Maximum Mean Discrepancy Gradient Flow

Michael Arbel [1] Anna Korba [1] Adil Salim [2] Arthur Gretton [1]

[1] Gatsby Computational Neuroscience Unit, UCL, London

[2] Visual Computing Center, KAUST, Saudi Arabia

Stuttgart Worskhop on Statistical Learning 2019

UCL

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Outline

Outline

Setting

- Let $\mathscr{X} \subset \mathbb{R}^d$ be the closure of a convex open set
- Let $\mathscr{P}_2(\mathscr{X})$ the set of probability measures on $\mathscr{X}$ with finite second moment

The space $\mathscr{P}_2(\mathscr{X})$ is endowed with the Wassertein-2 distance from **Optimal transport**:

$$W_2^2(\nu, \mu) = \inf_{\pi \in \Pi(\nu, \mu)} \int \|x - y\|^2 \, d\pi(x, y) \qquad \forall \nu, \mu \in \mathscr{P}_2(\mathscr{X})$$

where $\Pi(\nu, \mu)$ is the set of possible couplings between ν and μ.
In other words $\Pi(\nu, \mu)$ contains all possible distributions π on $\mathscr{X} \times \mathscr{X}$ such that if $(X, Y) \sim \pi$ then $X \sim \nu$ and $Y \sim \mu$.

Maximum Mean Discrepancy

- Let $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ a positive, semi-definite kernel
- $\mathscr{H}$ its corresponding RKHS. It is a Hilbert space with inner product $\langle ., . \rangle_{\mathscr{H}}$ and norm $\|.\|_{\mathscr{H}}$.

Suppose k is characteristic, ie the map:

$$\begin{aligned} \mathscr{P}_2(\mathscr{X}) &\to \mathscr{H} \\ \nu &\mapsto \int_{\mathscr{X}} k(x, .) d\nu(x) \end{aligned}$$

is injective.

Maximum Mean Discrepancy

Maximum Mean Discrepancy ([Gretton et al., 2012]) defines a distance on $\mathscr{P}_2(\mathscr{X})$:

$$MMD(\mu,\nu) = \|f_{\mu,\nu}\|_{\mathscr{H}}, \text{ where}$$
$$f_{\nu,\mu}(.) = \int k(x,.)d\nu(x) - \int k(x,.)d\mu(x)$$

$f_{\mu,\nu}$ is called the witness function and is the difference between the mean embeddings of ν and μ.

Now fix the (target) distribution μ. We consider the functional:

$$\mathscr{F}: \quad \mathscr{P}_2(\mathscr{X}) \to \mathbb{R}$$
$$\nu \mapsto \frac{1}{2} MMD^2(\mu,\nu)$$

Outline

Problem considered

Transport probability mass from a starting distribution ν to a target distribution μ, by finding a *continuous* path $(\nu_t)_{t\geq 0}$ decreasing $\mathscr{F}(\nu_t)$.

$\Longrightarrow$ **Gradient flows over the space of distributions** $\mathscr{P}_2(\mathscr{X})$

Problem considered

Transport probability mass from a starting distribution ν to a target distribution μ, by finding a *continuous* path $(\nu_t)_{t\geq 0}$ decreasing $\mathscr{F}(\nu_t)$.

$\Longrightarrow$ **Gradient flows over the space of distributions** $\mathscr{P}_2(\mathscr{X})$

This talk: Establish conditions for convergence of MMD gradient flow to its global optimum

- novel flow over the space of distributions
- can model the optimization of some overparameterized neural networks models
- we propose a trick to improve convergence

Continuous time flows

In a **euclidean** setting, a curve $x : [0, \infty] \to \mathbb{R}^d$ is the gradient flow, or steepest descent of a differentiable function $F : \mathbb{R}^d \to R$ if:

$$\frac{dx_t}{dt} = -\nabla F(x_t)$$

- Initial value problem: given x_0, find the gradient flow $(x_t)_{t \geq 0}$.

