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

Fields Institute
September 2020

Problem and Outline

Problem:

- Transport mass from a starting probability distribution to a target distribution
- How? By finding a *continuous* path on the space of distributions, decreasing some loss (**Wasserstein gradient flows**)
- **This work:** Minimize the Maximum Mean Discrepancy (MMD) on the space of probability distributions.

Application : Insights on the theoretical properties of some large neural networks and alteration of the dynamics to improve convergence.

Outline

Setting

Let $\mathcal{P}$ the set of probability measures on $\mathcal{Z} \subset \mathbb{R}^d$ with finite second moment :

$$\mathcal{P} = \{\mu \in \mathcal{P}(\mathcal{Z}), \int \|z\|^2 d\mu(z) < \infty\}$$

Setting

Let $\mathcal{P}$ the set of probability measures on $\mathcal{Z} \subset \mathbb{R}^d$ with finite second moment :

$$\mathcal{P} = \{\mu \in \mathcal{P}(\mathcal{Z}), \int \|z\|^2 d\mu(z) < \infty\}$$

The space $\mathcal{P}$ is endowed with the Wassertein-2 distance from

Optimal transport :

$$W_2^2(\nu, \mu) = \inf_{\pi \in \Pi(\nu,\mu)} \int_{\mathcal{Z}\times\mathcal{Z}} \left\|z - z'\right\|^2 d\pi(z, z') \qquad \forall \nu, \mu \in \mathcal{P}$$

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 $\mathcal{Z} \times \mathcal{Z}$ such that if $(Z, Z') \sim \pi$ then $Z \sim \nu$ and $Z' \sim \mu$.

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 ν^* (fixed), for any $\nu \in \mathcal{P}$:

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

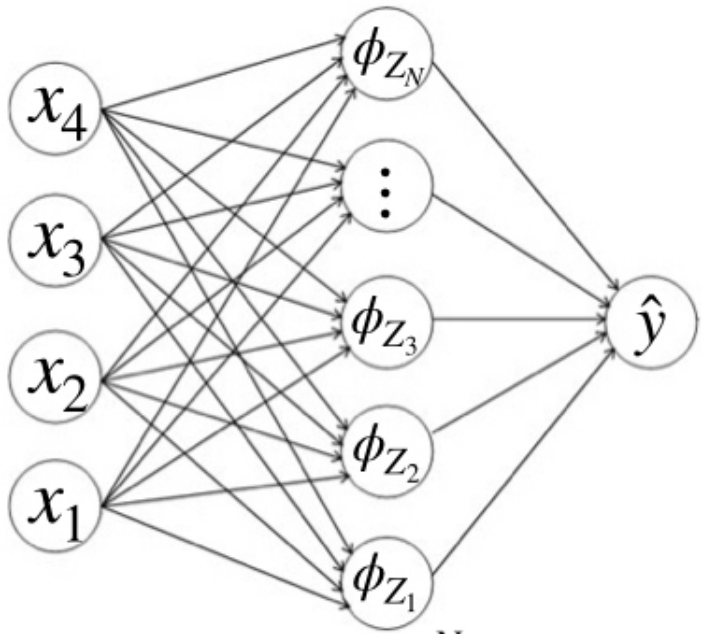
Proof : use the reproducing property with
$f_{\nu,\nu^*}(.) = \int k(z,.)d\nu(z) - \int k(z,.)d\nu^*(z)$

Appear as a loss when optimizing some large neural networks.

Regression with 1-hidden layer NN

Consider the following regression problem:

$$(x, y) \sim data$$

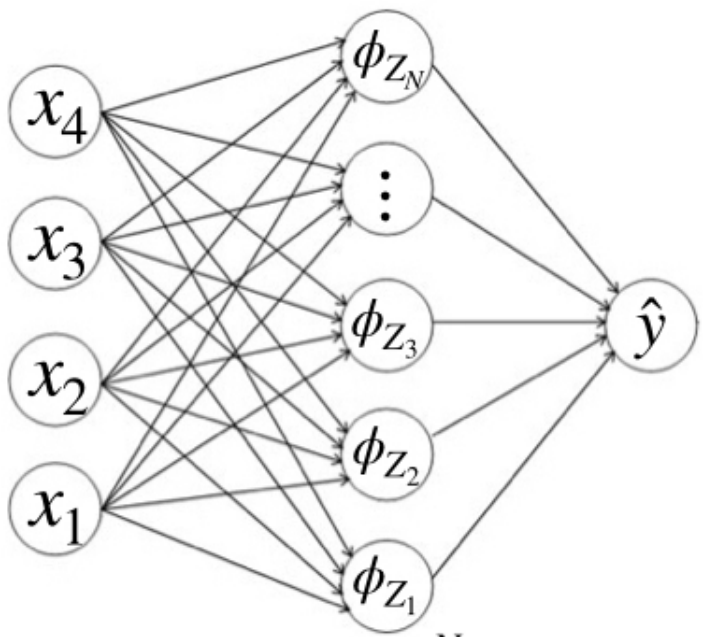


$$\min_{Z_1,\ldots,Z_N} \mathbb{E}_{data}[\|y - \frac{1}{N}\sum_{i=1}^{N} \phi_{Z_i}(x)\|^2]$$

