Wasserstein Proximal Gradient

Adil Salim [1] **Anna Korba** [2] Giulia Luise [3]

[1]VCC, KAUST, Saudi Arabia [2]CREST, ENSAE, Institut Polytechnique de Paris

[3]Department of Computer Science, University College London

Entropic regularization of OT and applications



Problem

Let $\mathcal{P}_2(\mathbb{R}^d) = \left\{\mu \in \mathcal{P}(\mathbb{R}^d), \int \|x\|^2 d\mu(x) < \infty\right\}$, and $V : \mathbb{R}^d \to \mathbb{R}$, $\mathcal{H} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$.

We consider the problem

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{G}(\mu) := \underbrace{\int V d\mu}_{\mathcal{E}_V(\mu)} + \mathcal{H}(\mu).$$

Examples:

1. Sampling (*e.g.* when $\mathcal{G}(\mu) = \mathrm{KL}(\mu|\pi)$, where π is a target distribution) [Cheng *et al.*'17, Wibisono'18, Bernton'18, Durmus *et al.*'19, Arbel *et al.*'19].
2. Optimization of overparametrized shallow neural networks (*e.g.* when $\mathcal{G}(\mu) = \mathrm{MMD}(\mu, \pi)$, where π is the optimal distribution over parameters) [Chizat *et al.*'18, Arbel *et al.*'19].

Contributions of this paper

- This problem is a free energy minimization for which **Wasserstein gradient flows** are well understood continuous time minimization dynamics [Ambrosio *et al.*'08].
- Various time-discretizations have been considered in the literature, see *e.g.* [Jordan *et al.*'98, Wibisono'18].

In this work, we propose a **Forward Backward (FB) discretization scheme** that can tackle the case where the objective function is the sum of a smooth and a nonsmooth terms.

We show that it has convergence guarantees similar to the analog scheme in Euclidean spaces, under mild assumptions on V and $\mathcal{H}$.

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.
- $W_2^2(\mu, \nu) = \|I - T_\mu^\nu\|_{L_2(\mu)}^2 = \inf_{T \in L_2(\mu)} \int (x - T(x))^2 d\mu(x)$
- Also if $\nu \ll Leb$, then $T_\mu^\nu \circ T_\nu^\mu = I$ ν-a.e. and $T_\nu^\mu \circ T_\mu^\nu = I$ μ-a.e.

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,

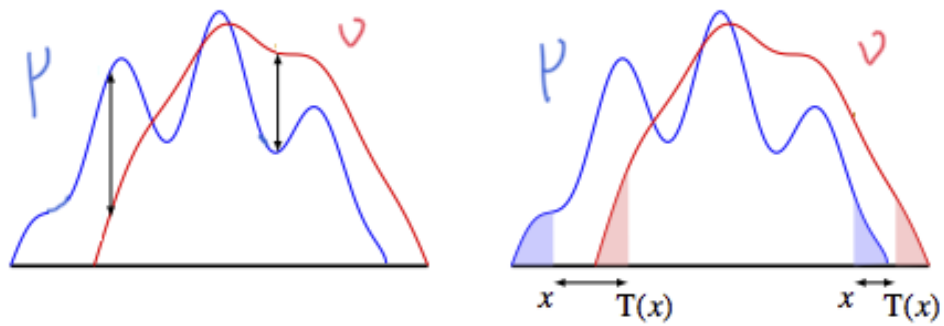
- Then $\exists!$ $T_\mu^\nu : \mathbb{R}^d \to \mathbb{R}^d$ s.t. $T_{\mu\#}^\nu \mu = \nu$, and a convex function g s.t. $T_\mu^\nu = \nabla g$ μ-a.e.
- $W_2^2(\mu, \nu) = \|I - T_\mu^\nu\|_{L_2(\mu)}^2 = \inf_{T \in L_2(\mu)} \int (x - T(x))^2 d\mu(x)$
- Also if $\nu \ll Leb$, then $T_\mu^\nu \circ T_\nu^\mu = I$ ν-a.e. and $T_\nu^\mu \circ T_\mu^\nu = I$ μ-a.e.

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:

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

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:
$\mathcal{T}_{\mu_t}\mathcal{P}_2(\mathbb{R}^d) \subset L^2(\mu_t) = \{f : \mathbb{R}^d \to \mathbb{R}^d, \int \|f(x)\|^2 d\mu_t(x) < \infty\}$.