Continuous time flows

In a **euclidean** setting, a curve $x : [0, \infty] \to \mathbb{R}^d$ is the gradient flow, or steepest descent of a differentiable function $F : \mathbb{R}^d \to R$ if:

$$\frac{dx_t}{dt} = -\nabla F(x_t)$$

- Initial value problem: given x_0, find the gradient flow $(x_t)_{t \geq 0}$.

By analogy, one can interpret the **gradient flow of a functional** $\mathscr{F} : \mathscr{P}_2(\mathscr{X}) \to \mathbb{R}$ to be a curve $\nu : [0, \infty] \to \mathscr{P}_2(\mathscr{X})$ that satisfies:

$$\frac{\partial \nu_t}{\partial t} = -\nabla_{W_2} \mathscr{F}(\nu_t)$$

for some generalized notion of gradient ∇_{W_2}, w.r.t. the W_2 metric.

Wassertein-2 gradient flows ([Ambrosio et al., 2008])

For a sufficiently regular $\mathscr{F}$ and ν, we can write:

$$-\nabla_{W_2}\mathscr{F}(\nu) = div(\nu\nabla\frac{\partial\mathscr{F}}{\partial\nu})$$

where $\frac{\partial\mathscr{F}}{\partial\nu}$ denotes the first variation of $\mathscr{F}$ at ν.

If it exists, it is the unique function such that for any $\nu, \nu' \in \mathscr{P}_2(\mathscr{X})$:

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}(\mathscr{F}(\nu+\varepsilon(\nu'-\nu))-\mathscr{F}(\nu)) = \int_{\mathscr{X}}\frac{\partial\mathscr{F}}{\partial\nu}(\nu)(d\nu'-d\nu)$$

Wassertein gradient flows

Since $-\nabla_{W_2}\mathscr{F}(\nu) = div(\nu\nabla\frac{\partial\mathscr{F}}{\partial\nu})$, all Wassertein gradient flows are of the form:

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

continuity equation

Ruling the density ρ_t of particles driven by a velocity field V_t ($-\nabla\frac{\partial\mathscr{F}}{\partial\nu}$).

Wassertein gradient flows

Since $-\nabla_{W_2}\mathscr{F}(\nu) = div(\nu\nabla\frac{\partial\mathscr{F}}{\partial\nu})$, all Wassertein gradient flows are of the form:

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

continuity equation

Ruling the density ρ_t of particles driven by a velocity field V_t ($-\nabla\frac{\partial\mathscr{F}}{\partial\nu}$).

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

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

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

Outline

MMD functional

For a target distribution μ (fixed), for any $\nu \in \mathscr{P}_2(\mathscr{X})$:

$$\mathscr{F}(\nu) = \frac{1}{2} MMD^2(\mu, \nu) = \frac{1}{2} \|f_{\mu,\nu}\|^2_{\mathscr{H}}$$

- Since $\mathscr{F}(\nu) = \frac{1}{2}(\int f_{\mu,\nu} d\mu - \int f_{\mu,\nu} d\nu)$, we have $\frac{\partial \mathscr{F}}{\partial \nu} = f_{\mu,\nu}$
- Then, $\mathscr{F}$ can be written as a free energy:

$$\mathscr{F}(\nu) = \underbrace{\int V(x) d\nu(x)}_{\mathscr{V}} + \underbrace{\frac{1}{2} \int W(x,y) d\nu(x) d\nu(y)}_{\mathscr{W}} + C.$$

where V is a confinement potential, W an interaction potential and C a constant defined by:

$$V(x) = -\int k(x,x') d\mu(x'), \;\; W(x,x') = k(x,x'), \;\; C = \frac{1}{2} \int k(x,x') d\mu(x) d\mu(x')$$

MMD Gradient flow

The MMD gradient flow w.r.t. W_2 is thus given by:

$$\frac{\partial \nu_t}{\partial t} = div(\nu_t \nabla f_{\mu,\nu_t}) = div(\nu_t \nabla (V + W * \nu_t)) \qquad (1)$$

where $\nabla f_{\mu,\nu_t}(z) = \int \nabla k(x,z) d\mu(x) - \int \nabla k(x,z) d\nu_t(x)$.