Regression with 1-hidden layer NN

Consider the following regression problem:

$$(x, y) \sim data$$



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

- $\phi_{Z_i}(x) = w_i g(x, \theta_i)$,
  $Z_i = (w_i, \theta_i) \in \mathbb{R} \times \mathbb{R}^d$
  ϕ_{Z_i} : non linearity
- Example:

$$\phi_Z(x) = wg(ax + b)$$

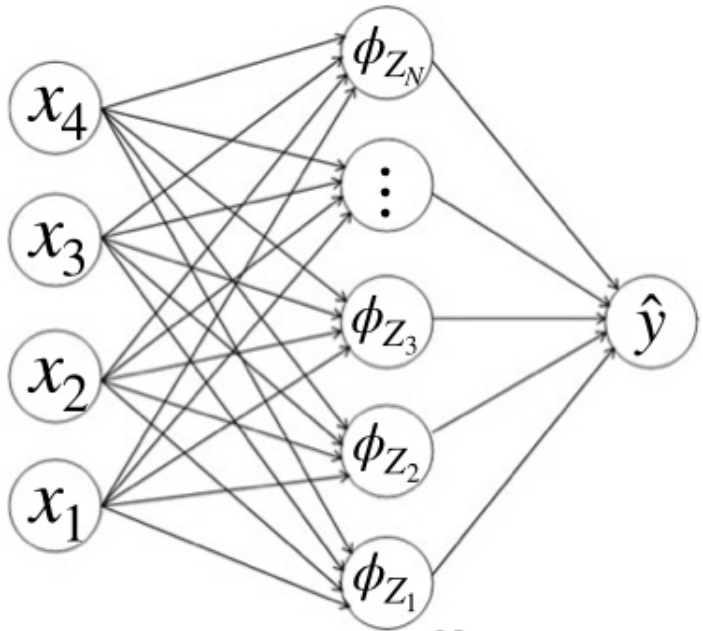
where $g : \mathbb{R} \to \mathbb{R}$
(sigmoid
$g(z) = 1/(1 + e^{-z})$, RelU
$(g(z) = max(0, z)...)$

Regression with 1-hidden layer NN

Finite dimensional non-convex optimization (regression setting):

$$\min_{z_1,\dots,z_N \in \mathcal{Z}} \mathcal{F}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{z_i}\right)$$

$(x, y) \sim data$



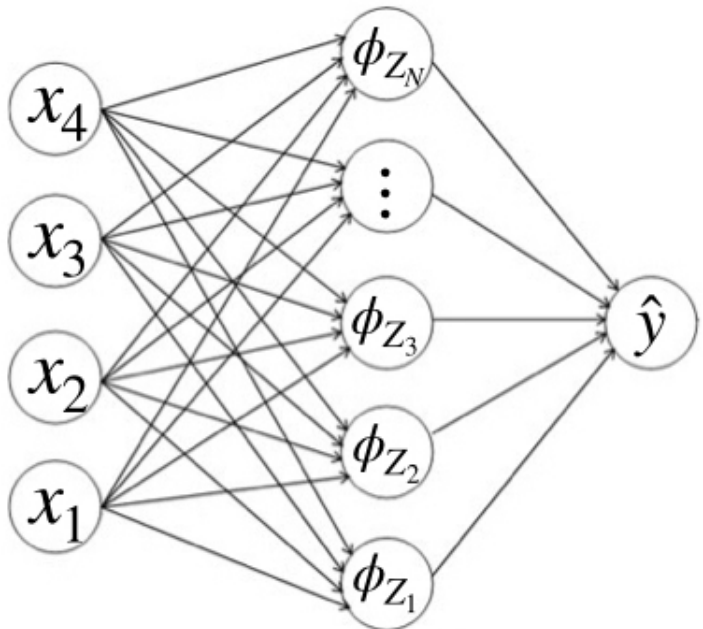
$$\min_{Z_1,\dots,Z_N} \mathbb{E}_{data}[\|y - \frac{1}{N}\sum_{i=1}^{N} \phi_{Z_i}(x)\|^2]$$

Regression with 1-hidden layer NN

Finite dimensional non-convex optimization (regression setting):

$$\min_{Z_1,\dots,Z_N \in \mathcal{Z}} \mathcal{F}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i}\right)$$

$(x, y) \sim data$



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

- Optimization using gradient descent (GD):

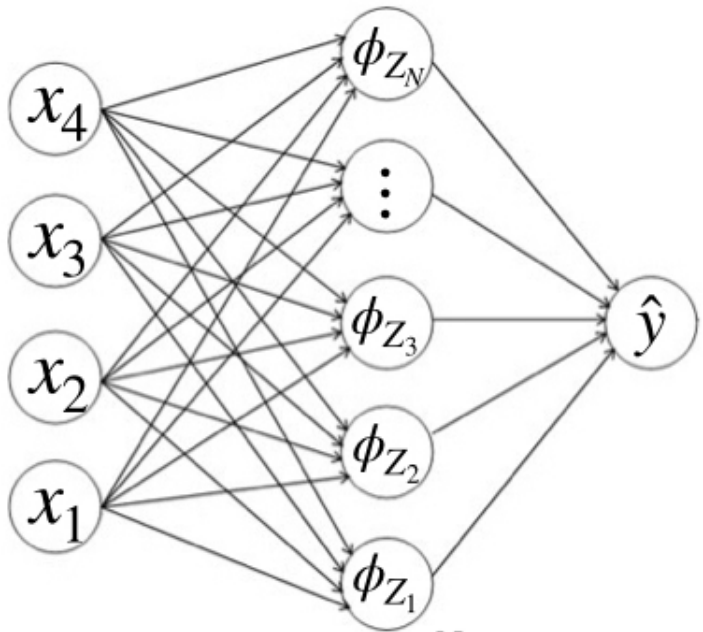
$$Z_i^{t+1} = Z_i^t - \gamma \nabla_{Z_i} \mathcal{F}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i^t}\right)$$

Regression with 1-hidden layer NN

Finite dimensional non-convex optimization (regression setting):

$$\min_{Z_1,\dots,Z_N \in \mathcal{Z}} \mathcal{F}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i}\right)$$

$(x, y) \sim data$



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

- Optimization using gradient descent (GD):

$$Z_i^{t+1} = Z_i^t - \gamma \nabla_{Z_i} \mathcal{F}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i^t}\right)$$

