_k : \overline{span(k(x, .), x \in \mathbb{R}^d)}^{\otimes d}$

  we assume : $\forall \mu, \int_{\mathbb{R}^d} k(x, x) d\mu(x) < \infty \implies \mathcal{H} \subset L^2(\mu)$.

- Define the **kernel integral operator** $S_\mu : L^2(\mu) \to \mathcal{H}_k$:

$$S_\mu f(\cdot) = \int k(x, .) f(x) d\mu(x) \quad \forall f \in L^2(\mu)$$

and denote $P_\mu = \iota_{\mathcal{H}_k \to L^2(\mu)} \circ S_\mu$.

> **SVGD trick:** under mild boundary conditions on k, π, applying this operator to the W_2 gradient of $\mathrm{KL}(\cdot|\pi)$ leads to
>
> $$P_\mu \nabla \log\left(\frac{\mu}{\pi}\right)(\cdot) = -\int [\nabla \log \pi(x) k(x, \cdot) + \nabla_x k(x, \cdot)] d\mu(x).$$

SVGD discrete time, infinite particles [Korba et al., 2020]

For the scheme:

$$\mu_{l+1} = \left(I - \gamma P_{\mu_l} \nabla \log \left(\frac{\mu_l}{\pi}\right)\right)_{\#} \mu_l$$

we showed a descent lemma, for a bounded of k, ∇k, Hessian of $V = \log \pi$, and gamma small enough :

$$\mathrm{KL}(\mu_{l+1}|\pi) - \mathrm{KL}(\mu_l|\pi) \leq -c_\gamma \underbrace{\left\| S_{\mu_l} \nabla \log \left(\frac{\mu_l}{\pi}\right) \right\|^2_{\mathcal{H}_k}}_{\mathsf{KSD}^2(\mu_l|\pi)}.$$

Rk: The KL is not smooth so such a descent lemma is specific to SVGD.

SVGD discrete time, infinite particles [Korba et al., 2020]

For the scheme:

$$\mu_{l+1} = \left(I - \gamma P_{\mu_l} \nabla \log \left(\frac{\mu_l}{\pi} \right) \right)_{\#} \mu_l$$

we showed a descent lemma, for a bounded of k, ∇k, Hessian of $V = \log \pi$, and gamma small enough :

$$\mathrm{KL}(\mu_{l+1}|\pi) - \mathrm{KL}(\mu_l|\pi) \leq -c_\gamma \underbrace{\left\| S_{\mu_l} \nabla \log \left(\frac{\mu_l}{\pi} \right) \right\|_{\mathcal{H}_k}^2}_{\mathsf{KSD}^2(\mu_l|\pi)}.$$

Rk: The KL is not smooth so such a descent lemma is specific to SVGD.

This descent lemma implies

$$\min_{k=1,\dots,l} \mathsf{KSD}^2(\mu_l|\pi) \leq \frac{1}{l} \sum_{k=1}^{l} \mathsf{KSD}^2(\mu_k|\pi) \leq \frac{\mathrm{KL}(\mu_0|\pi)}{c_\gamma l}.$$

Rk: Does not depend on the convexity of V.

Rates in terms of the KL objective?

To obtain rates, one may combine a descent lemma (1) of the form

$$\mathrm{KL}(\mu_{l+1}|\pi) - \mathrm{KL}(\mu_l|\pi) \leq -c_\gamma \left\| S_{\mu_l} \nabla \log \left(\frac{\mu_l}{\pi}\right) \right\|_{\mathcal{H}_k}^2$$

and the Stein log-Sobolev inequality (2) with constant λ:

$$\mathrm{KL}(\mu_{l+1}|\pi) - \mathrm{KL}(\mu_l|\pi) \underbrace{\leq}_{(1)} -c_\gamma \left\| S_{\mu_l} \nabla \log \left(\frac{\mu_l}{\pi}\right) \right\|_{\mathcal{H}_k}^2 \underbrace{\leq}_{(2)} -c_\gamma 2\lambda \, \mathrm{KL}(\mu_l|\pi).$$

Iterating this inequality yields $\mathrm{KL}(\mu_l|\pi) \leq (1 - 2c_\gamma\lambda)^l \, \mathrm{KL}(\mu_0|\pi)$.

Problem: In general, (2) does not hold [Duncan et al., 2019].