This type of equation is associated in the probability theory literature to the so-called **McKean-Vlasov process** [Kac, 1956]:

$$dX_t = -\nabla f_{\mu, \underbrace{\nu_t}_{\text{depends on the current distribution of the process!}}}(X_t) dt \qquad X_0 \sim \nu_0$$

whose distribution satisfy (1).

Outline

First strategy - Convexity on the space of distributions

Definition

A curve $\rho : [0,1] \to \mathscr{P}(\mathscr{X})$ is a geodesic between ν and μ if:
$\rho(0) = \nu,\ \rho(1) = \mu$, and
$L(\rho) = \min\{L(\widetilde{\rho}), \widetilde{\rho}(0) = \nu, \widetilde{\rho}(1) = \mu\} = W_2(\nu, \mu)$.

(λ)-Geodesic convexity: Convexity of the functional $\mathscr{F}$ on geodesic curves of $\mathscr{P}_2(\mathscr{X})$.

$$\mathscr{F}(\rho(t)) \leq (1-t)\mathscr{F}(\rho(0)) + t\mathscr{F}(\rho(1)) - t(1-t)\frac{\lambda}{2}d(\rho(0), \rho(1))^2$$

(classic λ-convexity on $\mathbb{R}^d$: $F((1-t)x+ty) \leq (1-t)F(x) + tF(y) - t(1-t)\frac{\lambda}{2}|x-y|^2$)

First strategy - Convexity on the space of distributions

Definition

A curve $\rho : [0,1] \to \mathscr{P}(\mathscr{X})$ is a geodesic between ν and μ if:
$\rho(0) = \nu,\ \rho(1) = \mu$, and
$L(\rho) = \min\{L(\widetilde{\rho}), \widetilde{\rho}(0) = \nu, \widetilde{\rho}(1) = \mu\} = W_2(\nu, \mu)$.

(λ)-Geodesic convexity: Convexity of the functional $\mathscr{F}$ on geodesic curves of $\mathscr{P}_2(\mathscr{X})$.

$$\mathscr{F}(\rho(t)) \leq (1-t)\mathscr{F}(\rho(0)) + t\mathscr{F}(\rho(1)) - t(1-t)\frac{\lambda}{2}d(\rho(0), \rho(1))^2$$

(classic λ-convexity on $\mathbb{R}^d$: $F((1-t)x+ty) \leq (1-t)F(x) + tF(y) - t(1-t)\frac{\lambda}{2}|x-y|^2$)

Our finding: The MMD is λ-convex with $\lambda < 0$.

Too bad... $\lambda > 0$ would have guaranteed that all gradient flows of $\mathscr{F}$ would converge the **unique** minimizer of $\mathscr{F}$.

Second strategy - Obtain a Lojasiewicz inequality

$$\frac{d\mathscr{F}(\nu_t)}{dt} \leq -C\mathscr{F}(\nu_t)^2 \tag{2}$$

Applying Gronwall's lemma results in: $\Rightarrow \mathscr{F}(\nu_t) = \mathscr{O}(\frac{1}{t})$.

Second strategy - Obtain a Lojasiewicz inequality

$$\frac{d\mathscr{F}(\nu_t)}{dt} \leq -C\mathscr{F}(\nu_t)^2 \tag{2}$$

Applying Gronwall's lemma results in: $\Rightarrow \mathscr{F}(\nu_t) = \mathscr{O}(\frac{1}{t})$.

- on the left we have the weighted Sobolev semi-norm:

$$\frac{d\mathscr{F}(\nu_t)}{dt} = -\int \|\nabla f_{\mu,\nu_t}(x)\|^2 \nu_t(x) = -\|f_{\mu,\nu_t}\|^2_{\dot{H}(\nu_t)}$$

- on the right the RKHS norm: $\mathscr{F}(\nu_t) = \frac{1}{2}\|f_{\mu,\nu_t}\|^2_{\mathscr{H}}$

Second strategy - Obtain a Lojasiewicz inequality

$$\frac{d\mathscr{F}(\nu_t)}{dt} \leq -C\mathscr{F}(\nu_t)^2 \tag{2}$$

Applying Gronwall's lemma results in: $\Rightarrow \mathscr{F}(\nu_t) = \mathscr{O}(\frac{1}{t})$.