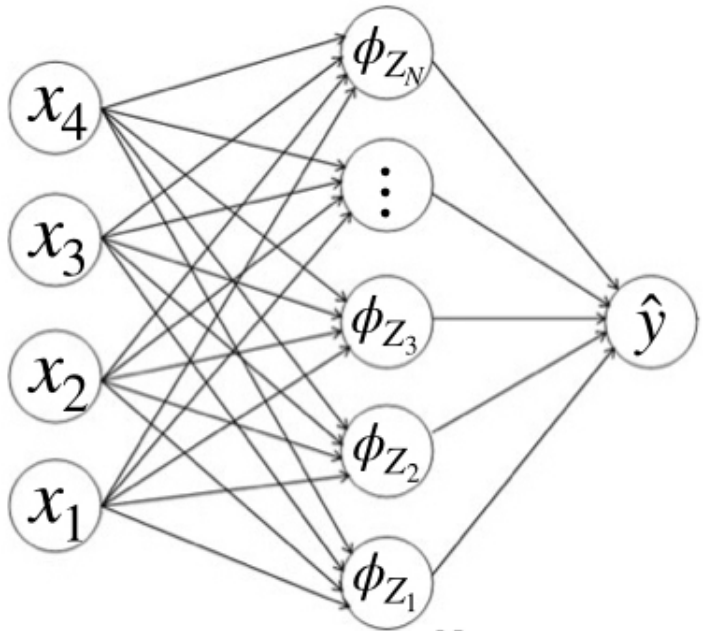
- Hard to describe the dynamics of GD!

Regression with 1-hidden layer NN

Finite dimensional non-convex optimization (regression setting):

$$\min_{Z_1,\dots,Z_N \in \mathcal{Z}} \mathcal{F}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i}\right)$$

$(x, y) \sim data$



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

- Optimization using gradient descent (GD):

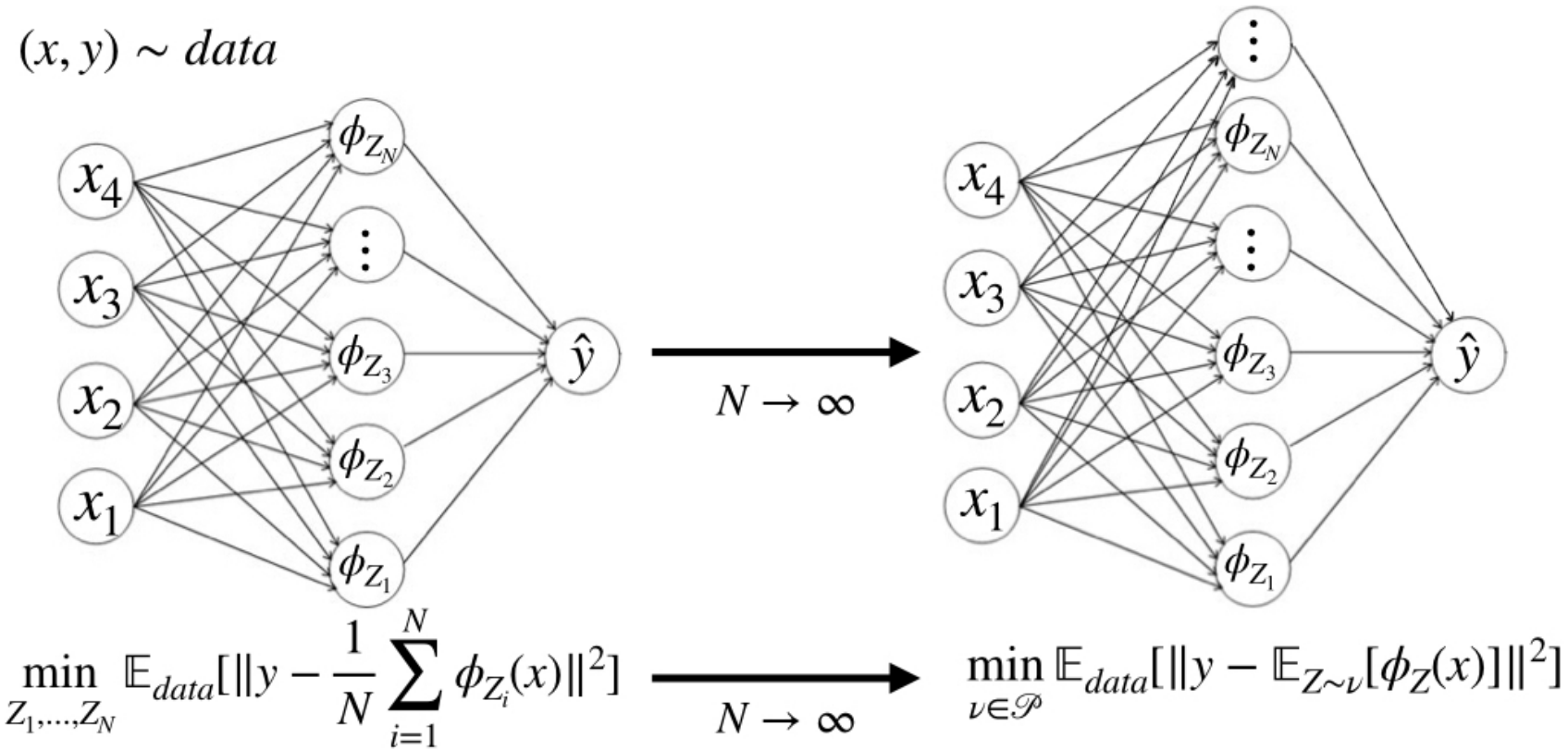
$$Z_i^{t+1} = Z_i^t - \gamma \nabla_{Z_i} \mathcal{F}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i^t}\right)$$

