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

Séminaire de Statistique - LPSM
March 3, 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
- **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.

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
- **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.

1. Background and motivation
2. Wasserstein gradient flow of the MMD
3. Convergence properties
4. A noise-injection algorithm for better convergence

Outline

Background and motivation

Wasserstein gradient flow of the MMD

Convergence properties of the MMD gradient flow

A noise-injection algorithm for better convergence

Reproducing Kernel Hilbert Spaces (RKHS)

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

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

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

Reproducing Kernel Hilbert Spaces (RKHS)

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

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

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

Let $\mathcal{P}$ the set of probability distributions on $\mathcal{Z}$ with finite second moment. Suppose k is characteristic, ie the map:

$$\begin{aligned} \mathcal{P} &\to \mathcal{H} \\ \nu &\mapsto \underbrace{\int_{\mathcal{Z}} k(z, .) d\nu(z)}_{\textit{"mean embedding of } \nu \textit{"}} \end{aligned}$$

is injective.

Maximum Mean Discrepancy (MMD)

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

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

$$f_{\mu,\nu}(.) = \int k(z, .)d\mu(z) - \int k(z, .)d\nu(z)$$

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

Maximum Mean Discrepancy (MMD)

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

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

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

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

$$\begin{aligned} \mathcal{L} : \quad & \mathcal{P} \to \mathbb{R} \\ & \nu \mapsto \frac{1}{2} MMD^2(\mu,\nu) \end{aligned}$$

MMD functional

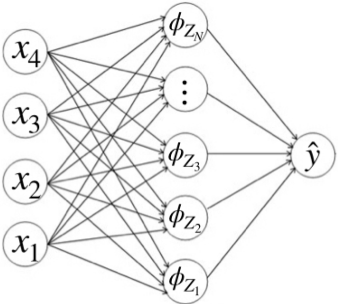
For a target distribution μ (fixed), for any $\nu \in \mathcal{P}$:

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

General problem

Consider the following regression problem:

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

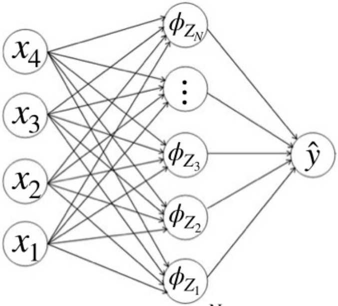


$$\min_{Z_1,\ldots,Z_N} \mathbb{E}_{data}[\|y - \frac{1}{N}\sum_{i=1}^{N} \phi_{Z_i}(x)\|^2]$$

General problem

Consider the following regression problem:

$(x, y) \sim data$



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

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

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

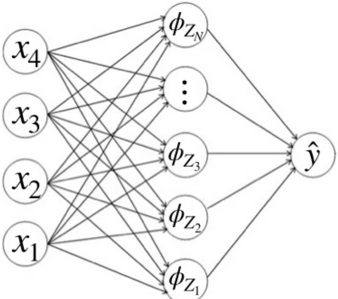
  where $g : \mathbb{R} \to \mathbb{R}$
  (sigmoid
  $g(z) = 1/(1 + e^{-z})$, RelU
  $(g(z) = max(0, z)$...)

General problem

Finite dimensional non-convex optimization (regression setting):

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

$(x, y) \sim data$



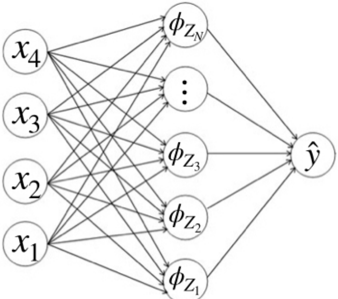
$$\min_{Z_1,\ldots,Z_N} \mathbb{E}_{data}[\|y - \frac{1}{N}\sum_{i=1}^{N} \phi_{Z_i}(x)\|^2]$$

General problem

Finite dimensional non-convex optimization (regression setting):

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

$(x, y) \sim data$



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

- Optimization using gradient descent (GD):

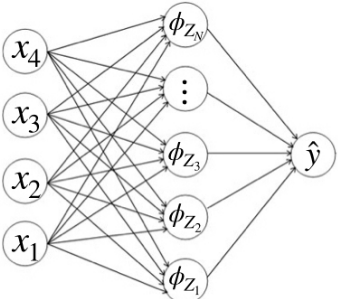
$$Z_i^{t+1} = Z_i^t - \gamma \nabla_{Z_i} \mathcal{L}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i^t}\right)$$