- on the left we have the weighted Sobolev semi-norm:

$$\frac{d\mathscr{F}(\nu_t)}{dt} = -\int \|\nabla f_{\mu,\nu_t}(x)\|^2 \nu_t(x) = -\|f_{\mu,\nu_t}\|^2_{\dot{H}(\nu_t)}$$

- on the right the RKHS norm: $\mathscr{F}(\nu_t) = \frac{1}{2}\|f_{\mu,\nu_t}\|^2_{\mathscr{H}}$

Let $L_2(\nu) = \{f, \int f(x)^2 d\nu(x) < \infty\}$, and $\|.\|_{\dot{H}^{-1}(\nu)}$ the weighted negative Sobolev norm, defined for $p, q \in \mathscr{P}_2(\mathscr{X})$ by:

$\|p - q\|_{\dot{H}^{-1}(\nu)} = \sup_{f \in L_2(\nu), \|f\|_{\dot{H}(\nu)} \leq 1} |\int f(x)dp(x) - \int f(x)dq(x)|$.

linearizes the W_2!

A condition for global convergence

It can be shown that:

$$\|f_{\mu,\nu_t}\|^2_{\mathcal{H}} \leq \|f_{\mu,\nu_t}\|_{\dot{H}(\nu_t)} \|\mu - \nu_t\|_{\dot{H}^{-1}(\nu_t)}.$$

Proof. Take $g = \|f_{\mu,\nu_t}\|^{-1}_{\dot{H}(\nu_t)} f_{\mu,\nu_t}$

- by def, $|\int g d\nu_t - \int g d\mu| = \|f_{\mu,\nu_t}\|^{-1}_{\dot{H}(\nu_t)} \left|\int f_{\mu,\nu_t} d\nu_t - \int f_{\mu,\nu_t} d\mu\right| = \|f_{\mu,\nu_t}\|^{-1}_{\dot{H}(\nu_t)} \|f_{\mu,\nu_t}\|^2_{\mathcal{H}}$
- $g \in L_2(\nu_t)$[1] and $\|g\|_{\dot{H}(\nu)} \leq 1$, so $|\int g d\nu_t - \int g d\mu| \leq \|\nu_t - \mu\|_{\dot{H}^{-1}(\nu_t)}$

Hence, provided $\|\mu - \nu_t\|^2_{\dot{H}^{-1}(\nu_t)} \leq \frac{4}{C}$, we obtain the Lojasiewicz inequality (2) **and the rate $\mathcal{O}(1/t)$ (thus global convergence).**

However in practice it is hard to guarantee this condition ([Peyre, 2018]).

[1] Under a Lipschitz assumption on ∇k, for all $\nu, \mu \in \mathcal{P}_2(\mathcal{X}), f_{\mu,\nu} \in L_2(\nu)$

Outline

Euler scheme (Time-discretization of the flow)

For any $T : \mathscr{X} \to \mathscr{X}$ a measurable map, and $\nu \in \mathscr{P}_2(\mathscr{X})$, we denote the pushforward measure by $T_{\#}\nu$.

$$T_{\#}\nu(A) = \nu(T^{-1}(A)) \text{ for every measurable set A,}$$

Euler scheme (Time-discretization of the flow)

For any $T : \mathscr{X} \to \mathscr{X}$ a measurable map, and $\nu \in \mathscr{P}_2(\mathscr{X})$, we denote the pushforward measure by $T_\# \nu$.

$$T_\# \nu(A) = \nu(T^{-1}(A)) \text{ for every measurable set A,}$$

Starting from $\nu_0 \in \mathscr{P}_2(\mathscr{X})$ and using a step-size $\gamma > 0$, a sequence $\nu_n \in \mathscr{P}_2(\mathscr{X})$ is given by iteratively applying

$$\nu_{n+1} = (I - \gamma \nabla f_{\mu,\nu_n})_\# \nu_n. \tag{3}$$

For all n, equation (3) is the distribution of the process defined by

$$X_{n+1} = X_n - \gamma \nabla f_{\mu,\nu_n}(X_n) \qquad X_0 \sim \nu_0.$$

A noisy update as regularization

The condition we exhibited for global convergence may not hold and $(\mathscr{F}(\nu_n))_{n\in\mathbb{N}}$ might be stuck at a local minima.

