Wasserstein Gradient Flows for Machine Learning

Anna Korba [1] Adil Salim [2]

[1]CREST, ENSAE, Paris, France

[2]VCC, KAUST, Saudi Arabia

Second Symposium on Machine Learning and Dynamical Systems
Fields Institute
September 2020

Problem: Transport an initial probability distribution $\mu_0 \in \mathcal{P}$ to a target distribution $\mu^* \in \mathcal{P}$.

Applications : sampling for Bayesian inference, optimizing wide neural networks

Can be written as an optimization problem on $\mathcal{P}$, e.g.

$$\min_{\mu \in \mathcal{P}} KL(\mu|\mu^*) \tag{1}$$

$\Longrightarrow$ Wasserstein GF find a *continuous* path on the space of distributions (equipped with the Wasserstein geometry).

Different algorithms result from different time-space discretizations and dynamical systems.

Outline

Motivations

From $\mu_0 \in \mathcal{P}$ to $\mu^* \in \mathcal{P}$

Preliminaries on gradient flows

$\frac{\partial \mu_t}{\partial t} = div(\mu_t \nabla \frac{\partial \mathcal{F}(\mu_t)}{\partial \mu})$ where $\mathcal{F} : \mathcal{P} \to \mathbb{R} \cup \{+\infty\}$

Maximum Mean Discrepancy Gradient flow (Neurips 2019)

M. Arbel, A. Korba, A. Salim, A. Gretton

$\min_{\mu \in \mathcal{P}} \mathcal{F}(\mu)$, with $\mathcal{F}(\mu) = \frac{1}{2} MMD^2(\mu, \mu^*)$

Primal Dual interpretation of the Proximal Gradient Langevin algorithm (submitted)

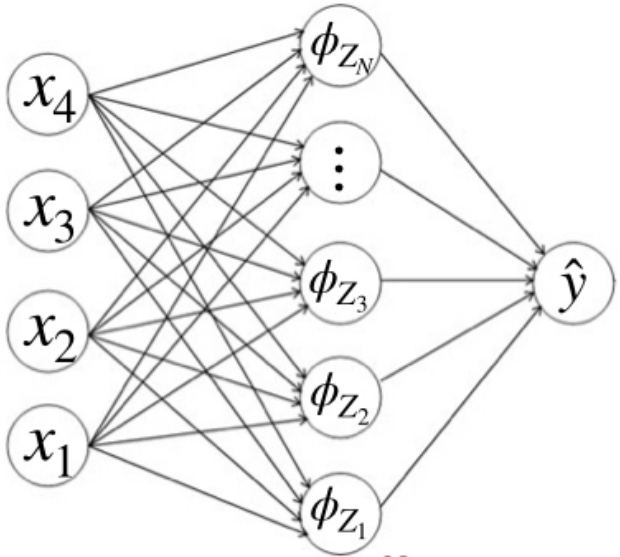
A. Salim, P. Richtarik

$\min_{\mu \in \mathcal{P}} \mathcal{F}(\mu)$, with $\mathcal{F}(\mu) = KL(\mu | \mu^*)$

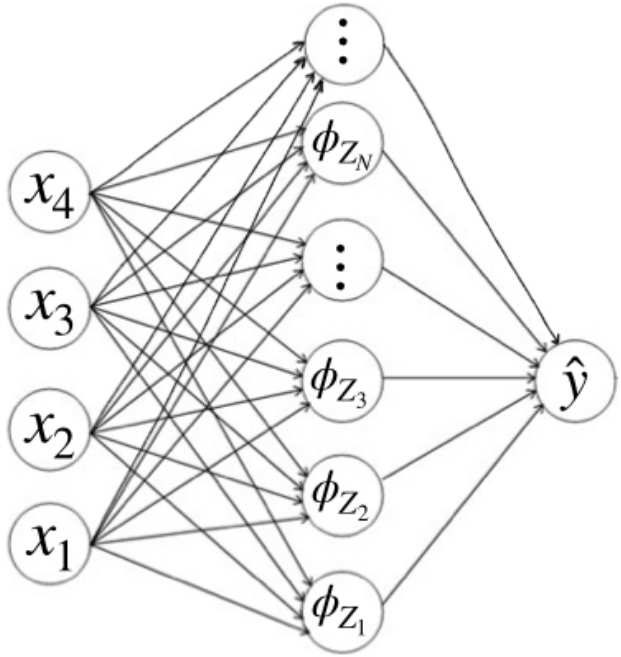
Conclusion and Research directions

Example 1 : Regression with infinite width NN

$(x, y) \sim data$



$N \to \infty$



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

Minimization of the MMD : the well-specified case

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

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

Then :

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

$$\Updownarrow$$

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

$$\Updownarrow$$

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

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

$$\Updownarrow$$

$$\min_{\mu\in\mathcal{P}} \textit{MMD}^2(\mu, \mu^*)$$

Example 2 : Bayesian statistics

- Let $\mathcal{D} = (x_i, y_i)_{i=1,\dots,N}$ observed data.
- Assume an underlying model parametrized by θ (e.g. $p(y|x,\theta)$ gaussian)
  $\Longrightarrow$ Likelihood: $p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(y_i|\theta, x_i)$
- The parameter $\theta \sim p$ the prior distribution.

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

How to sample from $\theta \mapsto p(\theta|\mathcal{D})$? ($Z$ unknown).