General problem

Finite dimensional non-convex optimization (regression setting):

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

$(x, y) \sim data$



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

- Optimization using gradient descent (GD):

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

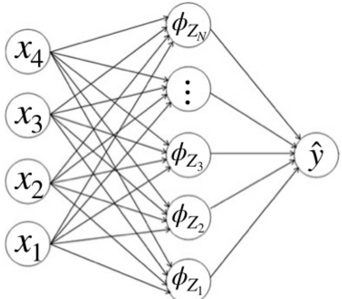
- Hard to describe the dynamics of GD!

General problem

Finite dimensional non-convex optimization (regression setting):

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

$(x, y) \sim data$



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

- Optimization using gradient descent (GD):

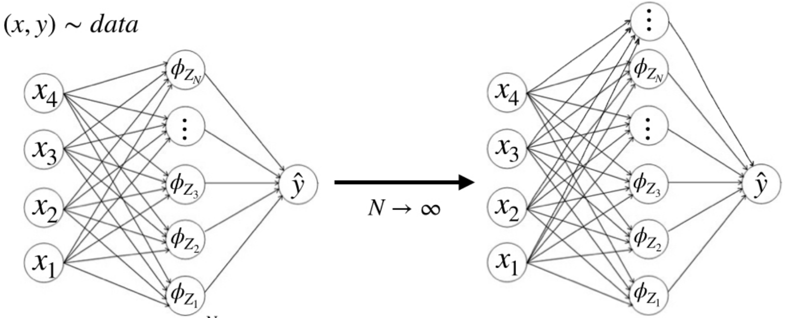
$$Z_i^{t+1} = Z_i^t - \gamma \nabla_{Z_i} \mathcal{L}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i^t}\right)$$

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

General problem

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

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

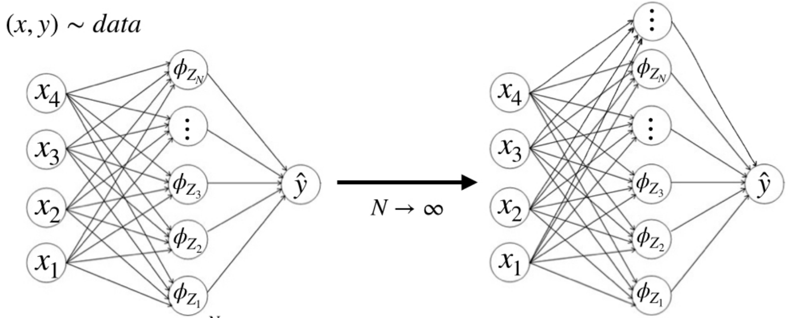


$$\min_{Z_1,\ldots,Z_N} \mathbb{E}_{data}[\|y - \frac{1}{N}\sum_{i=1}^{N} \phi_{Z_i}(x)\|^2] \xrightarrow[N\to\infty]{} \min_{\nu \in \mathscr{P}} \mathbb{E}_{data}[\|y - \mathbb{E}_{Z\sim\nu}[\phi_Z(x)]\|^2]$$

General problem

- Global convergence of Gradient descent[1] when $N \to \infty$ and $\phi_Z(x)$ of the form:

$$\phi_Z(x) = wg(x, \theta), \qquad Z = (w, \theta)$$



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

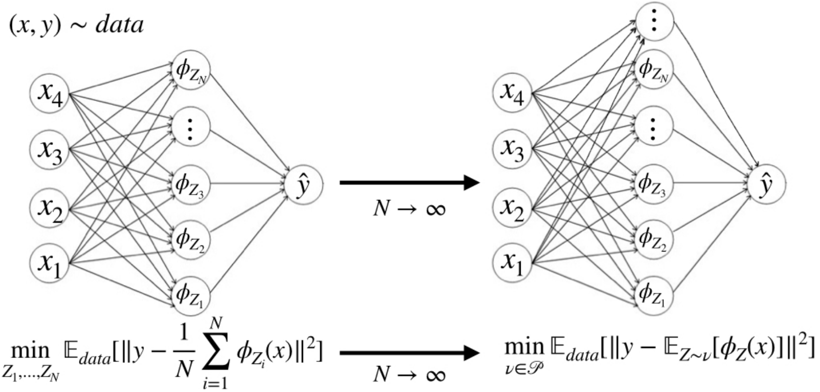
[1] [Rotskoff and Vanden-Eijnden, 2018, Chizat and Bach, 2018]