- Hard to describe the dynamics of GD!
- Idea: look at the distribution of the Z_i's

Infinite width regime

Infinite dimensional convex optimization [Chizat and Bach, 2018], [Mei et al., 2018]:

$$\min_{Z_1,\dots,Z_N \in \mathcal{Z}} \mathcal{F}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i}\right) \xrightarrow[N\to\infty]{} \min_{\nu \in \mathcal{P}} \mathcal{F}(\nu)$$



Minimization of the MMD : the well-specified case

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

Minimization of the MMD : the well-specified case

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

Then :

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

$$\Updownarrow$$

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

$$\Updownarrow$$

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

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

$$\Updownarrow$$

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

Minimization of the MMD : the well-specified case

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

Then :

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

$$\Updownarrow$$

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

$$\Updownarrow$$

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

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

$$\Updownarrow$$

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

Optimizing the parameters of a NN $\Leftrightarrow$ minimization of the MMD on $\mathcal{P}$ in the population limit ($N \to \infty$).

Outline

We consider

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

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

$$\frac{\partial \nu_t}{\partial t} = div(\nu_t \nabla \frac{\partial \mathcal{F}(\nu_t)}{\partial t}) = div(\nu_t \nabla f_{\nu_t, \nu^*})$$

  Can be obtained as the limit when $\tau \to 0$ of the **JKO scheme** [Jordan et al., 1998] :

$$\nu(n+1) = \operatorname*{argmin}_{\nu \in \mathcal{P}} \mathcal{F}(\nu) + \frac{1}{2\tau} W_2^2(\nu, \nu(n))$$

- Density of particles following a Mc-Kean Vlasov dynamic :

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

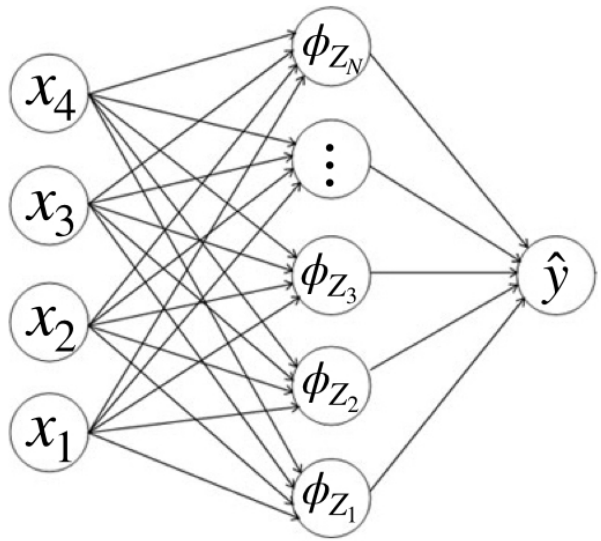
  where $\nabla_{Z_t} f_{\nu_t, \nu^*} = \int \nabla k(Z, Z_t) d\nu_t(Z) - \int \nabla k(Z, Z_t) d\nu^*(Z)$.

Example : Student-Teacher network

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

- the output of the Teacher network is deterministic and given by $y = \int \phi_Z(x) d\nu^*(Z)$ where $\nu^* = \frac{1}{M}\sum_{m=1}^{M} \delta_{U^m}$
- Student network parametrized by $\nu_0 = \frac{1}{N}\sum_{n=1}^{N} \delta_{Z_0^n}$ tries to learn the mapping $x \mapsto \int \phi_Z(x) d\nu^*(Z)$.

$(x, y) \sim data$



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

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:

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

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

Outline

Convergence of the MMD GF - Convexity

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

Convergence of the MMD GF - Convexity

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

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

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

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

Convergence of the MMD GF - Convexity

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

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

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

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

Convergence of the MMD GF - Convexity

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

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

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

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

- If $\mathcal{F}$ is λ-convex with $\lambda > 0$, all gradient flows of $\mathcal{F}$ converge to the **unique** minimizer of $\mathcal{F}$ [Carrillo et al., 2006]

Convergence of the MMD GF - Convexity

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

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

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

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

- If $\mathcal{F}$ is λ-convex with $\lambda > 0$, all gradient flows of $\mathcal{F}$ converge to the **unique** minimizer of $\mathcal{F}$ [Carrillo et al., 2006]

Our finding: The MMD is not λ-convex with $\lambda > 0$ in general.

Convergence of the MMD GF - a Lojasiewicz inequality

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

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

Convergence of the MMD GF - a Lojasiewicz inequality

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

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

- on the right, it's the RKHS norm:

$$\mathcal{F}(\nu_t) = \frac{1}{2}\|f_{\nu_t,\nu^*}\|^2_{\mathcal{H}}$$

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

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

Since:

$$\frac{\partial \nu_t}{\partial t} = div(\nu_t \nabla f_{\nu_t,\nu^*})$$

(dissipation of the MMD along the MMD flow)

A criterion for convergence

It can be shown that:

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

where on the r.h.s. we have the dual norm of $H^{\cdot}(\nu_t)$:

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

A criterion for convergence

It can be shown that:

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

where on the r.h.s. we have the dual norm of $H^{\cdot}(\nu_t)$:

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

A criterion for convergence

It can be shown that:

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

where on the r.h.s. we have the dual norm of $H^{\cdot}(\nu_t)$:

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

Assume that $\|\nu_t - \nu^*\|_{\dot{H}^{-1}(\nu_t)} \leq C$ **for all** t, then

$$MMD^2(\nu_t, \nu^*) \leq \frac{1}{MMD^2(\nu_0, \nu*) + 4C^{-1}t}$$

A criterion for convergence

It can be shown that:

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

where on the r.h.s. we have the dual norm of $H^{\cdot}(\nu_t)$:

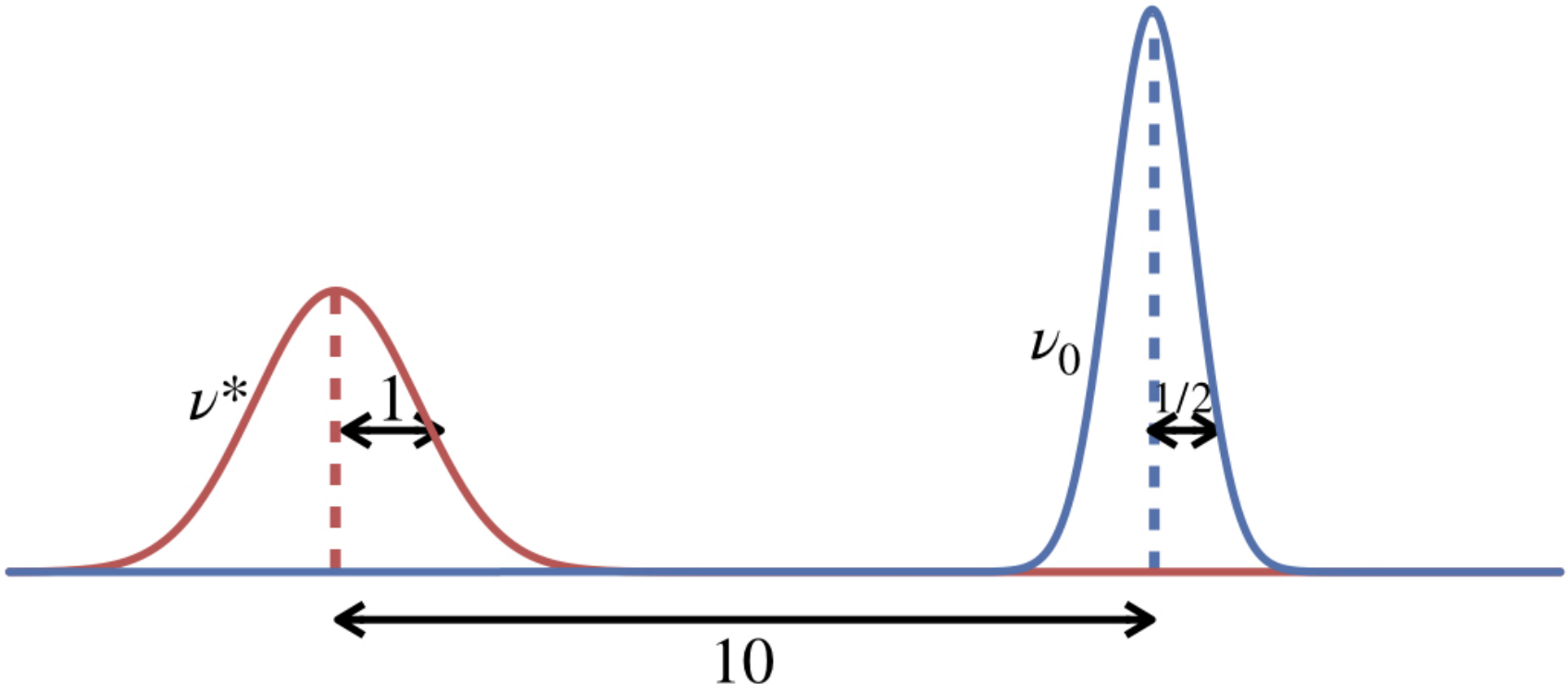
$$\|\nu_t - \nu^*\|_{\dot{H}^{-1}(\nu_t)} = \sup_{g,\ \mathbb{E}_{Z\sim\nu_t}[\|\nabla g(Z)\|^2]\leq 1} |\mathbb{E}_{Z\sim\nu_t}[g(Z)] - \mathbb{E}_{U\sim\nu^*}[g(U)]|$$