1. **MCMC methods** (LMC, HMC...)
   [Chopin et al., 2012], [Dalalyan, 2017]
2. **Sampling as optimization of the KL:**
   [Wibisono, 2018],[Alquier and Ridgway, 2017], [Zhang et al., 2018]

$$\min_{\mu \in \mathcal{P}} KL(\mu | \underbrace{p(\theta|\mathcal{D})}_{\mu^*})$$

Outline

Euclidean Gradient Flows

Let $V : \mathbb{R}^d \to \mathbb{R}$ smooth. The (Euclidean) Gradient Flow (GF) of V is given by the solution to

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

Continuous time version of gradient descent:

$$\frac{x_{n+1} - x_n}{\gamma} = -\nabla V(x_n)$$

Lyapunov functions for the GF

The GF tends to minimize V. Let x^* a minimizer of V. Denote $\mathcal{L}(t) = V(x(t)) - V(x^*)$.

$$\mathcal{L}'(t) = \langle x'(t), \nabla V(x(t)) \rangle = -\|\nabla V(x(t))\|^2 \leq 0,$$

therefore $V(x(t)) \searrow$. Moreover,

$$\frac{1}{T} \int_0^T \|\nabla V(x(t))\|^2 dt \leq \frac{V(x(0)) - V(x^*)}{T}.$$

Lyapunov functions for the GF

Denote $\mathcal{L}_c(t) = \|x(t) - x^*\|^2$. *Assume V convex.*

$$\mathcal{L}_c'(t) \leq -2(V(x(t)) - V(x^*)) \leq 0,$$

therefore $\|x(t) - x^*\|^2 \searrow$. Moreover,

$$V(x(T)) - V(x^*) \leq \frac{1}{T}\int_0^T (V(x(t)) - V(x^*))dt \leq \frac{\|x(0) - x^*\|^2}{2T}.$$

Further use of these two Lyapunov functions

1. Discrete versions $\mathcal{L}_n, \mathcal{L}_{c,n}$ of $\mathcal{L}(t), \mathcal{L}_c(t)$ can be used to prove $\mathcal{O}(1/n)$ convergence rates of gradient descent

2. $\mathcal{L}_n, \mathcal{L}_{c,n}, \mathcal{L}(t), \mathcal{L}_c(t)$ can be used to prove linear convergence (exponentially fast convergence) of gradient descent/ the gradient flow if V is strongly convex.

A dual point of view

Consider the gradient flow

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

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.

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

and

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

Therefore,

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

Wasserstein gradient flows

Let $\mathcal{P} = \{\mu \in \mathcal{P}(\mathcal{X}),\ \int \|x\|^2 d\mu(x) < \infty\}$[1] and $\mathcal{F} : \mathcal{P} \to \mathbb{R} \cup \{+\infty\}$ a regular functional.

Then $\mu : [0, \infty] \to \mathcal{P}, t \mapsto \mu_t$ satisfies a Wasserstein gradient flow of $\mathcal{F}$ if distributionnally:

$$\frac{\partial \mu_t}{\partial t} = div\left(\mu_t \nabla \frac{\partial \mathcal{F}(\mu_t)}{\partial \mu_t}\right),$$

where $\frac{\partial \mathcal{F}(\mu)}{\partial \mu} : \mathcal{X} \to \mathbb{R}$ is the differential of $\mu \mapsto \mathcal{F}(\mu)$ at μ and $\nabla_W \mathcal{F}(\mu) := \nabla \frac{\partial \mathcal{F}(\mu)}{\partial \mu} \in L^2(\mu)$[2] is called the Wasserstein gradient of $\mathcal{F}$.

---

[1] $\mathcal{X} = \mathbb{R}^d$

[2] Denote $L^2(\mu) = \{f : \mathcal{X} \to \mathcal{X},\ \int \|f(x)\|^2 d\mu(x) < \infty\}$ and $\langle f, g \rangle_\mu$ (resp. $\|f\|_\mu$) the associated inner product (resp. norm).

Free energies

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

$$\mathcal{F}(\mu) = \underbrace{\int U(\mu(x))\mu(x)dx}_{\text{internal potential } \mathcal{U}} + \underbrace{\int V(x)\mu(x)dx}_{\text{external potential } \mathcal{E}_V} + \underbrace{\int W(x,y)\mu(x)\mu(y)dxdy}_{\text{interaction energy } \mathcal{W}}$$

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

We recover the Euclidean GF if $U \equiv 0$, $W \equiv 0$.

Some free energies in Machine Learning

The **relative entropy** $\mathcal{F}(\mu) = KL(\mu|\mu^*)$ can be written:

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

$U(s) = s\log(s),\ V(x) = -log(\mu^*(x)),\ C = \mathcal{U}(\mu^*) + \mathcal{E}_V(\mu^*).$

Some free energies in Machine Learning

The **relative entropy** $\mathcal{F}(\mu) = KL(\mu|\mu^*)$ can be written:

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

$U(s) = s\log(s),\ V(x) = -log(\mu^*(x)),\ C = \mathcal{U}(\mu^*) + \mathcal{E}_V(\mu^*).$

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

Some free energies in Machine Learning

The **relative entropy** $\mathcal{F}(\mu) = KL(\mu|\mu^*)$ can be written:

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

$U(s) = s\log(s),\ V(x) = -log(\mu^*(x)),\ C = \mathcal{U}(\mu^*) + \mathcal{E}_V(\mu^*).$

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

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

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

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

Some free energies in Machine Learning

The **relative entropy** $\mathcal{F}(\mu) = KL(\mu|\mu^*)$ can be written:

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

$U(s) = s\log(s),\ V(x) = -log(\mu^*(x)),\ C = \mathcal{U}(\mu^*) + \mathcal{E}_V(\mu^*).$

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

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

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

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

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

Underlying structure

2-Wasserstein distance:

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

where $\Gamma(\nu,\mu)$ is the set of possible couplings between ν and μ.

