Wasserstein Gradient Flows for Machine Learning

Anna Korba

CREST, ENSAE, Institut Polytechnique de Paris

New Bridges between Mathematics and Data Science

Based on joint works with Adil Salim (KAUST), Giulia Luise (University College London), Arthur Gretton (Gatsby Unit, University College London), Michael Arbel (INRIA), Pierre-Cyril Aubin-Frankowski (Les Mines), Szymon Majewski (Polytechnique), Pierre Ablin (CNRS).

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

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

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

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

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

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

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 the choice of D, and different time-space discretizations.

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

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

- 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 [Chu et al., 2019],barycenters of distributions [Cuturi and Doucet, 2014]...

One can design new schemes/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,

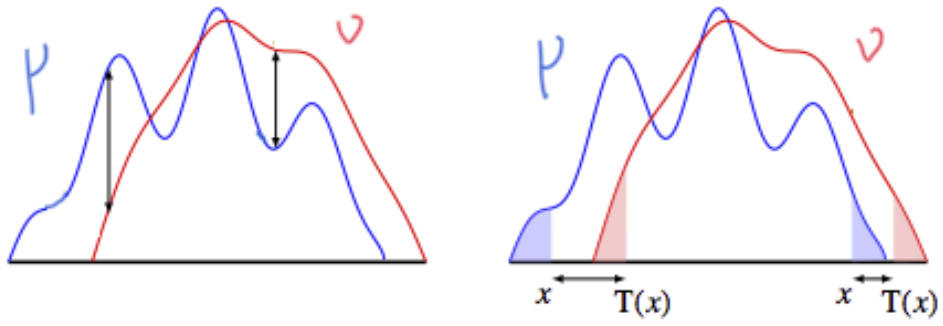
- Then $\exists!$ $T_\mu^\nu : \mathbb{R}^d \to \mathbb{R}^d$ s.t. $T_{\mu\#}^\nu \mu = \nu$, and a convex function g s.t. $T_\mu^\nu = \nabla g$ μ-a.e.
- $W_2^2(\mu, \nu) = \|I - T_\mu^\nu\|_{L_2(\mu)}^2 = \inf_{T \in L_2(\mu)} \int (x - T(x))^2 d\mu(x)$ where $L^2(\mu) = \{f : \mathbb{R}^d \to \mathbb{R}^d,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

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:

$$\boxed{\frac{\partial \mu_t}{\partial t} + div(\mu_t V_t) = 0}$$

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

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

Riemannian interpretation [Otto, 2001] :
The tangent space of $\mathcal{P}_2(\mathbb{R}^d)$ at μ_t verifies $\mathcal{T}_{\mu_t}\mathcal{P}_2(\mathbb{R}^d) \subset L^2(\mu_t)$.

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 differential of $\mathcal{G}$ evaluated at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is the unique function $\frac{\partial \mathcal{G}(\mu)}{\partial \mu}$:$\mathbb{R}^d \to \mathbb{R}$ 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 differential of $\mathcal{G}$ evaluated at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is the unique function $\frac{\partial \mathcal{G}(\mu)}{\partial \mu}$:$\mathbb{R}^d \to \mathbb{R}$ 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:

$$\frac{\partial \mu_t}{\partial t} - div\left(\mu_t \nabla \frac{\partial \mathcal{G}(\mu_t)}{\partial \mu_t}\right) = 0, \text{ i.e. } V_t = -\nabla_W \mathcal{G}(\mu)$$

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

WGF of Free energies

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

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

$$\text{Then : } \frac{\partial \mu_t}{\partial t} = div(\mu_t \underbrace{\nabla(H'(\mu_t) + V + W * \mu_t)}_{\nabla_W \mathcal{G}(\mu)}). \qquad (1)$$

For instance, if $\mathcal{H}(\mu)$ is the negative entropy ($H(s) = s\log(s)$), then (1) rules the density μ_t of particles $x_t \in \mathbb{R}^d$ driven by :

$$\frac{dx_t}{dt} = -\nabla V(x_t) - \int_{\mathbb{R}^d} \nabla W(x, x_t) d\mu_t(x) + \sqrt{2}dB_t,$$

$\mu_t = Law(x_t)$, B_t is a Brownian motion.