SVGD discrete time, finite particles

Algorithm : Starting from n i.i.d. samples $(X_0^i)_{i=1,\ldots,n} \sim \mu_0$, SVGD algorithm updates the n particles as follows :

$$X_{l+1}^i = X_l^i - \gamma \underbrace{\left[\frac{1}{n} \sum_{j=1}^n k(X_l^i, X_l^j) \nabla_{X_l^j} \log \pi(X_l^j) + \nabla_{X_l^j} k(X_l^j, X_l^i) \right]}_{P_{\hat{\mu}_l} \nabla \log\left(\frac{\hat{\mu}_l}{\pi}\right)(X_l^i)}$$

where $\hat{\mu}_l = \frac{1}{n} \sum_{j=1}^n \delta_{X_l^j}$. How far is $\hat{\mu}_l$ from $\bar{\mu}_l$ (empirical measure of n i.i.d. particles $\sim \mu_l$)?

SVGD discrete time, finite particles [Korba et al., 2020]

Algorithm : Starting from n i.i.d. samples $(X_0^i)_{i=1,\ldots,n} \sim \mu_0$, SVGD algorithm updates the n particles as follows :

$$X_{l+1}^i = X_l^i - \gamma \underbrace{\left[\frac{1}{n} \sum_{j=1}^n k(X_l^i, X_l^j) \nabla_{X_l^j} \log \pi(X_l^j) + \nabla_{X_l^j} k(X_l^j, X_l^i) \right]}_{P_{\hat{\mu}_l} \nabla \log\left(\frac{\hat{\mu}_l}{\pi}\right)(X_l^i)}$$

where $\hat{\mu}_l = \frac{1}{n} \sum_{j=1}^n \delta_{X_l^j}$. How far is $\hat{\mu}_l$ from $\bar{\mu}_l$ (empirical measure of n i.i.d. particles $\sim \mu_l$)?

Propagation of chaos result (non uniform in time)
Let $l \geq 0$ and $T > 0$. Under boundedness and Lipschitzness assumptions for all $k, \nabla k, V$; for any $0 \leq l \leq \frac{T}{\gamma}$ we have :

$$\mathbb{E}[W_2^2(\bar{\mu}_l, \hat{\mu}_l)] \leq \frac{1}{2} \left(\frac{1}{\sqrt{n}} \sqrt{var(\mu_0)} e^{LT} \right) (e^{2LT} - 1)$$

where L is a constant depending on k and π.

Open questions

Numerics;

► Closed-form or efficient subroutines for JKO (e.g. the JKO of the negative entropy)?

Theory:

► Rate of convergence in the KL objective for SVGD?
► uniform in time Propagation of chaos for a convex potential?

Outline

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, $err(x_1, \ldots, x_n) = \mathcal{O}(n^{-1/2})$
- ▶ **deterministic particle systems**, $err(x_1, \ldots, x_n)$?

Motivation

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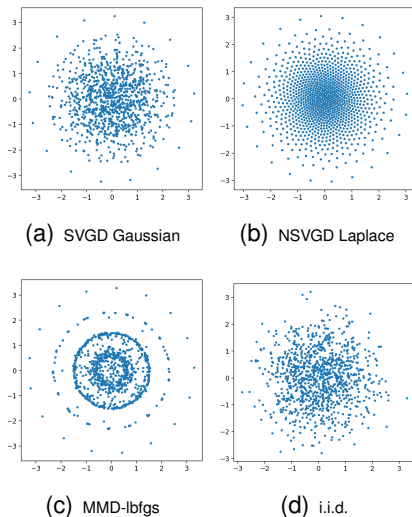
Figure: (a)-(c) Final states of the algorithms for 1024 particles, after 1e4 iterations. Ring structures tend to appear with the Gaussian kernel. The kernel bandwidth for all algorithm is set to 1.

We are interested in establishing bounds on the quantization error

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

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 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:

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

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

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

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):

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

- ▶ $\mathcal{D}(X_n, \pi) = 2^d \sup_{I = \prod_{i=1}^n [a_i, b_i]} |\pi(I) - \mu_n(I)|$ 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: 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.