Lyapunov functions for the Wasserstein GF

The Wasserstein GF tends to minimize $\mathcal{F}$. Let μ^* a minimizer of $\mathcal{F}$.

Denote $\mathcal{L}(t) = \mathcal{F}(\mu_t) - \mathcal{F}(\mu^*)$.

$$\mathcal{L}'(t) = \langle V_t, \nabla_W \mathcal{F}(\mu_t) \rangle_{\mu_t} = -\|\nabla_W \mathcal{F}(\mu_t)\|^2_{\mu_t} \leq 0,$$

therefore $\mathcal{F}(\mu_t) \searrow$. Moreover,

$$\frac{1}{T} \int_0^T \|\nabla_W \mathcal{F}(\mu_t)\|^2_{\mu_t} dt \leq \frac{\mathcal{F}(\mu_0) - \mathcal{F}(\mu^*)}{T}.$$

Lyapunov functions for the Wasserstein GF

Denote $\mathcal{L}_c(t) = W_2^2(\mu_t, \mu^*)$. *Assume $\mathcal{F}$ geodesically convex.*

$$\mathcal{L}_c'(t) \leq -2(\mathcal{F}(\mu_t) - \mathcal{F}(\mu^*)) \leq 0,$$

therefore $W_2^2(\mu_t, \mu^*) \searrow$. Moreover,

$$\mathcal{F}(\mu_t) - \mathcal{F}(\mu^*) \leq \frac{1}{T}\int_0^T \mathcal{F}(\mu_t) - \mathcal{F}(\mu^*)dt \leq \frac{W_2^2(\mu_0, \mu^*)}{2T}.$$

Our approach

Similarly to the transition

$$\textbf{\textit{Euclidean gradient flow}} \longrightarrow \textbf{\textit{gradient descent}},$$

we shall use the Wasserstein gradient flow point of view to analyze algorithms that can be seen as time *discretized Wasserstein gradient flows*.

If convexity is involved, we shall use the Lyapunov function $\mathcal{L}_c$, otherwise we use $\mathcal{L}$.

Outline

Motivations

From $\mu_0 \in \mathcal{P}$ to $\mu^* \in \mathcal{P}$

Preliminaries on gradient flows

$\frac{\partial \mu_t}{\partial t} = div(\mu_t \nabla \frac{\partial \mathcal{F}(\mu_t)}{\partial \mu})$ where $\mathcal{F} : \mathcal{P} \to \mathbb{R} \cup \{+\infty\}$

Maximum Mean Discrepancy Gradient flow (Neurips 2019)

M. Arbel, A. Korba, A. Salim, A. Gretton

$\min_{\mu \in \mathcal{P}} \mathcal{F}(\mu)$, with $\mathcal{F}(\mu) = \frac{1}{2} MMD^2(\mu, \mu^*)$

Primal Dual interpretation of the Proximal Gradient Langevin algorithm (submitted)

A. Salim, P. Richtarik

$\min_{\mu \in \mathcal{P}} \mathcal{F}(\mu)$, with $\mathcal{F}(\mu) = KL(\mu|\mu^*)$

Conclusion and Research directions

Maximum Mean Discrepancy

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

$$k(z, z') = \langle \phi(z), \phi(z') \rangle_{\mathcal{H}}, \quad \phi : \mathcal{Z} \to \mathcal{H}$$

- $\mathcal{H}$ its RKHS (Reproducing Kernel Hilbert Space).

Recall: $\mathcal{H}$ is a Hilbert space with inner product $\langle ., . \rangle_{\mathcal{H}}$ and norm $\|.\|_{\mathcal{H}}$. It satisfies the reproducing property:

$$\forall \quad f \in \mathcal{H},\ z \in \mathcal{Z}, \quad f(z) = \langle f, k(z, .) \rangle_{\mathcal{H}}$$

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

Maximum Mean Discrepancy

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

$$k(z, z') = \langle \phi(z), \phi(z') \rangle_{\mathcal{H}}, \quad \phi : \mathcal{Z} \to \mathcal{H}$$

- $\mathcal{H}$ its RKHS (Reproducing Kernel Hilbert Space).

Recall: $\mathcal{H}$ is a Hilbert space with inner product $\langle ., . \rangle_{\mathcal{H}}$ and norm $\|.\|_{\mathcal{H}}$. It satisfies the reproducing property:

$$\forall \quad f \in \mathcal{H},\ z \in \mathcal{Z}, \quad f(z) = \langle f, k(z, .) \rangle_{\mathcal{H}}$$

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

Maximum Mean Discrepancy ([Gretton et al., 2012]) defines a distance on $\mathcal{P}$ (probability distributions on $\mathcal{Z}$):

$$MMD(\mu, \nu) = \|f_{\mu,\nu}\|_{\mathcal{H}}, \text{ where}$$

$$\underbrace{f_{\mu,\nu}(.) = \int k(z, .) d\mu(z) - \int k(z, .) d\nu(z)}_{\textbf{"witness function"}}$$

MMD functional

For a target distribution μ^* (fixed), for any $\mu \in \mathcal{P}$:

$$\begin{aligned}\mathcal{F}(\mu) &= \frac{1}{2} MMD^2(\mu, \mu^*) \\ &= \frac{1}{2} \| f_{\mu,\mu^*} \|^2_{\mathcal{H}} \\ &= \frac{1}{2} \int k(z, z') d\mu(z) d\mu(z') + \frac{1}{2} \int k(z, z') d\mu^*(z) d\mu^*(z') \\ &\quad - \int k(z, z') d\mu(z) d\mu^*(z')\end{aligned}$$

Proof : use the reproducing property with
$f_{\mu,\mu^*}(.) = \int k(z, .) d\mu(z) - \int k(z, .) d\mu^*(z)$

Appear as a loss when optimizing some large neural networks.