Space discretization

If the vector field depends on the density of the particles at time t, replace μ_t by the empirical measure of a system of N interacting particles:

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

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

$$\frac{d\hat{x}_t^j}{dt} = -\nabla V(\hat{x}_t^j) - \frac{1}{N}\sum_{i=1}^{N} \nabla W(\hat{x}_t^i, \hat{x}_t^j) + \sqrt{2}dB_t.$$

Time discretizations

Time discretizations

1. Forward :

$$\mu_{n+1} = exp_{\mu_n}(-\gamma \nabla_W \mathcal{G}(\mu_n)) = (I - \gamma \nabla_W \mathcal{G}(\mu_n))_{\#} \mu_n$$

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

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

$$X_{n+1} = X_n - \gamma \nabla_W \mathcal{G}(\mu_n)(X_n) \sim \mu_{n+1}, \text{ if } X_n \sim \mu_n.$$

Time discretizations

1. Forward :

$$\mu_{n+1} = exp_{\mu_n}(-\gamma \nabla_W \mathcal{G}(\mu_n)) = (I - \gamma \nabla_W \mathcal{G}(\mu_n))_{\#} \mu_n$$

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

$$X_{n+1} = X_n - \gamma \nabla_W \mathcal{G}(\mu_n)(X_n) \sim \mu_{n+1}, \text{ if } X_n \sim \mu_n.$$

2. Backward :

$$\mu_{n+1} = JKO_{\gamma \mathcal{G}}(\mu_n)$$

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

Time discretizations

1. Forward :

$$\mu_{n+1} = exp_{\mu_n}(-\gamma \nabla_W \mathcal{G}(\mu_n)) = (I - \gamma \nabla_W \mathcal{G}(\mu_n))_{\#} \mu_n$$

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

$$X_{n+1} = X_n - \gamma \nabla_W \mathcal{G}(\mu_n)(X_n) \sim \mu_{n+1}, \text{ if } X_n \sim \mu_n.$$

2. Backward :

$$\mu_{n+1} = JKO_{\gamma \mathcal{G}}(\mu_n)$$

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

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

$$\nu_{n+1} = (I - \gamma \nabla_W \mathcal{G}_1)_{\#} \mu_n$$
$$\mu_{n+1} = JKO_{\gamma \mathcal{G}_2}(\nu_{n+1})$$

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

For any $\mu \in \mathcal{P}_2(\mathbb{R}^d), \mu \ll \pi$, 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, $\nabla_W \mathrm{KL}(\cdot|\pi)$ is:

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

Example 1 : Bayesian statistics

- Let $\mathcal{D} = (w_i, y_i)_{i=1,\dots,N}$ observed data.
- Assume an underlying model parametrized by $\theta \in \mathbb{R}^d$ (e.g. $p(y|w,\theta)$ gaussian)

  $\Longrightarrow$ Likelihood: $p(\mathcal{D}|\theta) = \prod_{i=1}^N p(y_i|\theta, w_i)$.
- The parameter $\theta \sim p$ the prior distribution.

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

π **is known up to a constant** since Z is untractable.
How to sample from π then?

1. **MCMC methods** (Langevin Monte Carlo [Roberts and Tweedie, 1996], Hamiltonian Monte Carlo [Neal et al., 2011]...)
2. **Sampling as optimization of the KL** [Wibisono, 2018]

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

Maximum Mean Discrepancy [Gretton et al., 2012]

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

$$k(z, z') = \langle \phi(z), \phi(z') \rangle_{\mathcal{H}}, \quad \phi : \mathbb{R}^d \to \mathcal{H}$$

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

Maximum Mean Discrepancy [Gretton et al., 2012]

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

$$k(z, z') = \langle \phi(z), \phi(z') \rangle_{\mathcal{H}}, \quad \phi : \mathbb{R}^d \to \mathcal{H}$$

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

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

$$\frac{1}{2}\,\mathrm{MMD}^2(\mu, \pi) = \frac{1}{2} \int k(z, z') d\mu(z) d\mu(z') + \frac{1}{2} \int k(z, z') d\pi(z) d\pi(z') - \int k(z, z') d\mu(z) d\pi(z').$$

Maximum Mean Discrepancy [Gretton et al., 2012]

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

$$k(z, z') = \langle \phi(z), \phi(z') \rangle_{\mathcal{H}}, \quad \phi : \mathbb{R}^d \to \mathcal{H}$$

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

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

$$\frac{1}{2} \mathrm{MMD}^2(\mu, \pi) = \frac{1}{2} \int k(z, z') d\mu(z) d\mu(z') + \frac{1}{2} \int k(z, z') d\pi(z) d\pi(z') - \int k(z, z') d\mu(z) d\pi(z').$$