General problem

- Global convergence of Gradient descent[1] when $N \to \infty$ and $\phi_Z(x)$ of the form:

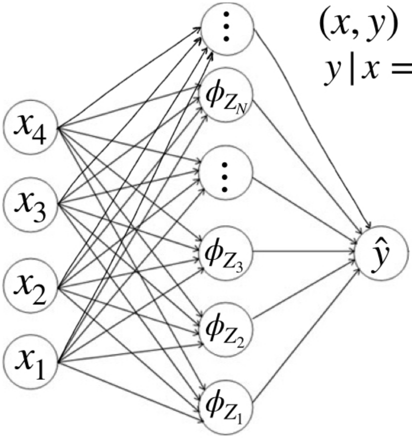
$$\phi_Z(x) = wg(x, \theta), \qquad Z = (w, \theta)$$

- Interested in more general form for $\phi_Z(x)$.



[1] [Rotskoff and Vanden-Eijnden, 2018, Chizat and Bach, 2018]

Minimization of the MMD: the well-specified case



$(x, y) \sim data$

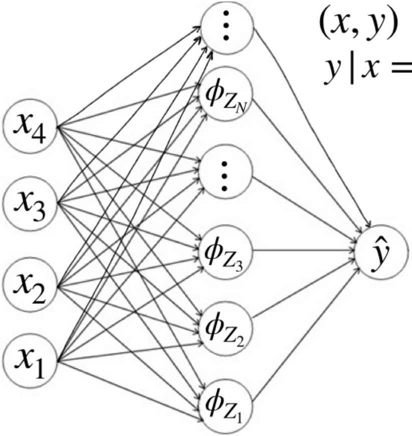
$y \,|\, x = \mathbb{E}_{U \sim \nu^*}[\phi_U(x)]$