We consider

$$\min_{\nu \in \mathcal{P}} \mathcal{F}(\mu) \text{ where } \mathcal{F}(\mu) = \frac{1}{2} MMD^2(\mu, \mu^*)$$

- Gradient descent dynamics in this setting takes the form of a PDE (gradient flow on $\mathcal{P}$)

$$\frac{\partial \mu_t}{\partial t} = div\left(\mu_t \nabla \frac{\partial \mathcal{F}(\mu_t)}{\partial t}\right) = div(\mu_t \nabla f_{\mu_t, \mu^*})$$

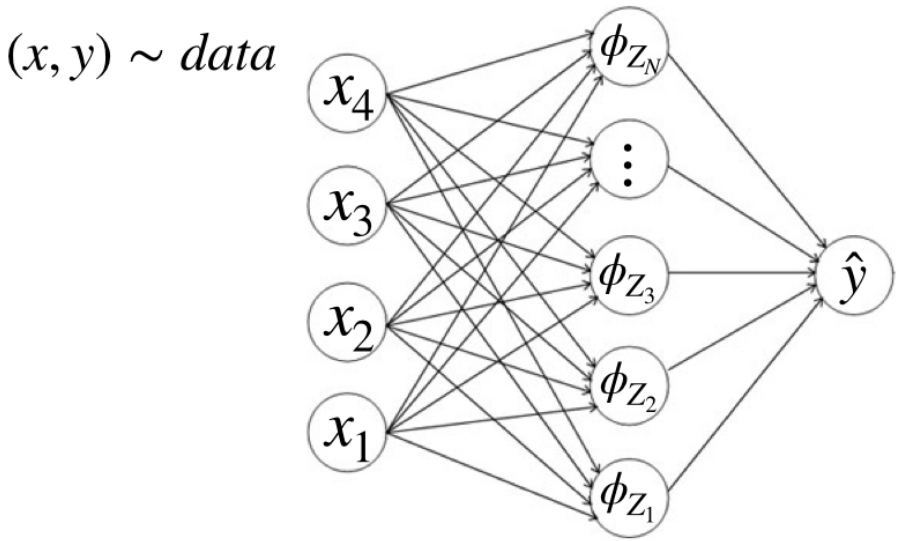
  where $\nabla_{Z_t} f_{\mu_t, \mu^*} = \int \nabla k(Z, Z_t) d\mu_t(Z) - \int \nabla k(Z, Z_t) d\mu^*(Z)$.
- Density of particles following a Mc-Kean Vlasov dynamic :

$$\frac{dZ_t}{dt} = -\nabla_{Z_t} f_{\mu_t, \mu^*}(Z_t), \qquad Z_t \sim \mu_t$$

Illustration : Student-Teacher network

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

- the output of the Teacher network is deterministic and given by $y = \int \phi_Z(x) d\mu^*(Z)$ where $\mu^* = \frac{1}{M}\sum_{m=1}^{M} \delta_{U^m}$
- Student network parametrized by $\mu_0 = \frac{1}{N}\sum_{n=1}^{N} \delta_{Z_0^n}$ tries to learn the mapping $x \mapsto \int \phi_Z(x) d\mu^*(Z)$.



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

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

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

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

$$z_{t+1}^n = z_t^n - \gamma \underbrace{\left(\frac{1}{N} \sum_{n'=1}^{N} \nabla_2 k(z_t^{n'}, z_t^n) - \frac{1}{M} \sum_{m=1}^{M} \nabla_2 k(u^m, z_t^n) \right)}_{\nabla f_{\mu^*, \mu_t}(z_t^n)}$$

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

Convergence of the MMD GF

$\mathcal{F}(\mu_t)$ along the MMD flow $\frac{\partial \mu_t}{\partial t} = div(\mu_t \nabla \frac{\partial \mathcal{F}(\mu_t)}{\partial \mu})$?

We know that $\frac{d\mathcal{F}(\mu_t)}{dt} = -\|\nabla_W \mathcal{F}(\mu_t)\|^2_{\mu_t}$. Do we have

$$\frac{d\mathcal{F}(\mu_t)}{dt} = -\|\nabla_W \mathcal{F}(\mu_t)\|^2_{\mu_t} \leq -C\mathcal{F}(\mu_t)^2?$$

Yes if $\|\mu_t - \mu^*\|_{\dot{H}^{-1}(\mu_t)} \leq C$ for all t, where $\|\cdot\|_{\dot{H}^{-1}(\mu_t)}$ is the weighted negative Sobolev distance (linearizes W_2)

$$\|\mu_t - \mu^*\|_{\dot{H}^{-1}(\mu_t)} = \sup_{g,\ \mathbb{E}_{Z\sim\mu_t}[\|\nabla g(Z)\|^2]\leq 1} |\mathbb{E}_{Z\sim\mu_t}[g(Z)] - \mathbb{E}_{U\sim\mu^*}[g(U)]|$$

This results in $\mathcal{F}(\mu_t) = \mathcal{O}(\frac{1}{t})$ (also true in discrete time)

Problem: Depends on the whole sequence μ_t; Hard to verify in general [Peyre, 2018]; and we've seen failure cases in practice.