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

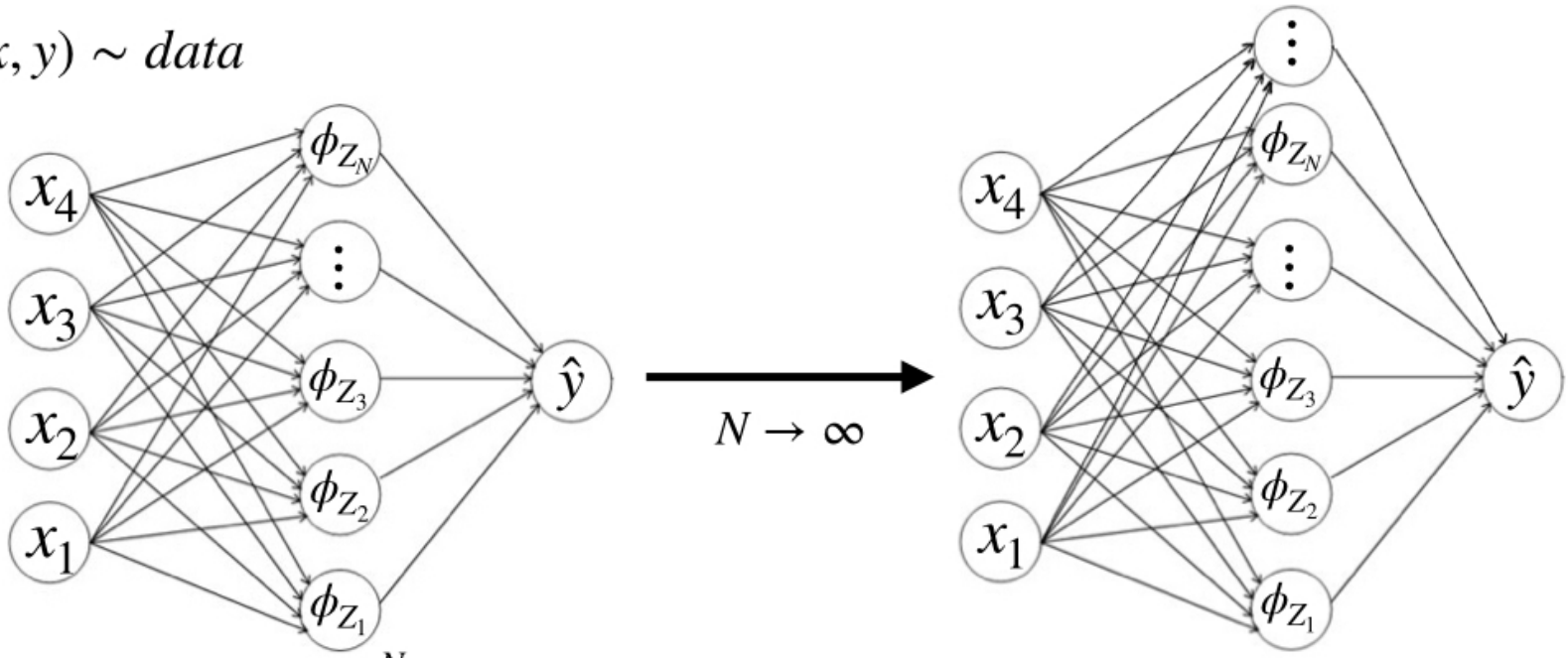
$$\int k(z, .) d\mu(z) - \int k(z, .) d\pi(z) : \mathbb{R}^d \to \mathbb{R}.$$

Hence, for k regular enough, $\nabla_W \frac{1}{2} \mathrm{MMD}^2(., \pi)$ 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