Wasserstein gradient flows [Ambrosio et al., 2008]

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

The differential of $\mu \mapsto \mathcal{G}(\mu)$ 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 $\mu, \mu' \in \mathcal{P}_2(\mathbb{R}^d)$, $\mu' - \mu \in \mathcal{P}_2(\mathbb{R}^d)$:

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

Wasserstein gradient flows [Ambrosio et al., 2008]

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

The differential of $\mu \mapsto \mathcal{G}(\mu)$ 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 $\mu, \mu' \in \mathcal{P}_2(\mathbb{R}^d)$, $\mu' - \mu \in \mathcal{P}_2(\mathbb{R}^d)$:

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

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 $\mathcal{G}$.

Free energies

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

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

$$\text{Then : } \frac{\partial \mu_t}{\partial t} = div(\mu_t \nabla(H'(\mu_t) + V + W * \mu_t)).$$

Here, we consider

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{G}(\mu) := \underbrace{\int V d\mu}_{\mathcal{E}_V(\mu)} + \mathcal{H}(\mu).$$

We study an unbiased algorithm/time-discretization of the Wasserstein gradient flow of $\mathcal{G}$ to minimize this functional.

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**
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:

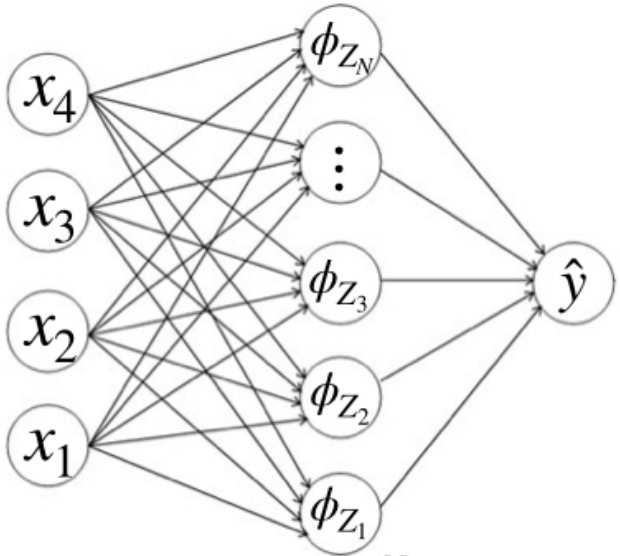
$$\int k(z, .) d\mu(z) - \int k(z, .) d\pi(z) : \mathbb{R}^d \to \mathbb{R}.$$

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

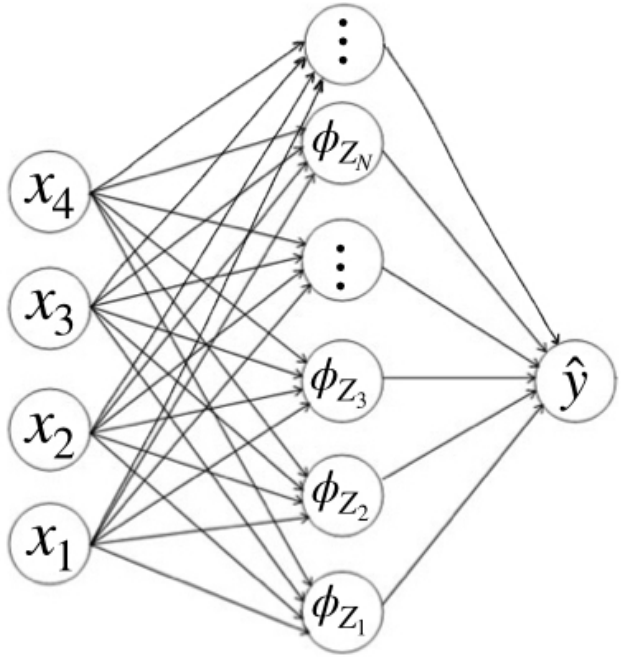
$$\int \nabla_2 k(z, .) d\mu(z) - \int \nabla_2 k(z, .) d\pi(z) : \mathbb{R}^d \to \mathbb{R}.$$

Example 2 : Regression with infinite width NN

$(x, y) \sim data$



$N \to \infty$



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

Minimization of the MMD : the well-specified case

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

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

Then :

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

$$\Updownarrow$$

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

$$\Updownarrow$$

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

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

$$\Updownarrow$$

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

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]

$$X_t = X_0 + \sqrt{2t}Z \tag{1}$$

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

Some time-discretizations of the KL gradient flow...