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

Minimization of the MMD: the well-specified case

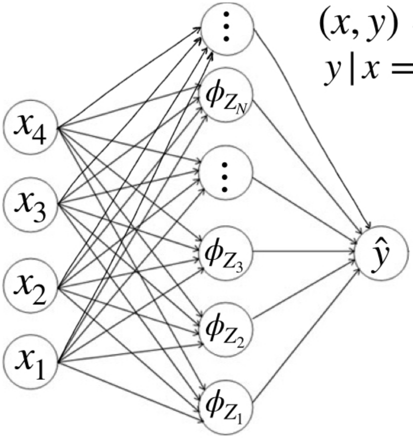
$$(x, y) \sim data$$

$$y \,|\, x = \mathbb{E}_{U\sim\nu^*}[\phi_U(x)]$$



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

Minimization of the MMD: the well-specified case

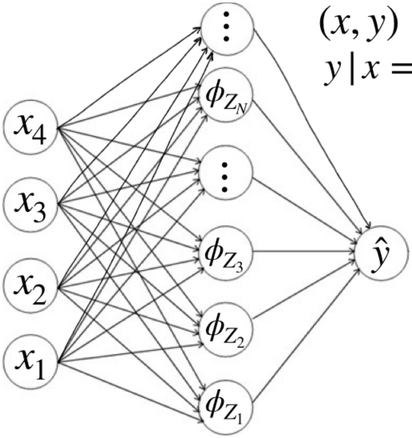


$(x, y) \sim data$

$y \,|\, x = \mathbb{E}_{U \sim \nu^*}[\phi_U(x)]$

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

Minimization of the MMD: the well-specified case



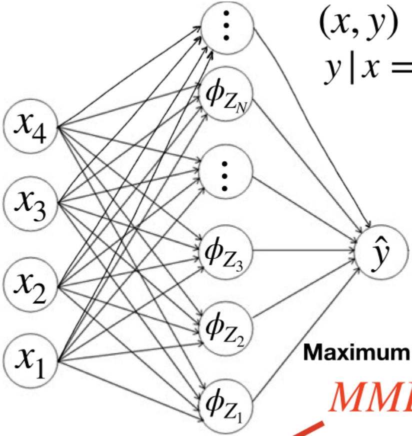
$(x, y) \sim data$

$y \,|\, x = \mathbb{E}_{U \sim \nu^*}[\phi_U(x)]$

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

$$k(Z, Z') = \mathbb{E}_{data}[\phi_Z(x)\phi_{Z'}(x)]$$

Minimization of the MMD: the well-specified case



$(x, y) \sim data$

$y \,|\, x = \mathbb{E}_{U\sim\nu^*}[\phi_U(x)]$

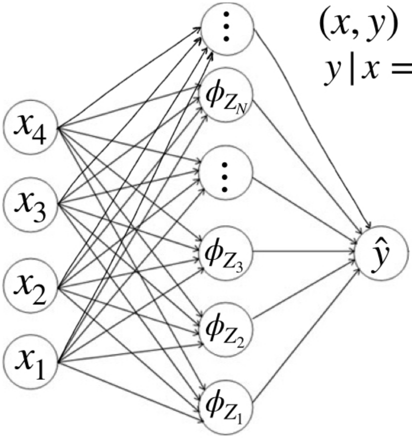
Maximum Mean Discrepancy

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

$$k(Z,Z') = \mathbb{E}_{data}[\phi_Z(x)\phi_{Z'}(x)]$$

[Gretton et al., 2012b]

Minimization of the MMD: the well-specified case

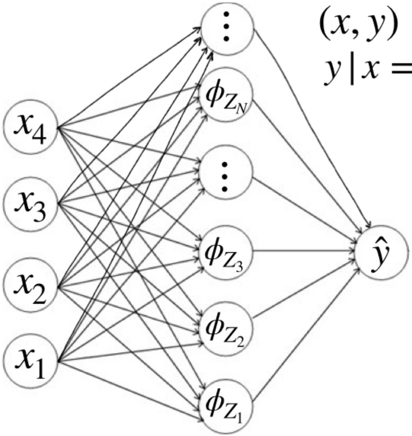


$$(x, y) \sim data$$
$$y \,|\, x = \mathbb{E}_{U \sim \nu^*}[\phi_U(x)]$$

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

$$k(Z, Z') = \mathbb{E}_{data}[\phi_Z(x)\phi_{Z'}(x)]$$

Minimization of the MMD: the well-specified case



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

$$y \,|\, x = \mathbb{E}_{U \sim \nu^*}[\phi_U(x)]$$

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

$$k(Z, Z') = \mathbb{E}_{data}[\phi_Z(x)\phi_{Z'}(x)]$$

MMD Gradient Flow

- The optimization over the parameters of a neural network can be seen as a minimization of the MMD on $\mathcal{P}$ in the population limit ($N \to \infty$).

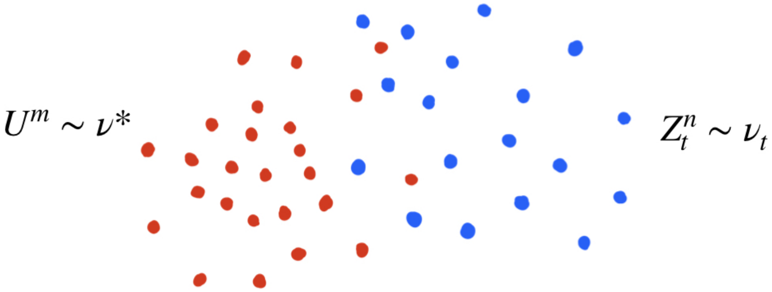
$$\min_{\nu} MMD^2(\nu, \nu^*)$$

$$\nu_{t+1} \approx \nu_t - \gamma \nabla_{\nu_t} MMD^2(\nu_t, \nu^*)$$

- Gradient descent dynamics in this setting takes the form of a PDE (gradient flow on $\mathcal{P}$)
- Powerful tool to analyze dynamics and possibly alterate them