$(x, y) \sim data$



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

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

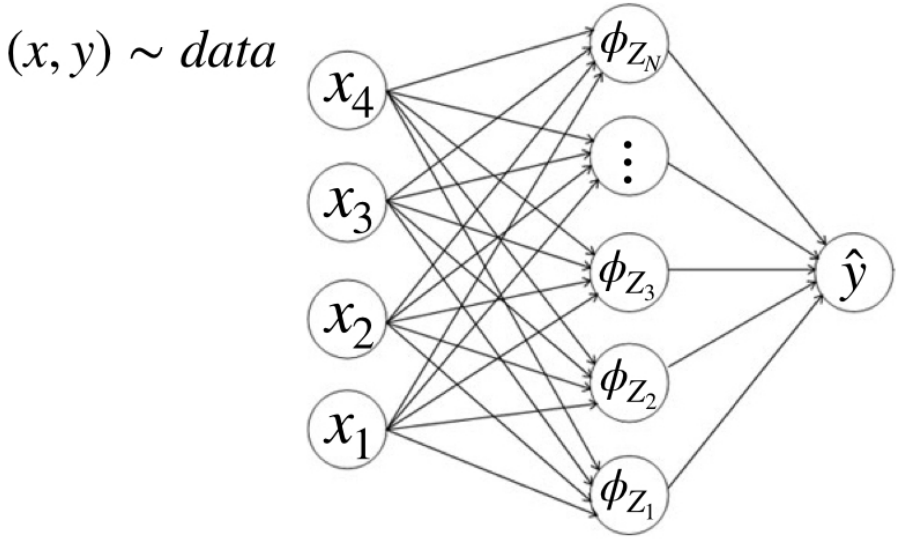
$$\Updownarrow$$

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

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



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

Gradient descent on each parameter $n \in \{1, \ldots, N\}$:

$$z_{t+1}^n = z_t^n - \gamma \mathbb{E}_{x \sim data} \left[\left(\frac{1}{N} \sum_{n'=1}^{N} \phi_{z_t^{n'}}(x) - \frac{1}{M} \sum_{m=1}^{M} \phi_{u^m}(x) \right) \nabla_{z_t^n} \phi_{z_t^n}(x) \right],$$

Re-arranging terms and recalling that
$k(Z, U) = \mathbb{E}_{x \sim data}[\phi_Z(x)^T \phi_U(x)]$, the update becomes:

$$z_{t+1}^n = z_t^n - \gamma \underbrace{\left(\frac{1}{N} \sum_{n'=1}^{N} \nabla_2 k(z_t^{n'}, z_t^n) - \frac{1}{M} \sum_{m=1}^{M} \nabla_2 k(u^m, z_t^n) \right)}_{\nabla_W \frac{1}{2} \mathrm{MMD}^2_{\pi, \hat{\mu}_t}(z_t^n)}$$

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

KL and MMD are free energies

The **relative entropy** $\mathcal{G}(\mu) = \mathrm{KL}(\mu|\pi)$ can be written:

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

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

Application : sampling from a posterior distribution $\pi \propto \exp(-V)$ in Bayesian inference.

KL and MMD are free energies

The **relative entropy** $\mathcal{G}(\mu) = \mathrm{KL}(\mu|\pi)$ can be written:

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

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

Application : sampling from a posterior distribution $\pi \propto \exp(-V)$ in Bayesian inference.

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

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

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

Application : optimizing infinite-width 1 hidden layer NN where π is the optimal distribution.

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:

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

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:

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

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 $\mu_t = \mu_0 * \mathcal{N}(0, 2tI_d)$.

In space, this is implemented by adding Gaussian noise [1]

$$X_t = X_0 + \sqrt{2t}Z \tag{2}$$

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

[1] 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)

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

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

Unadjusted Langevin Algorithm (ULA)

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

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

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

Unadjusted Langevin Algorithm (ULA)

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

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 :

$$Y_{n+1} = X_n - \gamma \nabla V(X_n) \quad \text{gradient descent/forward method for V}$$

$$X_{n+1} = Y_{n+1} + \sqrt{2\gamma}\xi_n \quad \text{exact solution for the heat flow}$$

$\Longrightarrow$ Forward-Flow discretization

Unadjusted Langevin Algorithm (ULA)

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

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 :

$$Y_{n+1} = X_n - \gamma \nabla V(X_n) \quad \text{gradient descent/forward method for V}$$

$$X_{n+1} = Y_{n+1} + \sqrt{2\gamma}\xi_n \quad \text{exact solution for the heat flow}$$

$\Longrightarrow$ Forward-Flow discretization

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

$$\nu_{n+1} = (I - \gamma \nabla V)_{\#}\mu_n \quad \text{gradient descent for } \mathcal{E}_V$$

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

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

Outline

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

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

We analyzed :

$$\nu_{n+1} = (I - \gamma \nabla V)_{\#} \mu_n$$
$$\mu_{n+1} = JKO_{\gamma \mathcal{H}}(\nu_{n+1})$$

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

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

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

We analyzed :

$$\nu_{n+1} = (I - \gamma \nabla V)_{\#} \mu_n$$
$$\mu_{n+1} = JKO_{\gamma \mathcal{H}}(\nu_{n+1})$$

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

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 $n \geq 0$:

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

$\Longrightarrow$ faster than ULA ($1/\sqrt{n}$ for $\lambda = 0$ and $1/n$ 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].