$$\frac{d\mathscr{F}(\nu_t)}{dt} = -\int \|\nabla f_{\mu,\nu_t}(x)\|^2 d\nu_t(x) \text{ at equilibrium } \implies \int \|\nabla f_{\mu,\nu^*}(x)\|^2 d\nu^*(x) = 0$$

If ν^* positive everywhere this implies $f_{\mu,\nu^*} = cte = 0$ as soon as $0 \notin \mathscr{H}$. But ν^* might be singular...

Our proposal: Inject noise into the gradient during updates:

$$X_{n+1} = X_n - \gamma \nabla f_{\mu,\nu_n}(X_n + \beta_n U_n), \quad n \geq 0,$$

where $U_n \sim \mathscr{N}(0,1)$ and β_n is the noise level at n.

⚠Different from adding a noise outside the gradient (i.e. diffusion)!

Guarantees

Proposition

For a choice of β_n such that:

$$8\lambda^2\beta_n^2\mathscr{F}(\nu_n) \leq \int \|\nabla f_{\mu,\nu_n}(x+\beta_n u)\|^2 d\nu_n(x))dg(u) \quad (4)$$

the following inequality holds:

$$\mathscr{F}(\nu_{n+1}) - \mathscr{F}(\nu_n) \leq -\frac{\gamma}{2}(1-\frac{3}{2}\gamma L)\int \|\nabla f_{\mu,\nu_n}(x+\beta_n u)\|^2 d\nu_n(x))dg(u)$$

where λ and L are Lipschitz constants on the first derivatives of k.

Guarantees

Proposition

For a choice of β_n such that:

$$8\lambda^2\beta_n^2\mathscr{F}(\nu_n) \leq \int \|\nabla f_{\mu,\nu_n}(x+\beta_n u)\|^2 d\nu_n(x))dg(u) \tag{4}$$

the following inequality holds:

$$\mathscr{F}(\nu_{n+1}) - \mathscr{F}(\nu_n) \leq -\frac{\gamma}{2}(1-\frac{3}{2}\gamma L)\int \|\nabla f_{\mu,\nu_n}(x+\beta_n u)\|^2 d\nu_n(x))dg(u)$$

where λ and L are Lipschitz constants on the first derivatives of k.

Moreover under (4)

$$\mathscr{F}(\nu_n) \leq \mathscr{F}(\nu_0)e^{-\Gamma\sum_{i=0}^{n}\beta_i^2}.$$

where $\Gamma = 4\lambda^2\gamma(1-\frac{3}{2}\gamma L$.

So $\sum_{i=0}^{n}\beta_i^2 \to \infty$ with* (4) *implies global convergence.

The sample-based approximate scheme

How can we simulate

$$X_{n+1} = X_n - \gamma \nabla f_{\mu,\nu_n}(X_n + \beta_n U_n), \quad n \geq 0?$$

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

$$X_{n+1} = X_n - \gamma \nabla f_{\mu,\nu_n}(X_n + \beta_n U_n), \quad n \geq 0?$$

It depends on:

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

$$\widehat{\nu}_{n+1} \begin{cases} X_{n+1}^1 = X_n^1 - \gamma \nabla f_{\widehat{\mu},\widehat{\nu}_n}(X_n^1 + \beta_n U_n^1) \\ \dots \\ X_{n+1}^N = X_n^N - \gamma \nabla f_{\widehat{\mu},\widehat{\nu}_n}(X_n^N + \beta_n U_n^N) \end{cases}$$

The sample-based approximate scheme

How can we simulate

$$X_{n+1} = X_n - \gamma \nabla f_{\mu,\nu_n}(X_n + \beta_n U_n), \quad n \geq 0?$$

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 ($\widehat{\mu}$)

$$\widehat{\nu}_{n+1} \begin{cases} X_{n+1}^1 = X_n^1 - \gamma \nabla f_{\widehat{\mu},\widehat{\nu}_n}(X_n^1 + \beta_n U_n^1) \\ \dots \\ X_{n+1}^N = X_n^N - \gamma \nabla f_{\widehat{\mu},\widehat{\nu}_n}(X_n^N + \beta_n U_n^N) \end{cases}$$

Our guarantees: For any iteration $n \in \mathbb{N}$ and $T > 0$, if $\beta_n < B$:

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

Outline

Overparameterized single-layer neural network

Single-layer neural network, parameterized by $\theta \in \mathscr{X}$. Let (x, y) denote the input/output data.