Convergence issues

- The condition we exhibited for global convergence may not hold and $(\mathcal{F}(\mu_t))_t$ might be stuck at a local minima.

$$\frac{d\mathcal{F}(\mu_t)}{dt} = -\int \|\nabla f_{\mu_t,\mu^*}(x)\|^2 d\mu_t(x) \text{ at equilibrium}$$

$$\Longrightarrow \int \|\nabla f_{\mu^\infty,\mu^*}(x)\|^2 d\mu^\infty(x) = 0$$

If μ^∞ positive everywhere this implies $f_{\mu^\infty,\mu^*} = cte = 0$[3]

But μ^∞ might be singular...

- **Idea :** Evaluate $\nabla f_{\mu_t,\mu^*}$ outside of the support of μ_t to get a better signal!

[3] as soon as the RKHS $\mathcal{H}$ does not contain non-zero constant functions.

Noise Injection

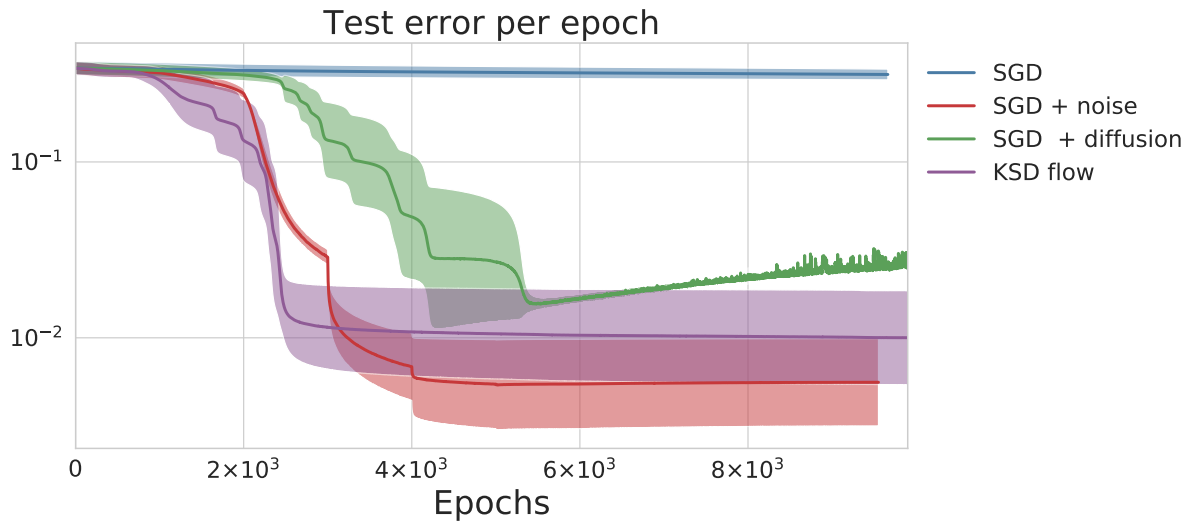
At each iteration n, sample $\xi_n \sim \mathcal{N}(0, 1)$ and β_n is the noise level:

$$Z_{n+1} = Z_n - \gamma \nabla f_{\mu_n, \mu^*}(Z_n + \beta_n \xi_n)$$

Different from adding noise outside ("diffusion")

$$Z_{n+1} = Z_n - \gamma \nabla f_{\mu_n, \mu^*}(Z_n) + \beta_n \xi_n$$

which corresponds to an entropic regularization of the loss.



Contributions and openings

- Provided a convergence criterion for the Wasserstein gradient flow of the MMD.
- Proposed a pertubation of the dynamics with a noise injection and showed it effectiveness on simple examples.
- new insights for training large neural networks.

Openings:

- Control of the weighted negative Sobolev norm?
- Stronger guarantees for the convergence for the noise injection algorithm.

Outline

Sampling as optimization of the KL

Problem : sampling from a distribution $\mu^* \propto exp(-V)$ where $V : \mathcal{X} \to \mathbb{R}$ convex[4].

$$\min_{\mu \in \mathcal{P}} \mathcal{F}(\mu)$$

where $\mathcal{F}(\mu) = KL(\mu|\mu^*)$

$$KL(\mu|\mu^*) = \int \log\left(\frac{d\mu}{d\mu^*}(x)\right) d\mu(x) = \mathcal{E}_V(\mu) + \mathcal{U}(\mu),$$

if $\mu \ll \mu^*$ with density $\frac{d\mu}{d\mu^*}$, and $KL(\mu|\mu^*) = +\infty$ else.

[4] $\mathcal{X} = \mathbb{R}^d$

Sampling as optimization of the KL

Problem : sampling from a distribution $\mu^* \propto exp(-V)$ where $V : \mathcal{X} \to \mathbb{R}$ convex[4].

$$\min_{\mu \in \mathcal{P}} \mathcal{F}(\mu)$$

where $\mathcal{F}(\mu) = KL(\mu|\mu^*)$

$$KL(\mu|\mu^*) = \int \log\left(\frac{d\mu}{d\mu^*}(x)\right) d\mu(x) = \mathcal{E}_V(\mu) + \mathcal{U}(\mu),$$

if $\mu \ll \mu^*$ with density $\frac{d\mu}{d\mu^*}$, and $KL(\mu|\mu^*) = +\infty$ else.

Assume V smooth.