../../../20-wass-prox-grad/neurips/Sim

Forward discretization for the KL

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

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

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

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

$$X_{n+1} = X_n - \gamma \nabla \log\left(\frac{\mu_n}{\pi}\right)(X_n) \sim \mu_{n+1} \text{ if } X_n \sim \mu_n.$$

Forward discretization for the KL

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

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

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

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

$$X_{n+1} = X_n - \gamma \nabla \log\left(\frac{\mu_n}{\pi}\right)(X_n) \sim \mu_{n+1} \text{ if } X_n \sim \mu_n.$$

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

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 : $k(x, y) = \exp(-\frac{\|x-y\|^2}{h})$
- $\mathcal{H}$ its RKHS : $\overline{\left\{f : \mathbb{R}^d \to \mathbb{R}, f(.) = \sum_{i=1}^n a_i k(x_i, .)\right\}}^{\otimes d}$
  Hilbert space of functions equipped with $\langle \cdot, \cdot \rangle_{\mathcal{H}}, \| \cdot \|_{\mathcal{H}}$.
  we assume : $\forall \mu$, $\int_{\mathbb{R}^d} k(x, x) d\mu(x) < \infty \Longrightarrow \mathcal{H} \subset L^2(\mu)$.
- Define the **kernel integral operator** $S_\mu : L^2(\mu) \to \mathcal{H}$:

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

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

SVGD trick: under mild boundary conditions, 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:

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

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

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

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:

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

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

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

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

This descent lemma implies

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

Rk: Does not depend on the convexity of V.

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

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

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

where $\hat{\mu}_n = \frac{1}{N} \sum_{j=1}^{N} \delta_{X_n^j}$. How far is $\hat{\mu}_n$ from $\mu_n^{\otimes N}$?

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

$$\mathbb{E}[W_2^2(\mu_n^{\otimes N}, \hat{\mu}_n)] \leq \frac{1}{2} \left(\frac{1}{\sqrt{N}} \sqrt{var(\mu_0)} e^{LT} \right) (e^{2LT} - 1)$$

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

MMD gradient flow [Arbel et al., 2019]

We noticed convergence issues for the regular dynamics.

We proposed the following perturbation.

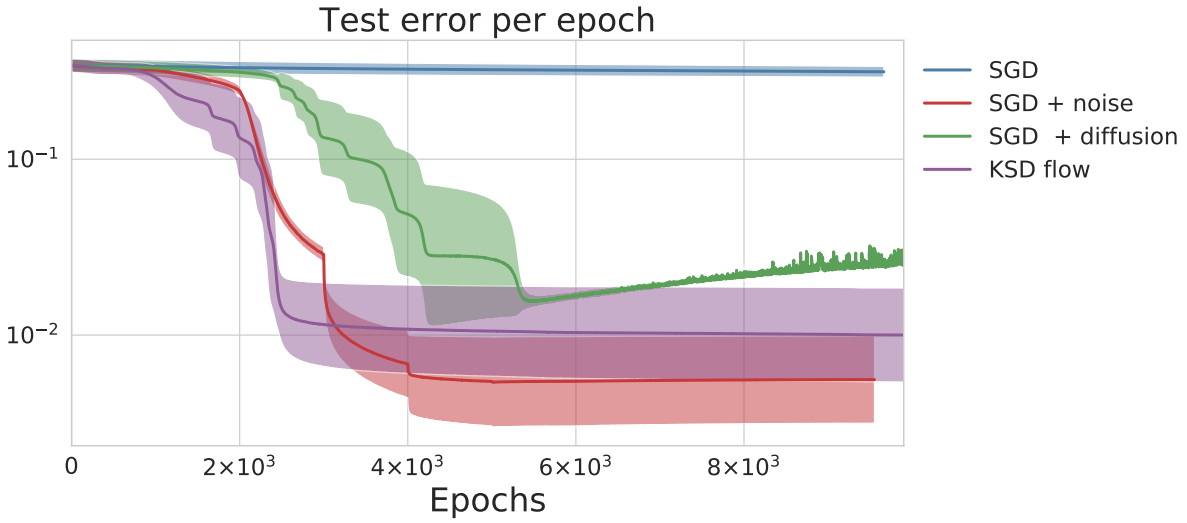
At each iteration n, sample $\xi_n \sim \mathcal{N}(0,1)$ and β_n is the noise level:

$$X_{n+1} = X_n - \gamma \nabla_W \mathrm{MMD}^2(X_n + \beta_n \xi_n)$$

Different from adding noise outside ("diffusion")

$$X_{n+1} = X_n - \gamma \nabla_W \mathrm{MMD}^2(X_n) + \beta_n \xi_n$$

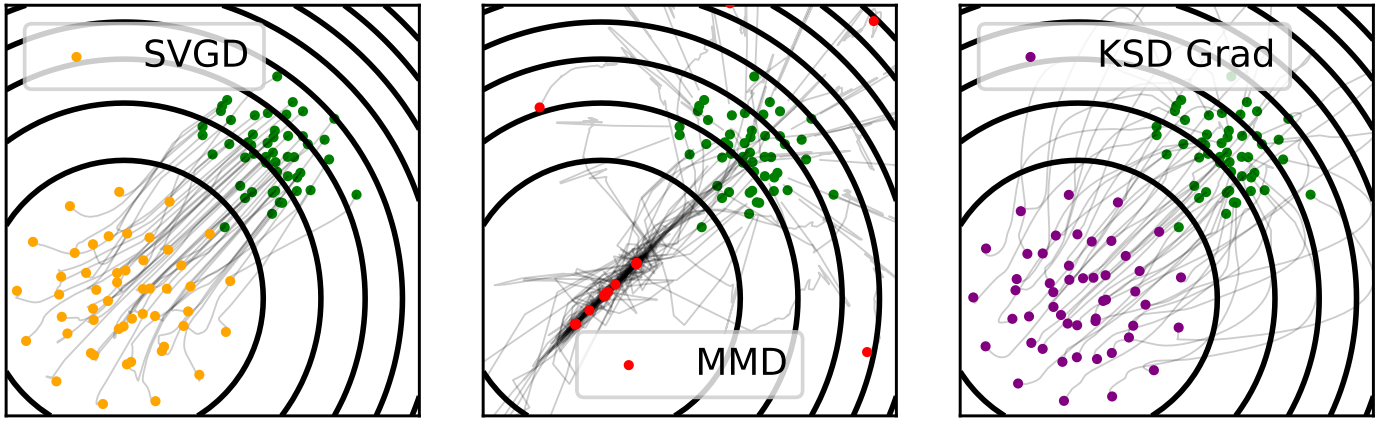
(which corresponds to an entropic regularization of the loss.)



Training of a neural network.

KSD gradient flow [Korba et al., 2021]

Idea: implement a forward discretization for the KSD.



The green points represent the initial positions of the particles.
The light grey curves correspond to their trajectories.

Conclusion, open questions

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

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

Conclusion, open questions

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

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

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 (obtain finer guarantees)

Thank you!

References I

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.

Boursier, E. and Perchet, V. (2019).
Utility/privacy trade-off through the lens of optimal transport.
In *International Conference on Artificial Intelligence and Statistics (AISTATS), Palermo, Italie*.

References II

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.

Cuturi, M. and Doucet, A. (2014).
Fast computation of wasserstein barycenters.

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

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

References III

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

Neal, R. M. et al. (2011).
Mcmc using hamiltonian dynamics.
Handbook of markov chain monte carlo, 2(11):2.

References IV

Otto, F. (2001).
The Geometry of Dissipative Evolution Equations: The Porous Medium Equation.
Communications in Partial Differential Equations, 26(1-2):101–174.

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

Roberts, G. O. and Tweedie, R. L. (1996).
Exponential convergence of langevin distributions and their discrete approximations.
Bernoulli, pages 341–363.

References V

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

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

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

Identification of the optimal transport maps

From μ_n to $\nu_{n+1} = (I - \gamma \nabla V)_{\#} \mu_n$:

Assumption : V is L-smooth i.e. $\forall\, (x, y) \in \mathbb{R}^d$,

$$V(y) \leq V(x) + \langle \nabla V(x), y - x \rangle + \frac{L}{2} \|x - y\|^2.$$

Then : If $\mu_0 \ll Leb$ and $\gamma < 1/L$, the OT map from μ_n to ν_{n+1} corresponds to :

$$T_{\mu_n}^{\nu_{n+1}} = (I - \gamma \nabla V)$$

and $\nu_{n+1} \ll Leb$.

Proof : $(I - \gamma \nabla V)$ is the gradient of a convex function for $\gamma < 1/L$.