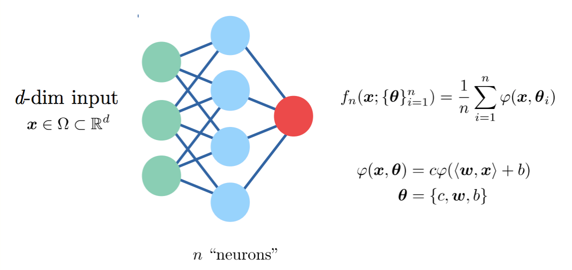


Figure: [Rotskoff et al., 2019]

Consider the supervised learning problem:

$$\min_{(\theta_1,\ldots,\theta_n)\in \underbrace{\mathscr{X}}_{\text{Parameter space}}} \mathbb{E}_{(x,y)\sim p}\left[\left\| y - \frac{1}{n}\sum_{i=1}^n \phi(x, \theta_i)\right\|^2\right]$$

Motivation

If $n \to \infty$, the previous problem can be rewritten:

$$\underbrace{\min_{\nu \in \mathscr{P}_2(\mathscr{X})}}_{\text{Distributions over the parameter space}} \mathscr{L}(\nu) \quad := \quad \mathbb{E}_{(x,y)\sim p}\left[\left\| y - \underbrace{\int \psi(x,\theta) d\nu(\theta)}_{\Psi(x,\nu)} \right\|^2\right]$$

If $\exists \mu \in \mathscr{P}_2(\mathscr{X})$ s.t. $\mathbb{E}_{y\sim p(.|x)}[y] = \int \psi(x,\theta) d\mu(\theta)$,

$\Longrightarrow \mathscr{L}(\nu) = MMD^2(\nu,\mu)$ with $k(\theta,\theta') = \mathbb{E}_{x\sim p}\left[\psi(x,\theta)^T \psi(x,\theta')\right]$.

[Chizat and Bach, 2018], [Rotskoff et al., 2019]: gradient descent on the parameters of a neural network can be seen as a particle transport problem.

Experiments - training a student-teacher network

- the teacher network $\Psi_T(x,\mu)$ is given by M particles $\mathscr{X} = (\xi_1, ..., \xi_M)$ which are fixed during training $\Longrightarrow$ $\mu = \frac{1}{M}\sum_{j=1}^{M} \delta_{\xi_j}$
- the student network $\Psi_S(x,\nu_\Theta)$ has N particles $\Theta = (\theta_1, ..., \theta_N)$ that are initialized randomly $\Longrightarrow$ $\nu_\Theta = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_j}$

Experiments - training a student-teacher network

- the teacher network $\Psi_T(x, \mu)$ is given by M particles $\mathscr{X} = (\xi_1, ..., \xi_M)$ which are fixed during training $\Longrightarrow$ $\mu = \frac{1}{M} \sum_{j=1}^{M} \delta_{\xi_j}$
- the student network $\Psi_S(x, \nu_\Theta)$ has N particles $\Theta = (\theta_1, ..., \theta_N)$ that are initialized randomly $\Longrightarrow$ $\nu_\Theta = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_j}$

Performing gradient descent to minimize

$$\min_{\Theta} \mathbb{E}_{x \sim p} \left[(\Psi_T(x, \mu) - \Psi_S(x, \nu_\Theta))^2 \right]$$

can be seen as a particle version of the gradient flow of the MMD with a kernel given by $k(\theta, \theta') = \mathbb{E}_{x \sim p}[\psi(x, \theta')\psi(z, \theta)]$