[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 (1)

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 :

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

$X_{n+1} = Y_{n+1} + \sqrt{2\gamma}\xi_n$ 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].

Unbiased time discretizations (or algorithms)

1. Forward :

$$\mu_{n+1} = (I - \gamma \nabla_W \mathrm{KL}(\mu_n|\pi))_{\#}\, \mu_n$$

2. Backward :

$$\mu_{n+1} = JKO_{\gamma\, \mathrm{KL}(.|\pi)}(\mu_n)$$

where $JKO_{\gamma\, \mathrm{KL}(.|\pi)}(\mu_n) = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\mathrm{argmin}}\ \mathrm{KL}(.|\pi)(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \mu_n).$

3. Forward-Backward :

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

It is unbiased because the backward method is the adjoint of the forward method, so the minimizer is conserved.

Outline

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

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 $\nu \ll Leb$, $t \in [0,1]$:

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

$T_{\mu^*}^{\nu}$ 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. $\forall\ (x, y) \in \mathbb{R}^d$,

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

Results : 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 $\lambda > 0$)

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 very close from the entropic-regularized OT problem, since:

$$\begin{aligned}
&\min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \gamma \mathcal{H}(\nu) + \frac{1}{2} W^2(\nu, \mu) \\
&\quad = \min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \min_{s \in \Gamma(\mu,\nu)} \gamma \mathcal{H}(\nu) + \frac{1}{2} \int \|x - y\|^2 ds(x, y) \\
&\quad = \min_{s : P_{1\#} s = \mu} \gamma \mathcal{H}(P_{2\#} s) + \frac{1}{2} \int \|x - y\|^2 ds(x, y)
\end{aligned}$$

where $P_1 : (x, y) \mapsto x$ and $P_2 : (x, y) \mapsto y$.

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:

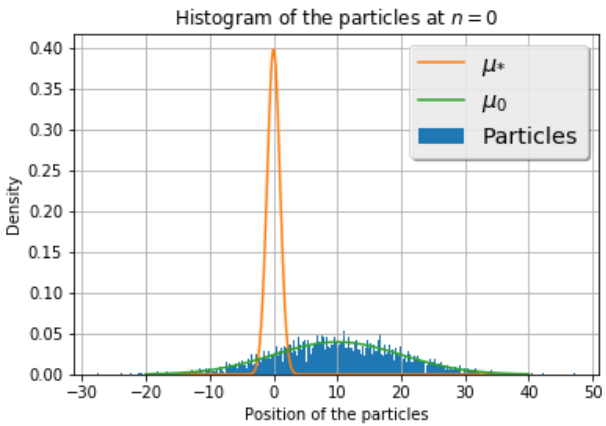
$$y_{n+1} = m + (I - \gamma\Sigma^{-1})(x_n - m)$$
$$x_{n+1} = m_{n+1} + (I - \gamma\Sigma_{n+1}^{-1})^{-1}(y_{n+1} - \mu_n)$$