Identification of the optimal transport maps

From ν_{n+1} to $\mu_{n+1} \in JKO_{\gamma\mathcal{H}}(\nu_{n+1})$:

There exists a strong Fréchet subgradient at ν_{n+1} denoted $\nabla_W \mathcal{H}(\mu_{n+1})$, such that the OT map from ν_{n+1} to μ_{n+1} corresponds to :

$$T_{\mu_{n+1}}^{\nu_{n+1}} = I + \gamma \nabla_W \mathcal{H}(\mu_{n+1})$$

and $\mu_{n+1} \ll Leb$ [Ambrosio et al., 2008].

By Brenier's theorem ($T_{\mu_{n+1}}^{\nu_{n+1}} \circ T_{\nu_{n+1}}^{\mu_{n+1}} = I$) this also means

$$\mu_{n+1} = (I - \gamma \nabla_W \mathcal{H}(\mu_{n+1}) \circ T_{\nu_{n+1}}^{\mu_{n+1}})_{\#} \nu_{n+1}.$$

Generalized geodesic convexity of $\mathcal{H}$

Key assumption : $\mathcal{H}$ is convex along *generalized geodesics* defined by W_2, i.e. for any $\mu, \pi, \nu \in \mathcal{P}$ with $\nu \ll Leb$, $t \in [0, 1]$:

$$\mathcal{H}((tT_\nu^\pi + (1-t)T_\nu^\mu)_\# \nu) \leq t\mathcal{H}(\pi) + (1-t)\mathcal{H}(\mu)$$

where T_ν^π and T_ν^μ are the OT maps from ν to π and from ν to μ.

Generalized geodesic convexity of $\mathcal{H}$

Key assumption : $\mathcal{H}$ is convex along *generalized geodesics* defined by W_2, i.e. for any $\mu, \pi, \nu \in \mathcal{P}$ with $\nu \ll Leb$, $t \in [0, 1]$:

$$\mathcal{H}((tT_\nu^\pi + (1-t)T_\nu^\mu)_\# \nu) \leq t\mathcal{H}(\pi) + (1-t)\mathcal{H}(\mu)$$

where T_ν^π and T_ν^μ are the OT maps from ν to π and from ν to μ.

This enables us to prove a **descent lemma** for V being L-smooth and $\gamma < 1/L$:

$$\mathrm{KL}(\mu_{n+1}|\pi) \leq \mathrm{KL}(\mu_n|\pi) - \gamma\left(1 - \frac{L\gamma}{2}\right)\|\nabla V + \nabla_W \mathcal{H}(\mu_{n+1}) \circ X_{n+1}\|^2_{L_2(\mu_n)},$$

where $X_{n+1} = T_{\nu_{n+1}}^{\mu_{n+1}} \circ (I - \gamma \nabla V)$.

A dual point of view

Consider the gradient flow of $V : \mathbb{R}^d \to \mathbb{R}$

$$x'(t) = -\nabla V(x(t))$$

for $V : \mathbb{R}^d \to \mathbb{R}$ smooth and assume $x(0)$ random with density μ_0. What is the dynamics of the density μ_t of $x(t)$?

[2] $\mathcal{C}^\infty$ function from $\mathbb{R}^d$ to $\mathbb{R}$ with compact support.

A dual point of view

Consider the gradient flow of $V : \mathbb{R}^d \to \mathbb{R}$

$$x'(t) = -\nabla V(x(t))$$

for $V : \mathbb{R}^d \to \mathbb{R}$ smooth and assume $x(0)$ random with density μ_0. What is the dynamics of the density μ_t of $x(t)$?

Let $\phi : \mathbb{R}^d \to \mathbb{R}$ a test function[2].

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

and

$$\frac{d}{dt}\mathbb{E}(\phi(x(t))) = -\int \langle \nabla\phi, \nabla V\rangle \mu_t(x)dx = \int \phi(x) div(\mu_t \nabla V)(x)dx,$$

Therefore,

$$\frac{\partial \mu_t}{\partial t} = div(\mu_t \nabla V).$$

[2] $\mathcal{C}^\infty$ function from $\mathbb{R}^d$ to $\mathbb{R}$ with compact support.

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

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

$\Longrightarrow$ We propose to analyze [Wibisono, 2018] :

$$\nu_{n+1} = (I - \gamma \nabla V)_{\#}\mu_n$$
$$\mu_{n+1} = JKO_{\gamma\mathcal{H}}(\nu_{n+1})$$

where $JKO_{\mathcal{H}}(\nu_{n+1}) = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\text{argmin}} \ \mathcal{H}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \nu_{n+1})$.