Assume that $\|\nu_t - \nu^*\|_{\dot{H}^{-1}(\nu_t)} \leq C$ **for all** t, then

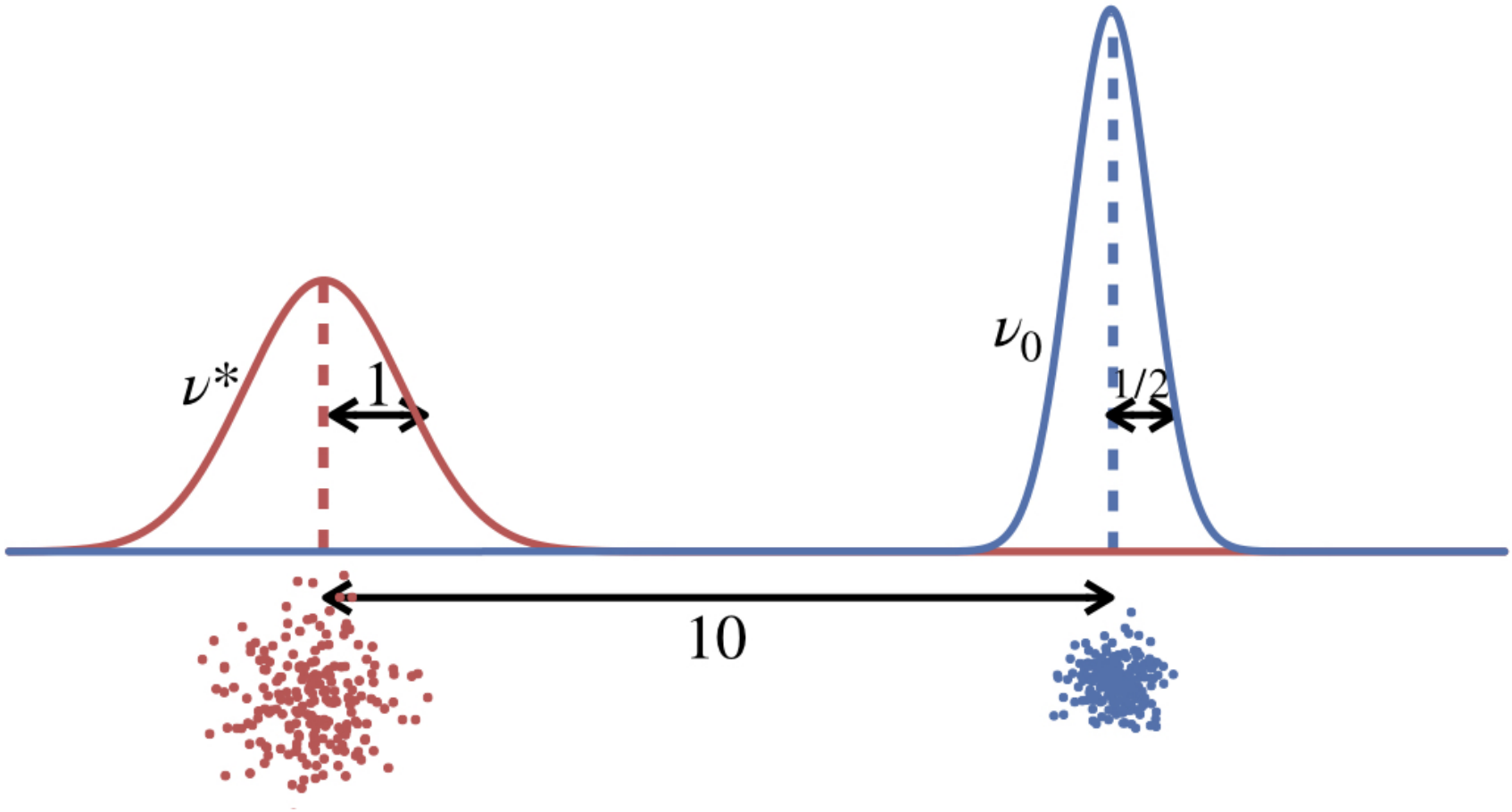
$$MMD^2(\nu_t, \nu^*) \leq \frac{1}{MMD^2(\nu_0, \nu*) + 4C^{-1}t}$$

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

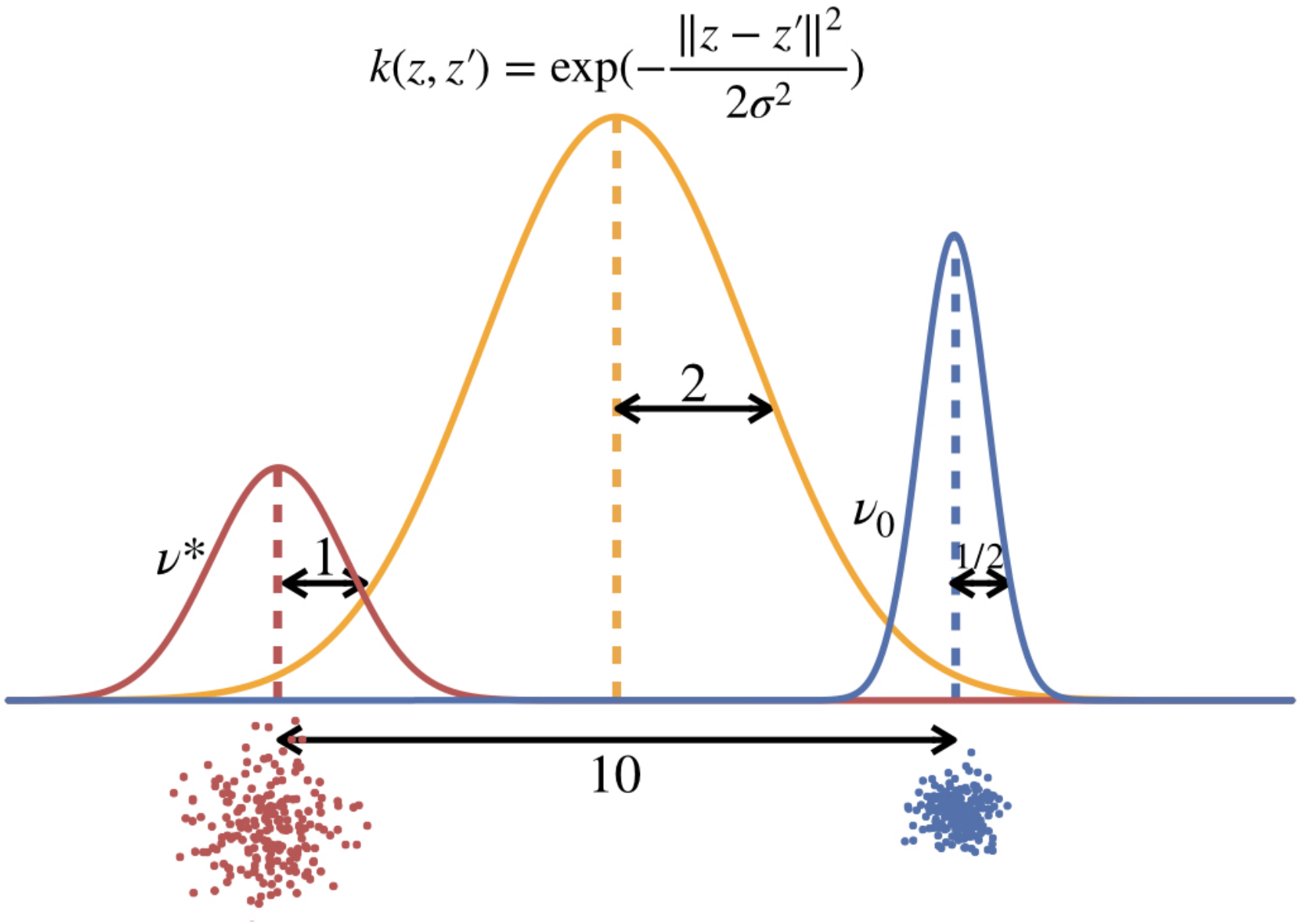
Convergence: Failure case



Convergence: Failure case



Convergence: Failure case



Convergence: Failure case

Convergence issues

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

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

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

If ν^∞ positive everywhere this implies $f_{\nu^\infty,\nu^*} = cte = 0$[1]

But ν^∞ might be singular...