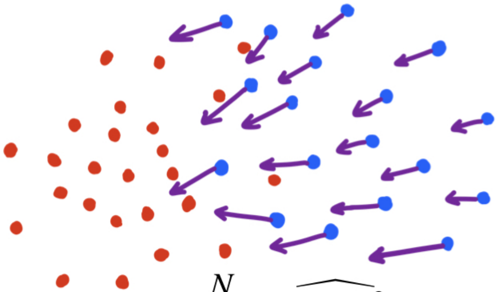
Outline

Gradient descent of the MMD



Gradient descent of the MMD

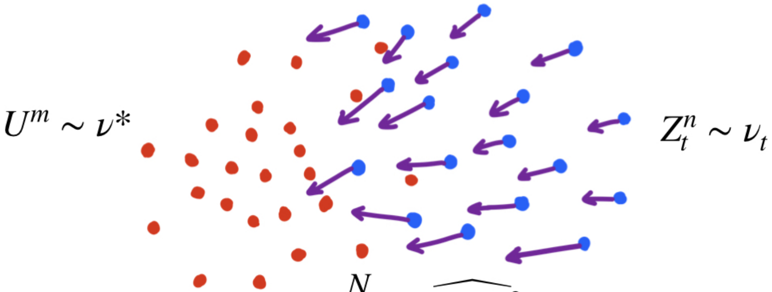
$U^m \sim \nu^*$



$Z_t^n \sim \nu_t$

$$Z_{t+1}^n = Z_t^n - \gamma \frac{N}{2} \nabla_{Z_t^n} \widehat{MMD}^2(\nu^*, \nu_t)$$

Gradient descent of the MMD



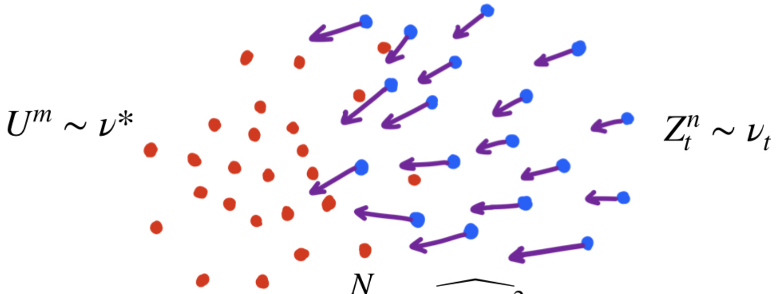
$$Z_{t+1}^n = Z_t^n - \gamma \frac{N}{2} \nabla_{Z_t^n} \widehat{MMD}^2(\nu^*, \nu_t)$$

$$\frac{N}{2} \nabla_{Z_t^n} \widehat{MMD}^2(\nu^*, \nu_t) = \frac{1}{N-1} \sum_{n' \neq n} \partial_1 k(Z_t^n, Z_t^{n'}) - \frac{1}{M} \sum_{m=1}^{M} \partial_1 k(Z_t^n, U^m)$$

$$\Big\downarrow N, M \to \infty$$

$$\nabla_{Z_t} \Big(\mathbb{E}_{Z' \sim \nu_t}[k(Z_t, Z')] - \mathbb{E}_{U \sim \nu^*}[k(Z_t, U)] \Big)$$

Gradient descent of the MMD



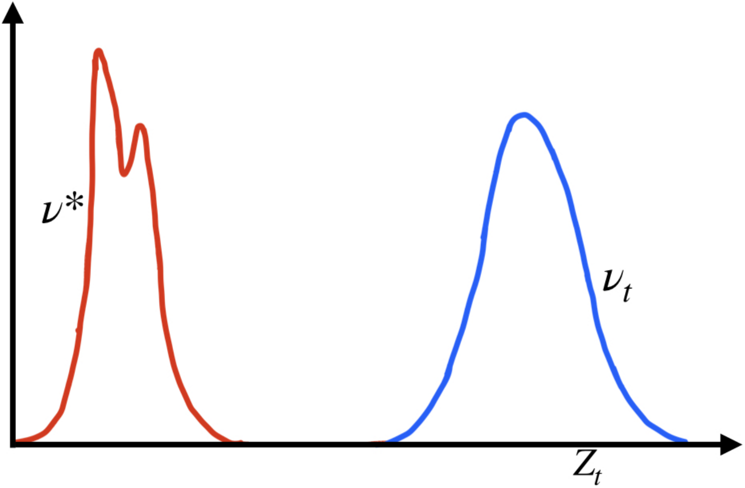
$$Z_{t+1}^n = Z_t^n - \gamma \frac{N}{2} \nabla_{Z_t^n} \widehat{MMD}^2(\nu^*, \nu_t)$$

$$\frac{N}{2} \nabla_{Z_t^n} \widehat{MMD}^2(\nu^*, \nu_t) = \frac{1}{N-1} \sum_{n' \neq n} \partial_1 k(Z_t^n, Z_t^{n'}) - \frac{1}{M} \sum_{m=1}^{M} \partial_1 k(Z_t^n, U^m)$$