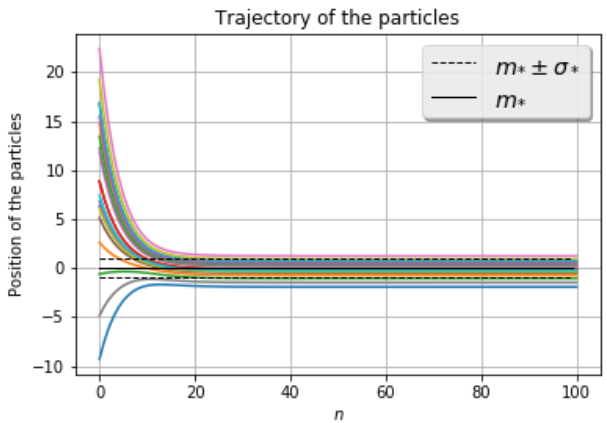
where

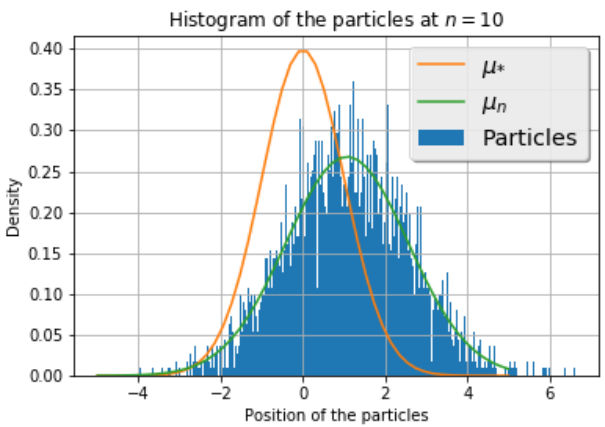
$$\mu_{n+1} = m + (I - \gamma\Sigma^{-1})(\mu_n - m)$$
$$\Sigma_{n+1}(I - \gamma\Sigma_{n+1}^{-1})^2 = \Sigma_n(I - \gamma\Sigma^{-1})^2$$

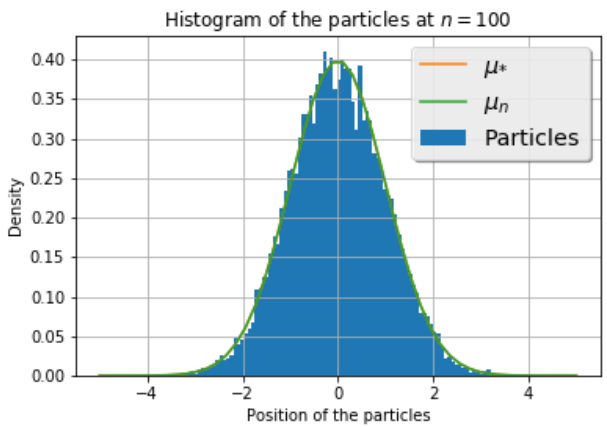
Experiments (d=1)

- $\pi = \mathcal{N}(0, 1)$ (hence $V(x) = 0.5x^2$ and $\lambda = 1$); $\mu_0 = \mathcal{N}(10, 100)$
- we use the closed-form particle implementation for the FB scheme









Linear rate (d=1000)

multi dimensional extension : $V(x) = 0.5\|x\|^2$, target $\mu^{*\otimes d}$ and initial distribution $\mu_0^{\otimes d}$