Tools for the proof :

- Identification of OT maps
- use geodesic convexity (convexity of V and generalized geodesic convexity of $\mathcal{H}$)

Descent Lemma in the smooth case

Key assumption : $\mathcal{H}$ is convex along *generalized geodesics* defined by W_2, i.e. for any $\mu, \nu, \mu^* \in \mathcal{P}$ with $\nu \ll Leb$, $t \in [0, 1]$:

$$\mathcal{H}((tT^{\nu}_{\mu^*} + (1-t)T^{\mu}_{\mu^*})_{\#}\mu^*) \leq t\mathcal{H}(\nu) + (1-t)\mathcal{H}(\mu).$$

$T^{\nu}_{\mu^*}$ and $T^{\mu}_{\mu^*}$ are the OT maps from μ^* to ν and from μ^* to μ.

Descent Lemma in the smooth case

Key assumption : $\mathcal{H}$ is convex along *generalized geodesics* defined by W_2, i.e. for any $\mu, \nu, \mu^* \in \mathcal{P}$ with $\nu \ll Leb$, $t \in [0,1]$:

$$\mathcal{H}((tT^{\nu}_{\mu^*} + (1-t)T^{\mu}_{\mu^*})_{\#}\mu^*) \leq t\mathcal{H}(\nu) + (1-t)\mathcal{H}(\mu).$$

$T^{\nu}_{\mu^*}$ and $T^{\mu}_{\mu^*}$ are the OT maps from μ^* to ν and from μ^* to μ.

Result: A **descent lemma** for V being L-smooth[a] and $\gamma < 1/L$:

$$\mathcal{G}(\mu_{n+1}) \leq \mathcal{G}(\mu_n) - \gamma\left(1 - \frac{L\gamma}{2}\right)\|\nabla V + \nabla_W \mathcal{H}(\mu_{n+1}) \circ X_{n+1}\|^2_{L_2(\mu_n)},$$

where $X_{n+1} = T^{\mu_{n+1}}_{\nu_{n+1}} \circ (I - \gamma\nabla V)$.

[a] i.e. $\forall\,(x,y) \in \mathbb{R}^d$, $V(y) \leq V(x) + \langle \nabla V(x), y - x\rangle + \frac{L}{2}\|x-y\|^2$.

Rates of convergence in the convex case

Assumption : V is λ-strongly convex, i.e. $\forall\ (x, y) \in \mathbb{R}^d$,

$$V(x) + \langle \nabla V(x), y - x \rangle + \frac{\lambda}{2}\|x - y\|^2 \leq V(y).$$

Rates of convergence in the convex case

Assumption : V is λ-strongly convex, i.e. $\forall\,(x,y) \in \mathbb{R}^d$,

$$V(x) + \langle \nabla V(x), y - x \rangle + \frac{\lambda}{2}\|x - y\|^2 \leq V(y).$$

Theorem : Assume the step size $\gamma < 1/L$ and $\mu_0 \ll Leb$. Then for all $n \geq 0$

$$W_2^2(\mu_{n+1}, \pi) \leq (1 - \gamma\lambda) W_2^2(\mu_n, \pi) - 2\gamma(\mathcal{G}(\mu_{n+1}) - \mathcal{G}(\pi)).$$

which implies:

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

$\Longrightarrow$ same rates than proximal gradient in the euclidean setting!
$\Longrightarrow$ faster than ULA ($1/\sqrt{n}$ for $\lambda = 0$ and $1/n$ for $\lambda > 0$)

Closed-form for the Gaussian case

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].

Assume $\pi = \mathcal{N}(m, \Sigma)$.

Let $\mu_0 = \mathcal{N}(m_0, \Sigma_0)$ and let $\Sigma_0 = I$ for simplicity, so Σ_0 commutes with Σ. Along FB, $\mu_n = \mathcal{N}(m_n, \Sigma_n)$ stays Gaussian, and:

$$
\begin{aligned}
y_{n+1} &= m + (I - \gamma\Sigma^{-1})(x_n - m) \\
x_{n+1} &= m_{n+1} + (I - \gamma\Sigma_{n+1}^{-1})^{-1}(y_{n+1} - \mu_n)
\end{aligned}
$$

where

$$
\begin{aligned}
\mu_{n+1} &= m + (I - \gamma\Sigma^{-1})(\mu_n - m) \\
\Sigma_{n+1}(I - \gamma\Sigma_{n+1}^{-1})^2 &= \Sigma_n(I - \gamma\Sigma^{-1})^2
\end{aligned}
$$