Algorithms

we investigate numerically the quantization properties of :

- ▶ SVGD
- ▶ MMD descent
- ▶ KSD Descent
- ▶ Kernel Herding (KH) : greedy minimization of $MMD(.|\pi)$
- ▶ Stein points (SP) : greedy minimization of $KSD(.|\pi)$

Hyperparameters:

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

SVGD

We found that

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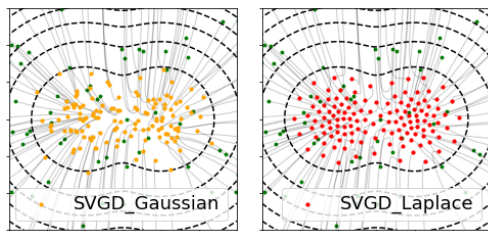
Figure: Example of a 2D Gaussian mixture. The configuration of 128 particles are plotted in green at initialization, and in different colors after convergence. The light grey curves correspond to their trajectories. From left to right: SVGD with Gaussian and Laplace kernel, γ=0.5, after 1000 iters

Quantization rates of the algorithms, $\pi = \mathcal{N}(0, \frac{1}{d}I_d)$
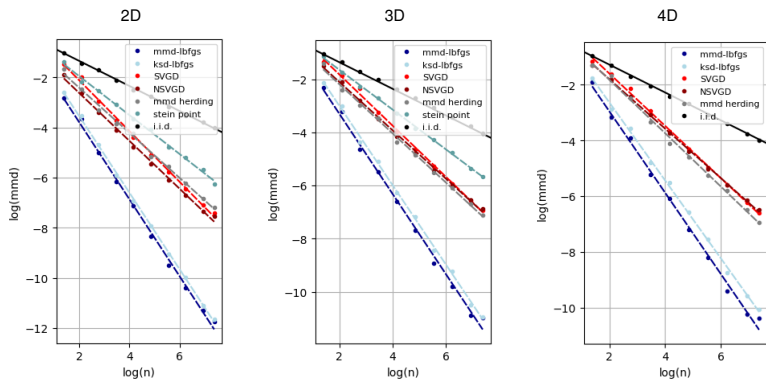
Figure: Each point is the result of averaging 3 runs of each algorithm run for 1e4 iterations, where the initial particles are i.i.d. samples of π. MMD/KSD Descent use bandwidth 1; SVGD use Laplace kernel with median trick; NSVGD use Laplace kernel with adaptive choice of bandwidth. Stein points use 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
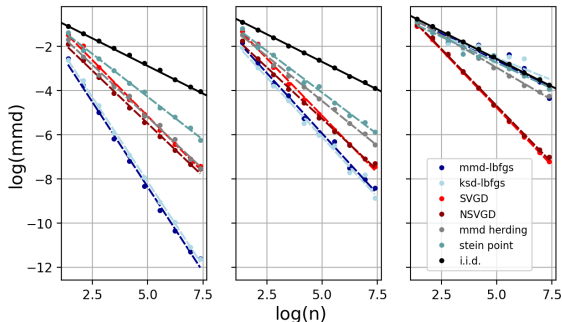
Figure: Importance of the choice of the bandwidth in the MMD evaluation metric when evaluating the final states, in 2D. From Left to Right: (evaluation) MMD bandwidth = 1, 0.7, 0.3.

- ▶ 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, open questions

Many problems in ML can be formulated as the minimization of a functional on probability distributions :

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{G}(\mu)$$

Conclusion, open questions

Many problems in ML can be formulated as the minimization of a functional on probability distributions :

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{G}(\mu)$$

Many ideas from optimization can be useful in this setting

(perturbation of dynamics, adapted discretizations...)

Open questions:

- ▶ numerics (improve the convergence of the schemes)
- ▶ theory : accurate rates of convergence in time and number of particles

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

📄 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.

📄 Chu, C., Blanchet, J., and Glynn, P. (2019).
Probability functional descent: A unifying perspective on gans, variational inference, and reinforcement learning.
arXiv preprint arXiv:1901.10691.

📄 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.

References III

📄 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. 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).

References IV

Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. (2021).
Kernel stein discrepancy descent.
arXiv preprint arXiv:2105.09994.

Korba, A., Salim, A., Arbel, M., Luise, G., and Gretton, A. (2020).
A non-asymptotic analysis for stein variational gradient descent.
arXiv preprint arXiv:2006.09797.

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.

References V

📄 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.

📄 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.

📄 Peyré, G. (2015).
Entropic approximation of wasserstein gradient flows.
SIAM Journal on Imaging Sciences, 8(4):2323–2351.

📄 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.

📄 Salim, A., Korba, A., and Luise, G. (2020).
Wasserstein proximal gradient.
arXiv preprint arXiv:2002.03035.

References VII

Santambrogio, F. (2017).
{Euclidean, metric, and Wasserstein} gradient flows: an overview.
Bulletin of Mathematical Sciences, 7(1):87–154.

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.

📄 Wibisono, A. (2018).
Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem.
arXiv preprint arXiv:1802.08089.