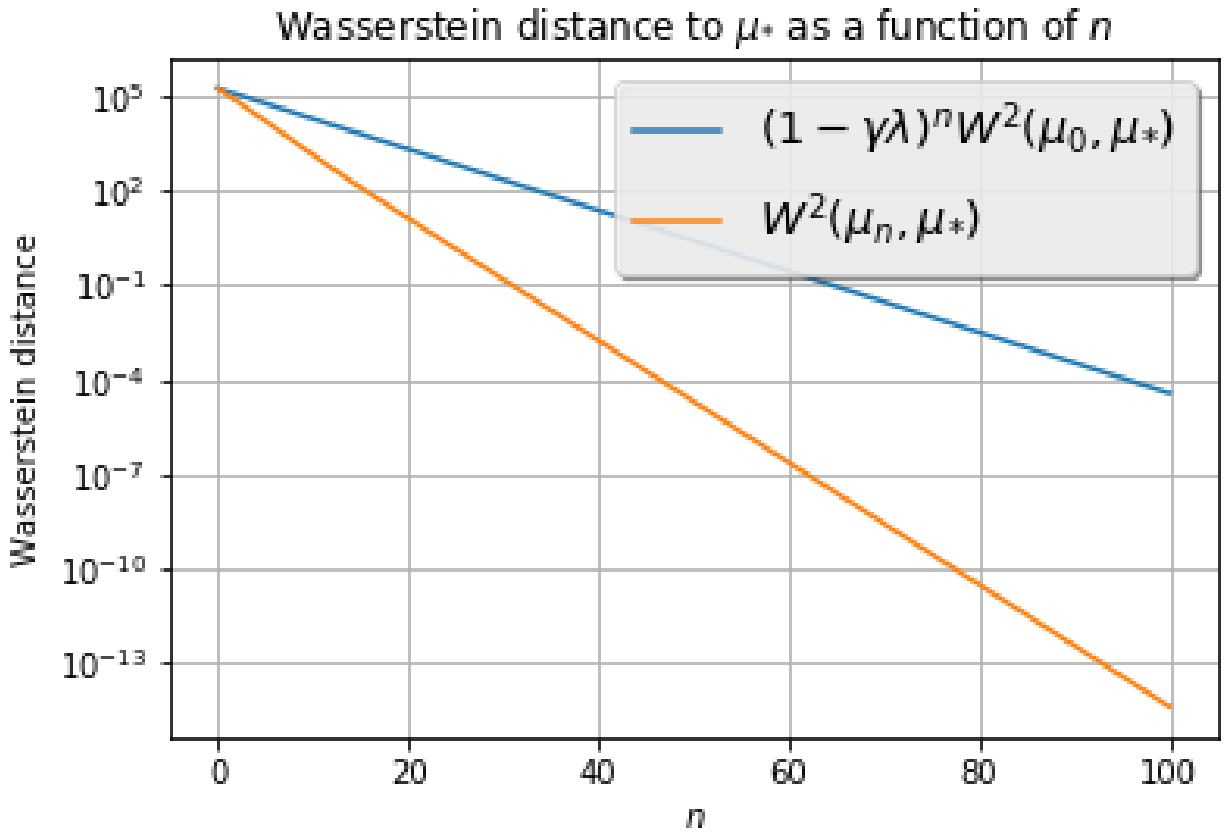


Figure: Linear convergence of μ_n to π in dimension $d = 1000$.

Contributions

- FB scheme is faster in nb of iterations compared to the Unadjusted Langevin algorithm (converges at rate $\mathcal{O}(1/\sqrt{n})$) at the cost of a higher iteration complexity.
- Our proof works for any functional $\mathcal{H}$ that is **convex along generalized geodesics**, and that works for entropies, but also for

$$\text{potential energies } \mathcal{H}(\mu) = \int F(x)\mu(x)dx$$

for F convex, or

$$\text{interaction energies } \mathcal{H}(\mu) = \int W(x,y)\mu(x)\mu(y)dxdy$$

for W convex.

Open questions

- The JKO of entropy deserves more investigation to find an efficient subroutine.
- Results in the non-convex case?

Thank you for listening !

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.


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


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

References II


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.


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

References III


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.

Wasserstein Gradient descent for the KL

Let $\mu_0 \in \mathcal{P}$. Gradient descent on $(\mathcal{P}, W_2)$ is written:

$$\mu_{n+1} = \left(I - \gamma \nabla \frac{\partial KL(\mu_n|\pi)}{\partial \mu} \right)_{\#} \mu_n$$

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

(Particle version) i.e. given $X_0 \sim \mu_0$,

$$X_{n+1} = X_n - \gamma \nabla \frac{\partial \, \mathrm{KL}(\mu_n|\pi)}{\partial \mu}(X_n) \sim \mu_{n+1}$$

Wasserstein Gradient descent for the KL

Let $\mu_0 \in \mathcal{P}$. Gradient descent on $(\mathcal{P}, W_2)$ is written:

$$\mu_{n+1} = \left(I - \gamma \nabla \frac{\partial KL(\mu_n|\pi)}{\partial \mu} \right)_{\#} \mu_n$$

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

(Particle version) i.e. given $X_0 \sim \mu_0$,

$$X_{n+1} = X_n - \gamma \nabla \frac{\partial \, \mathrm{KL}(\mu_n|\pi)}{\partial \mu}(X_n) \sim \mu_{n+1}$$

Problem: the W_2 gradient of $\mathrm{KL}(\cdot|\pi)$ at μ_n is the function $\nabla \log(\frac{\mu_n}{\pi})$. While $\nabla \log \pi$ is known, we do not know what μ_n is at each n, we only have X_{n+1}
$\Longrightarrow \nabla \log \mu_n$ has to be estimated from samples.

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

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

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

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

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

- $\mathcal{H}$ 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: applying this operator to the W_2 gradient of $KL(\cdot|\pi)$ leads to (if $\lim_{\|x\| \to \infty} k(x, \cdot)\pi(x) \to 0$)

$$P_\mu \nabla \log \left(\frac{\mu}{\pi} \right) (\cdot) = - \int [\nabla \log \pi(x) k(x, \cdot) + \nabla_x k(x, \cdot)] d\mu(x),$$

Stein Variational Gradient Descent (SVGD)

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 $\hat{\mu}_n = \frac{1}{N} \sum_{j=1}^{N} \delta_{X_n^j}$.

- "non parametric" VI, only depends on the choice of some kernel k
- uses a set of interacting particles to approximate π: `https://chi-feng.github.io/mcmc-demo/app.html?algorithm=HamiltonianMC&target=banana`