Experiments - training a student-teacher network

- the teacher network $\Psi_T(x,\mu)$ is given by M particles $\mathscr{X} = (\xi_1,\ldots,\xi_M)$ which are fixed during training $\Longrightarrow$ $\mu = \frac{1}{M}\sum_{j=1}^{M}\delta_{\xi_j}$
- the student network $\Psi_S(x,\nu_\Theta)$ has N particles $\Theta = (\theta_1,\ldots,\theta_N)$ that are initialized randomly $\Longrightarrow$ $\nu_\Theta = \frac{1}{N}\sum_{i=1}^{N}\delta_{\theta_j}$

Performing gradient descent to minimize

$$\min_{\Theta}\mathbb{E}_{x\sim p}\left[(\Psi_T(x,\mu)-\Psi_S(x,\nu_\Theta))^2\right]$$

can be seen as a particle version of the gradient flow of the MMD with a kernel given by $k(\theta,\theta') = \mathbb{E}_{x\sim p}[\psi(x,\theta')\psi(z,\theta)]$

$\Longrightarrow$ approximated by

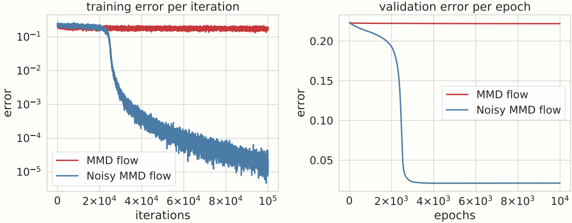
$$\hat{k}(\theta,\theta') = \frac{1}{n_b}\sum_{b=1}^{n_b}\psi(x_b,\theta)^T\psi(x_b,\theta').$$

where $(x_1,\ldots,x_{n_b})$ are n_b samples from the data distribution.

Experiments

Leads to the approximate update:

$$\theta_{n+1}^{i} = \theta_{n}^{i} - \gamma \nabla \hat{f}_{\mu,\nu_n}(\theta_n^i)$$



$\Longrightarrow$ adding noise to the gradient seems to lead to global convergence.

Outline

Summary and openings

What we have done:

- novel flow over the space of distributions
- theoretical results on the MMD flow
- trick to improve convergence

Summary and openings

What we have done:

- novel flow over the space of distributions
- theoretical results on the MMD flow
- trick to improve convergence

Future work:

- Deeper understanding of the regularization proposed (continuous formulation?) and of the choice of the kernel
- Other regularizations to improve convergence?
- Other gradient flows? (here: $\nabla_{W_2}\mathscr{F}$ vs [Rotskoff et al., 2019] who get a global convergence for $\rightarrow \nabla_{WFR}\mathscr{F}$).

Ambrosio, L., Gigli, N., and Savaré, G. (2008).
Gradient flows: in metric spaces and in the space of probability measures.
Springer Science & Business Media.

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

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012).
A kernel two-sample test.
Journal of Machine Learning Research.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998).
The variational formulation of the fokker–planck equation.
SIAM journal on mathematical analysis, 29(1):1–17.

Kac, M. (1956).
Foundations of kinetic theory.

In *Proceedings of The third Berkeley symposium on mathematical statistics and probability*, volume 3, pages 171–197. University of California Press Berkeley and Los Angeles, California.

Mroueh, Y., Sercu, T., and Raj, A. (2019).
Sobolev descent.
In *AISTATS*.

Peyre, R. (2018).
Comparison between w2 distance and $\dot{H}$- 1 norm, and localization of wasserstein distance.
ESAIM: Control, Optimisation and Calculus of Variations, 24(4):1489–1501.

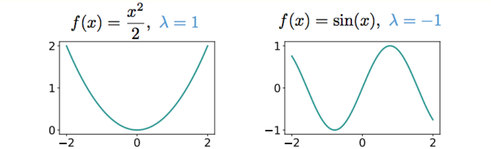
Rotskoff, G., Jelassi, S., Bruna, J., and Vanden-Eijnden, E. (2019).
Global convergence of neuron birth-death dynamics.
In *ICML*.

Convexity on vector spaces

Existence, uniqueness results on gradient flows rely on the notion of **convexity**.