- **Idea :** Evaluate $\nabla f_{\nu_t,\nu^*}$ outside of the support of ν_t to get a better signal!

[1] as soon as the RKHS $\mathcal{H}$ does not contain non-zero constant functions, e.g. for a gaussian kernel.

Outline

Noise injection

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

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

[2] [Duchi et al., 2012]

[3] [Mei et al., 2018]

Noise injection

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

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

- Similar to *randomized smoothing*[2], but extended to interacting particles.

[2] [Duchi et al., 2012]

[3] [Mei et al., 2018]

Noise injection

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

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

- Similar to *randomized smoothing*[2], but extended to interacting particles.
- Different from adding noise outside ("diffusion")

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

which corresponds to an entropic regularization of the original loss [3].

[2] [Duchi et al., 2012]

[3] [Mei et al., 2018]

Noise Injection: Theory (discrete time)

Tradeoff for the level of noise β_n

- Too large β_n: ν_{n+1} not a descent direction anymore: $\mathcal{F}(\nu_{n+1}, \nu^*) > \mathcal{F}^2(\nu_n, \nu^*)$

$$\implies \beta_n^2 \mathcal{F}^2(\nu_n) \leq C_k \mathbb{E}_{\substack{Z_n \sim \nu_n \\ u_n \sim \mathcal{N}(0,1)}} [\|\nabla f_{\nu_n, \nu^*}(Z_n + \beta_n \xi_n)\|^2] \quad (1)$$

Noise Injection: Theory (discrete time)

Tradeoff for the level of noise β_n

- Too large β_n: ν_{n+1} not a descent direction anymore: $\mathcal{F}(\nu_{n+1}, \nu^*) > \mathcal{F}^2(\nu_n, \nu^*)$

$$\implies \beta_n^2 \mathcal{F}^2(\nu_n) \leq C_k \mathbb{E}_{\substack{Z_n \sim \nu_n \\ u_n \sim \mathcal{N}(0,1)}} [\|\nabla f_{\nu_n, \nu^*}(Z_n + \beta_n \xi_n)\|^2] \quad (1)$$

- Too small β_n: Back to the failure mode: $\nabla f_{\nu_n, \nu^*}(Z_n + \beta_n u_n) \simeq 0$.

$$\implies \sum_{n=1}^{N} \beta_n^2 \to \infty \quad (2)$$

Under (1) and (2) :

$$\mathcal{F}^2(\nu_N, \nu^*) \leq \mathcal{F}^2(\nu_0, \nu^*) e^{-C_k \gamma (1 - \gamma C_k') \sum_{n=1}^{N} \beta_n^2}$$

Noise Injection: Experiments (Gaussians)

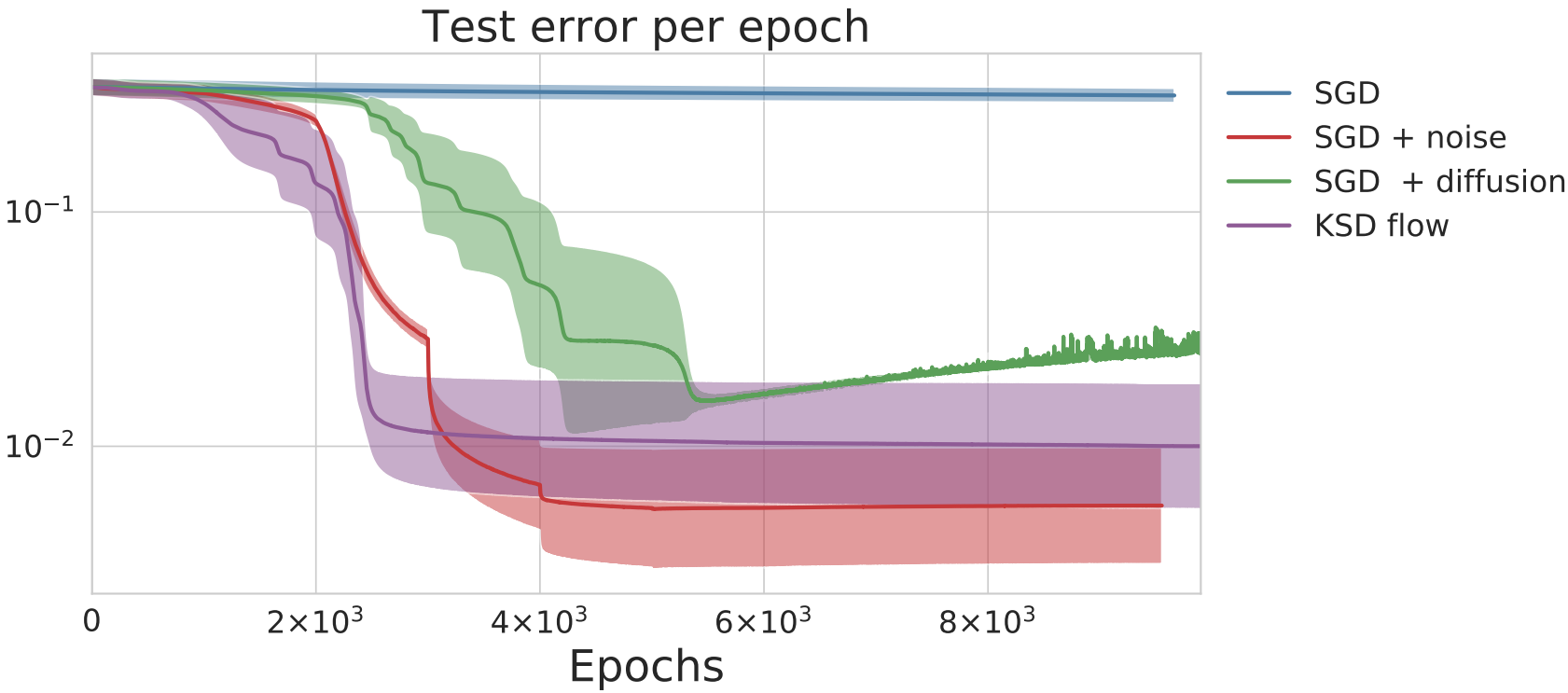
Noise Injection: Experiments (Student-Teacher)