Langevin Monte Carlo (LMC) is a sampling algorithm:

$$x_{n+1} = x_n - \gamma \nabla V(x_n) + \sqrt{2\gamma}\xi_{n+1},$$

where $\gamma > 0$ and $(\xi_n)_n$ i.i.d. r.v. with standard Gaussian distribution [Dalalyan, 2017, Durmus et al., 2019, Wibisono, 2018, Bernton, 2018].

[4] $\mathcal{X} = \mathbb{R}^d$

Proximal Gradient Langevin Algorithm

Assume $V = F + G$ where F is smooth λ-strongly convex and G nonsmooth convex[5].
Proximal Gradient Langevin Algorithm (PGLA) is a sampling algorithm:

$$x_{n+1} = \mathrm{prox}_{\gamma G}\left(x_n - \gamma \nabla F(x_n) + \sqrt{2\gamma}\xi_{n+1}\right),$$

where $\gamma > 0$, $(\xi_n)_n$ i.i.d. r.v. with standard Gaussian distribution and

$$\mathrm{prox}_G(x) := \underset{y \in \mathcal{X}}{\mathrm{argmin}} \frac{1}{2}\|x - y\|^2 + G(y). \tag{2}$$

If $G = \iota_C$[6] where C convex set, $\mathrm{prox}_G(x) = \mathrm{proj}_C(x)$.

More generally, $\mathrm{prox}_G(x) \in \mathrm{dom}(G)$, hence $x_{n+1} \in \mathrm{dom}(G)$.

[5] *i.e.* G convex, l.s.c., proper
[6] *i.e.* $G(x) = 0$ if $x \in C$ and $G(x) = +\infty$ else.

Non-asymptotic analysis of PGLA

1. Constrained sampling: $G = \iota_C$ where C convex body [Bubeck et al., 2018].
   In this case, $\mathrm{Supp}(\mu^*) = \mathrm{dom}(G) = C$.

   PGLA is called Projected Langevin algorithm:
   $x_{n+1} = \mathrm{proj}_C\left(x_n - \gamma \nabla F(x_n) + \sqrt{2\gamma}\xi_{n+1}\right)$.

   Complexity: $n = \mathcal{O}(1/\varepsilon^{12})$ in Total Variation distance[7].
2. G Lipschitz continuous [Durmus et al., 2019].
   In this case, $\mathrm{Supp}(\mu^*) = \mathcal{X}$.

   Complexity: $n = \mathcal{O}(1/\varepsilon^2)$ in 2-Wasserstein distance.
3. G l.s.c. proper and μ^* locally Sobolev 1,1 (**This work**).
   In this case, $\mathrm{Supp}(\mu^*) = \mathrm{dom}(G)$.

   Complexity: $n = \mathcal{O}(1/\varepsilon^2)$ in 2-Wasserstein distance.

[7] This result also holds if $\lambda = 0$.

Proximal Gradient algorithm

1. Consider the problem

$$\min_{x \in \mathcal{X}} F(x) + G(x). \tag{3}$$

2. The Forward Backward (FB) Euler discretization of GF is

$$x_{n+1} = \mathrm{prox}_{\gamma G}(x_n - \gamma \nabla F(x_n)), \quad n \geq 0. \tag{4}$$

3. The FB discretization satisfies (for γ small enough):
   $x^* \in \operatorname{argmin} V, \forall n \geq 0,$

   $$\|x_{n+1} - x^*\|^2 - \|x_n - x^*\|^2 \leq -2\gamma\,(V(x_{n+1}) - V(x^*)) - \lambda\gamma\|x_n - x^*\|^2.$$

4. The FB solves (3): $\|x_n - x^*\|^2 \leq (1 - \gamma\lambda)^n \|x_0 - x^*\|^2$.

Proximal Gradient algorithm, Revisited

1. Consider the problem

$$\min_{x \in \mathcal{X}} F(x) + G(x).$$

2. The Forward Backward (FB) Euler discretization of GF is

$$x_{n+1} = \text{prox}_{\gamma G}(x_n - \gamma \nabla F(x_n)), \quad n \geq 0. \tag{5}$$

3. The FB discretization satisfies (for γ small enough):
$x^* \in \text{argmin } V, y^* \in \partial G(x^*), \forall n \geq 0,$

$$\|x_{n+1} - x^*\|^2 - \|x_n - x^*\|^2 \leq -2\gamma \left(\mathcal{L}(x_{n+1}, y^*) - \mathcal{L}(x^*, y_{n+1})\right) - \lambda\gamma \|x_n - x^*\|^2,$$

where $\mathcal{L}(x, y) = F(x) - G^*(y) + \langle x, y \rangle$.

4. The FB solves (3): $\|x_n - x^*\|^2 \leq (1 - \gamma\lambda)^n \|x_0 - x^*\|^2$.

Proximal Gradient Langevin algorithm

1. Consider the problem

$$\min_{\mu\in\mathcal{P}} \mathcal{F}(\mu) = \mathcal{E}_V(\mu) + \mathcal{U}(\mu) = \mathcal{E}_F(\mu) + \mathcal{E}_G(\mu) + \mathcal{U}(\mu). \quad (6)$$

2. The Forward Flow Backward (FFB) discretization of GF [Wibisono, 2018] is

$$x_{n+1} = \mathrm{prox}_{\gamma G}(x_n - \gamma\nabla F(x_n) + \sqrt{2\gamma}\xi_{n+1}), \quad n \geq 0. \quad (7)$$

3. The FFB discretization satisfies (γ small enough) [Durmus et al., 2018] (**assumes G Lipschitz**): $\forall n \geq 0$,

$$W_2^2(\mu_{n+1},\mu^*) - W_2^2(\mu_n,\mu^*) \leq -2\gamma KL(\tilde{\mu_n}|\mu^*) - \lambda\gamma W_2^2(\mu_n,\mu^*) + \gamma^2 C.$$

4. $W_2^2(\mu_n,\mu^*) \leq (1-\gamma\lambda)^n W_2^2(\mu_0,\mu^*) + \gamma C$.
5. Complexity to obtain $W(\mu_n,\mu^*) \leq \varepsilon$: $n = \mathcal{O}(1/\varepsilon^2)$.