A function F defined on $\mathbb{R}^d$ is λ-convex if $D^2 F \geq \lambda I_{d\times d}$ or equivalently if for any $x, y \in \mathbb{R}^d$ and $t \in [0,1]$:

$$F((1-t)x + ty) \leq (1-t)F(x) + tF(y) - t(1-t)\frac{\lambda}{2}|x-y|^2$$



- Uniqueness when $\lambda > 0$: any gradient flow $x(t)$ converges to some $x^\star$.

W_2 distance

$$W_2^2(\nu,\mu) = \inf_{\pi\in\Pi(\nu,\mu)} \int \|x-y\|^2 \, d\pi(x,y) \qquad \forall \nu,\mu \in \mathscr{P}_2(\mathscr{X})$$

where $\Pi(\nu,\mu)$ is the set of possible couplings between ν and μ. In other words $\Pi(\nu,\mu)$ contains all possible distributions π on $\mathscr{X} \times \mathscr{X}$ such that if $(X,Y) \sim \pi$ then $X \sim \nu$ and $Y \sim \mu$.

W_2 vs L_2?
L_2 geodesic: $\rho(t) = (1-t)\rho(0) + t\rho(1)$
W_2 geodesic: $\rho(t) = ((1-t)Id + tT_{\rho(0),\rho(1)})_{\#}\rho(0)$

Informally, L^p distances are ”vertical” (values of the distributions) whereas W_p distances are ”horizontal” (mass of the distributions).

Gradient flows - comparison

	Euclidean	W_2
Metric (X,d)	$(\mathbb{R}^d, \|.\|)$	$(\mathscr{P}_2(\mathbb{R}^d), W_2)$
Definition of ∇_X	$\langle \nabla F(x), v \rangle =$	$\langle \nabla \mathscr{F}(\nu), -div(\xi\nu) \rangle_{Tan_\mu \mathscr{P}_2(\mathbb{R}^d)} =$
	$\lim_{h\to 0} \frac{F(x+hv)-F(x)}{h}$	$\lim_{h\to 0} \frac{\mathscr{F}((I+h\xi)_\# \nu)-\mathscr{F}(\nu)}{h}$
Formula for ∇_X	$\nabla_{\mathbb{R}^d} F(x) = \nabla F(x)$	$\nabla_{W_2}\mathscr{F} = -div(\nu \nabla \frac{\partial \mathscr{F}}{\partial \nu})$

Comparison with Langevin

Seminal work of [Jordan et al., 1998] who revealed that the Fokker-Planck equation is a gradient flow of the Kullback-Leibler divergence:

$$\frac{\partial \nu}{\partial t} - div(\nu V) = 0, \text{ where the vector field } V = \nabla_{W_2} KL(\nu) = \nabla log(\frac{\nu}{\mu}).$$

Results in the Langevin Monte-Carlo algorithm (requires the knowledge of $\nabla log(\mu)$):

$$X_{n+1} = X_n - \gamma \nabla log(\mu)(X_n) + \varepsilon_n$$

where $\varepsilon_n \sim \mathcal{N}(0,1)$.

Free energies

1. $\mathscr{F}(\nu) = KL(\mu, \nu)$ admits a free-energy expression:

$$\mathscr{F}(\nu) = \underbrace{\int U(\nu(x))dx}_{\mathscr{U}} + \underbrace{\int V(x)\nu(x)dx}_{\mathscr{V}}$$

   with $U(s)$ the internal potential (entropy function) and V confinement potential defined as:

$$U(s) = s\log(s), \;\; V(x) = -log(\mu(x))$$

2. $\mathscr{F}(\nu) = \frac{1}{2}MMD^2(\mu, \nu)$ also:

$$\mathscr{F}(\nu) = \underbrace{\int V(x)d\nu(x)}_{\mathscr{V}} + \underbrace{\frac{1}{2}\int W(x,y)d\nu(x)d\nu(y)}_{\mathscr{W}} + C.$$

   where V is a confinement potential, W an interaction potential and C a constant defined by:

$$V(x) = -\int k(x,x')d\mu(x'), \;\; W(x,x') = k(x,x'), \;\; C = \frac{1}{2}\int k(x,x')d\mu(x)d\mu(x')$$