Methods:

- SGD
- SGD + Noise injection
- SGD + diffusion
- KSD [4]: SGD using a (regularized) Negative Sobolev distance as a loss function; also decreases the MMD.

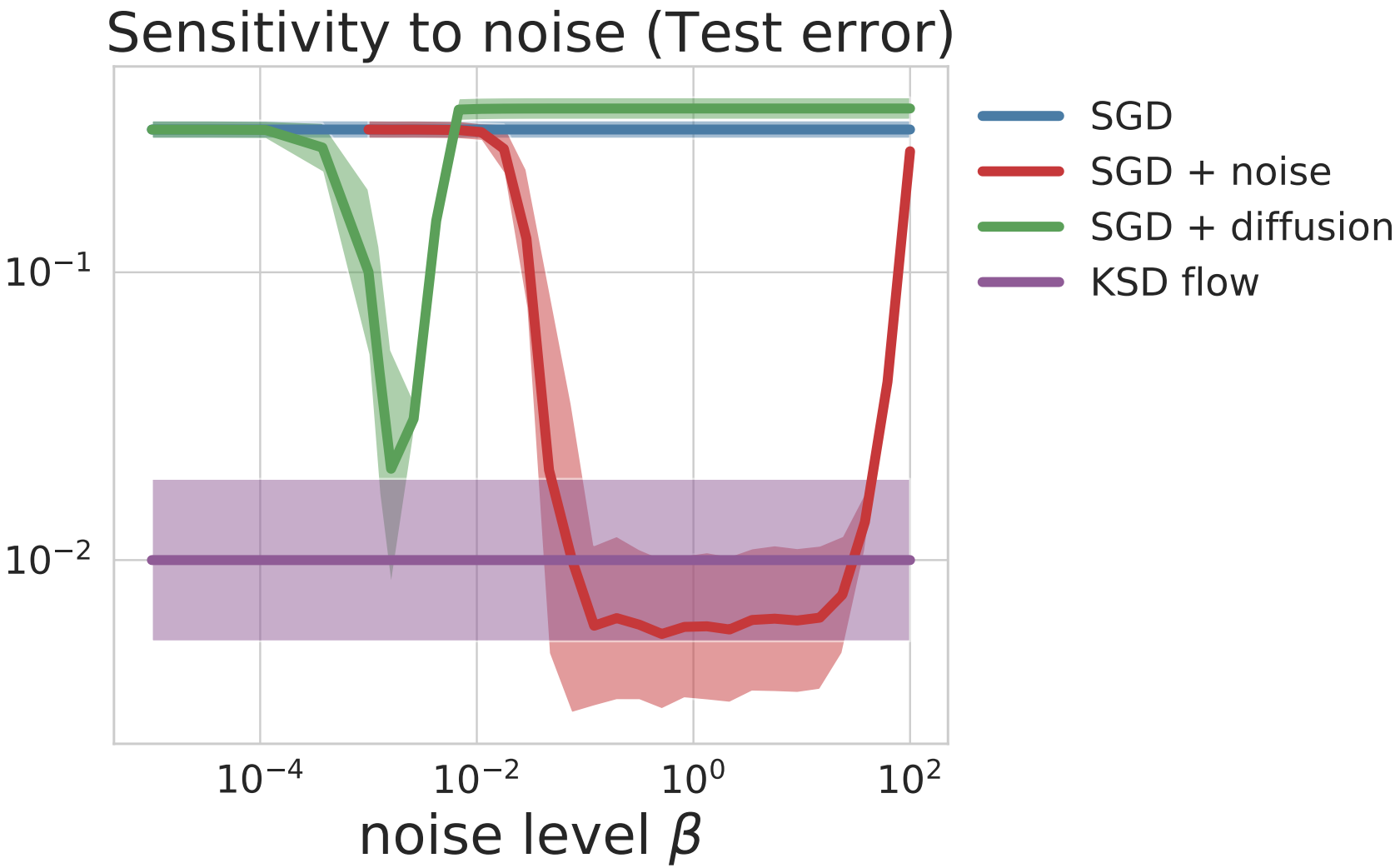
[4] "Kernel Sobolev Descent" [Mroueh et al., 2019]

Noise Injection: Experiments



dimension = 50

Noise Injection: Experiments



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.


Carrillo, J. A., McCann, R. J., and Villani, C. (2006).
Contractions in the 2-wasserstein length space and thermalization of granular media.
Archive for Rational Mechanics and Analysis, 179(2):217–263.


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


Duchi, J. C., Bartlett, P. L., and Wainwright, M. J. (2012).
Randomized smoothing for stochastic optimization.
SIAM Journal on Optimization, 22(2):674–701.


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


Jordan, R., Kinderlehrer, D., and Otto, F. (1998).
The variational formulation of the fokker–planck equation.