Proximal Gradient Langevin algorithm, Revisited

1. Consider the problem

$$\min_{\mu\in\mathcal{P}} \mathcal{F}(\mu) = \mathcal{E}_V(\mu) + \mathcal{U}(\mu) = \mathcal{E}_F(\mu) + \mathcal{E}_G(\mu) + \mathcal{U}(\mu). \quad (8)$$

2. The Forward Flow Backward (FFB) discretization of GF is

$$x_{n+1} = \mathrm{prox}_{\gamma G}(x_n - \gamma\nabla F(x_n) + \sqrt{2\gamma}W_{n+1}), \quad n \geq 0. \quad (9)$$

3. The FFB discretization satisfies (γ small enough) (**This work**) (**does not assume G Lipschitz**): $\forall n \geq 0$,

$$W_2^2(\mu_{n+1}, \mu^*) - W_2^2(\mu_n, \mu^*) \leq -2\gamma \underbrace{(\mathscr{L}(\tilde{\mu}_n, y^*) - \mathscr{L}(\mu^*, y_{n+1}))}_{\geq 0} - \lambda\gamma W_2^2(\mu_n, \mu^*) + \gamma^2 C,$$

where $\mathscr{L}(\mu, y) = \mathcal{E}_F(\mu) + \mathcal{H}(\mu) - \mathcal{E}_{G^*}(\mu) + \mathbb{E}\langle x, y\rangle$, and $x = T_{\mu^*}^{\mu}(x^*)$.

4. $W_2^2(\mu_n, \mu^*) \leq (1-\gamma\lambda)^n W_2^2(\mu_0, \mu^*) + \gamma C$.
5. Complexity to obtain $W(\mu_n, \nu^*) \leq \varepsilon$: $n = \mathcal{O}(1/\varepsilon^2)$.

Outline

Advertisement

1. Non asymptotic analysis of SVGD: A sampling algorithm that iteratively transports a set of particles to a target distribution [Korba et al., 2020].
2. Non asymptotic analysis of the Wasserstein proximal gradient algorithm: a proximal gradient algorithm to minimize $\mathcal{F}(\mu) = \mathcal{E}_F(\mu) + \mathcal{G}(\mu)$ where F smooth convex and $\mathcal{G}$ geodesically convex nonsmooth [Salim et al., 2020].

Conclusion and Research directions

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

$$\min_{\mu \in \mathcal{P}} \mathcal{F}(\mu), \qquad \text{where } \mathcal{F}(\mu) = d(\mu, \mu^*)$$

- sampling
- optimizing wide NN
- many others : generative modelling [Chu et al., 2019], online learning [Boursier and Perchet, 2019], barycenters of distributions [Cuturi and Doucet, 2014]...

Many ideas from optimization can be useful in this setting (perturbation of dynamics, adapted discretizations...)

Conclusion and Research directions

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

$$\min_{\mu \in \mathcal{P}} \mathcal{F}(\mu), \qquad \text{where } \mathcal{F}(\mu) = d(\mu, \mu^*)$$

- sampling
- optimizing wide NN
- many others : generative modelling [Chu et al., 2019], online learning [Boursier and Perchet, 2019], barycenters of distributions [Cuturi and Doucet, 2014]...

Many ideas from optimization can be useful in this setting (perturbation of dynamics, adapted discretizations...)

Thank you!

References I


Alquier, P. and Ridgway, J. (2017).
Concentration of tempered posteriors and of their variational approximations.
arXiv preprint arXiv:1706.09293.


Bernton, E. (2018).
Langevin monte carlo and jko splitting.
arXiv preprint arXiv:1802.08671.


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


Bubeck, S., Eldan, R., and Lehec, J. (2018).
Sampling from a log-concave distribution with projected langevin monte carlo.
Discrete & Computational Geometry, 59(4):757–783.


Chizat, L. and Bach, F. (2018).
On the global convergence of gradient descent for over-parameterized models using optimal transport.
NIPS.


Chopin, N., Lelièvre, T., and Stoltz, G. (2012).
Free energy methods for Bayesian inference: efficient exploration of univariate Gaussian mixture posteriors.
Stat. Comput., 22(4):897–916.

References III


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.


Dalalyan, A. S. (2017).
Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent.
arXiv preprint arXiv:1704.04752.


Durmus, A., Eberle, A., Guillin, A., and Zimmer, R. (2018).
An elementary approach to uniform in time propagation of chaos.
arXiv preprint arXiv:1805.11387.

References IV


Durmus, A., Majewski, S., and Miasojedow, B. (2019).
Analysis of langevin monte carlo via convex optimization.
Journal of Machine Learning Research, 20(73):1–46.


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


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.

References V


Peyre, R. (2018).
Comparison between w2 distance and h1 norm, and localization of wasserstein distance.
ESAIM: Control, Optimisation and Calculus of Variations, 24(4):1489–1501.


Rotskoff, G., Jelassi, S., Bruna, J., and Vanden-Eijnden, E. (2019).
Global convergence of neuron birth-death dynamics.
In *ICML*.


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

References VI


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


Zhang, C., Butepage, J., Kjellstrom, H., and Mandt, S. (2018).
Advances in variational inference.
IEEE transactions on pattern analysis and machine intelligence.

The sample-based approximate scheme

How can we simulate

$$Z_{n+1} = Z_n - \gamma \nabla f_{\mu,\mu_n}(Z_n + \beta_n U_n), \quad n \geq 0?$$

It depends on:

- the current distribution $\nu_n \Longrightarrow$ approximate it by the empirical distribution of a system of N interacting particles
- the target distribution $\mu \Longrightarrow$ replace it by the empirical distribution of the M samples that we have access to ($\widehat{\mu}$)

The sample-based approximate scheme

How can we simulate

$$Z_{n+1} = Z_n - \gamma \nabla f_{\mu,\mu_n}(Z_n + \beta_n U_n), \quad n \geq 0?$$

It depends on:

- the current distribution $\nu_n \Longrightarrow$ approximate it by the empirical distribution of a system of N interacting particles
- the target distribution $\mu \Longrightarrow$ replace it by the empirical distribution of the M samples that we have access to ($\widehat{\mu}$)

$\Longrightarrow$ **create a system of interacting particles**

$$\widehat{\nu}_{n+1} \begin{cases} Z^1_{n+1} = Z^1_n - \gamma \nabla f_{\widehat{\mu},\widehat{\nu}_n}(Z^1_n + \beta_n U^1_n) \\ \cdots \\ Z^N_{n+1} = Z^N_n - \gamma \nabla f_{\widehat{\mu},\widehat{\nu}_n}(Z^N_n + \beta_n U^N_n) \end{cases}$$

Theoretical guarantees

(Propagation of chaos type of result)

Theorem

Let $n \geq 0$ and $T > 0$. Let ν_n and $\hat{\nu}_n$ defined by the (theoretical) Euler-scheme and the practical algorithm. Suppose $\|\nabla k\|_{Lip} = L$ and that $\beta_n < B$ for all n, for some $B > 0$. Then for any $\frac{T}{\gamma} \geq n$:

$$\mathbb{E}[W_2(\hat{\nu}_n, \nu_n)] \leq \frac{C_1(\nu_0, B, T, L)}{\sqrt{N}} + \frac{C_2(\mu, T, L)}{\sqrt{M}}$$

where N is the number of interacting particles and M is the number of samples from the target distribution.

MMD flow and related work

- [Rotskoff et al., 2019],[Chizat and Bach, 2018] identify the same pb that when the norm of the vector field is zero, i.e. for a fixed point of the GF, if the support of the density of the particles is too small, it might not be a global minimizer
- that's why [Chizat and Bach, 2018] needs the initial distribution to be supported everywhere; and gradient flow dynamics avoid spurious local minimas (ie they converge to the global minimizer) under appropriate conditions on the functions activations
- our global convergence criterion does not depend on the activation functions!
- also, we propose a perturbed dynamic ($\neq$ [Rotskoff et al., 2019]) and a discrete-time analysis

Neural tangent kernel

- common point : NN in the infinite width regime
- $k(x, x') = \mathbb{E}_{Z\sim\nu}[\frac{\partial\phi(Z,x)}{\partial Z}^T \frac{\partial\phi(Z,x')}{\partial\partial Z}]$ kernel between data points
- the time varying kernel $(\nu(t))$ is actually close to a deterministic one (NTK, ie the kernel at ν_0)
- A properly randomly initialized sufficiently wide deep neural network trained by gradient descent with infinitesimal step size (a.k.a. gradient flow) is equivalent to a kernel regression predictor with a deterministic kernel called neural tangent kernel (NTK).

Update for Student-teacher network

the gradient descent on the parameters can be written:

$$Z_{l+1}^i = Z_l^i - \gamma \underbrace{\mathbb{E}_{x\sim data}[(\frac{1}{N}\sum_{j=1}^{N}\phi(x, Z_l^j) - \frac{1}{M}\sum_{j=1}^{M}\phi(x, U^j))\nabla_{Z_l^i}\phi(x, Z_l^i)]}_{\nabla f_{\nu_l,\nu^*}(Z_l^i)},$$

where $(Z_l^i)_{1\leq i\leq N}$ are the particles at iteration l with distribution ν_l and
$\nabla f_{\nu_l,\nu^*}(Z_l^i) = \left(\frac{1}{N}\sum_{j=1}^{N}\nabla_2 k(Z_l^j, Z_l^i) - \frac{1}{M}\sum_{j=1}^{M}\nabla_2 k(U^j, Z_l^i)\right)$.

Proof for Wass. Proximal Gradient

- *Step 1 :* identify the optimal transport maps between $\mu_n, \nu_{n+1}, \mu_{n+1}$, show that $\forall\ n \geq 0$, $\nu_n, \mu_n \ll Leb$. (requires $\gamma < 1/L$)
- *Step 2:* prove a descent lemma (due to smoothness of V)

$$KL(\mu_{n+1}|\nu^*) \leq KL(\mu_n|\nu^*) - \gamma\left(1 - \frac{L\gamma}{2}\right)\|\nabla F + \nabla_{W_2}\mathcal{U}(\mu_{n+1}) \circ X_{n+1}\|^2_{L_2(\mu_n)},$$

  where $X_{n+1} = T^{\mu_{n+1}}_{\nu_{n+1}} \circ (I - \gamma\nabla V)$
- *Step 3 :* Discrete EVI for the entropy using its generalized geo. convexity (weaker than geo. convexity):

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

- *Step 4:* Discrete EVI for the potential using strong convexity.