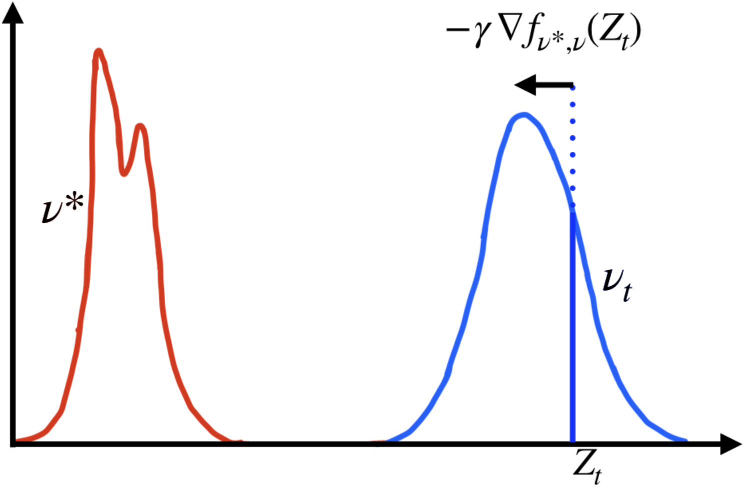
$$\Big\downarrow N, M \to \infty$$

$$\nabla_{Z_t} \Big(\underbrace{\mathbb{E}_{Z' \sim \nu_t}[k(Z_t, Z')] - \mathbb{E}_{U \sim \nu^*}[k(Z_t, U)]}_{f_{\nu^*, \nu_t}(Z_t)} \Big)$$

Wasserstein gradient descent

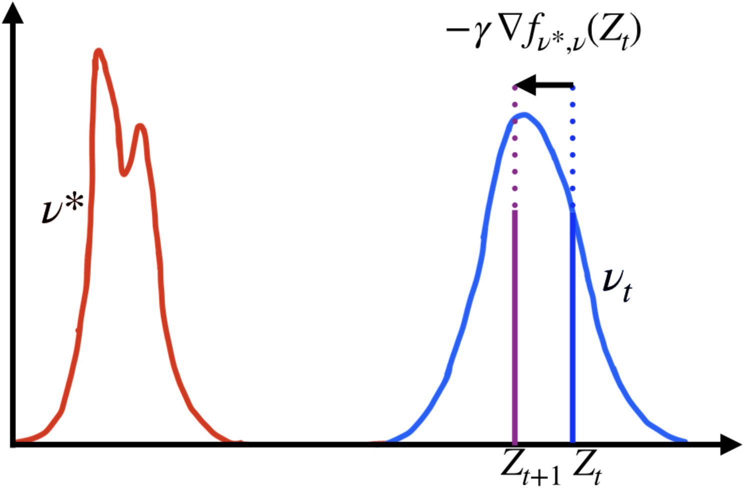


Wasserstein gradient descent



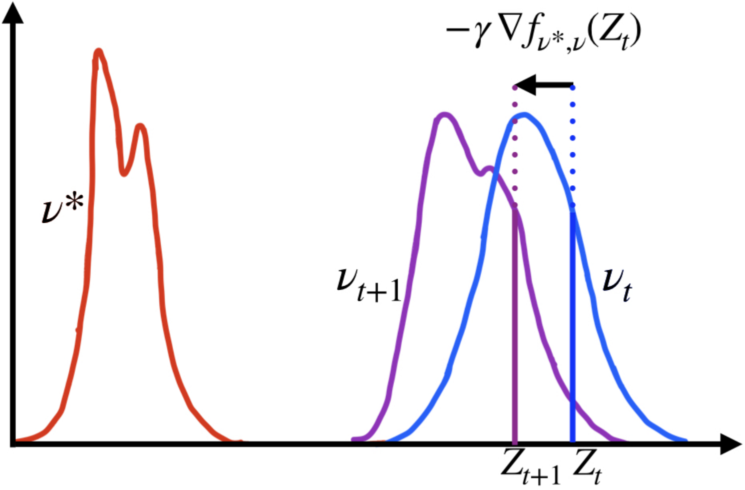
$$f_{\nu^*,\nu_t}(Z_t) = \mathbb{E}_{Z' \sim \nu_t}[k(Z_t, Z')] - \mathbb{E}_{U \sim \nu^*}[k(Z_t, U)]$$

Wasserstein gradient descent



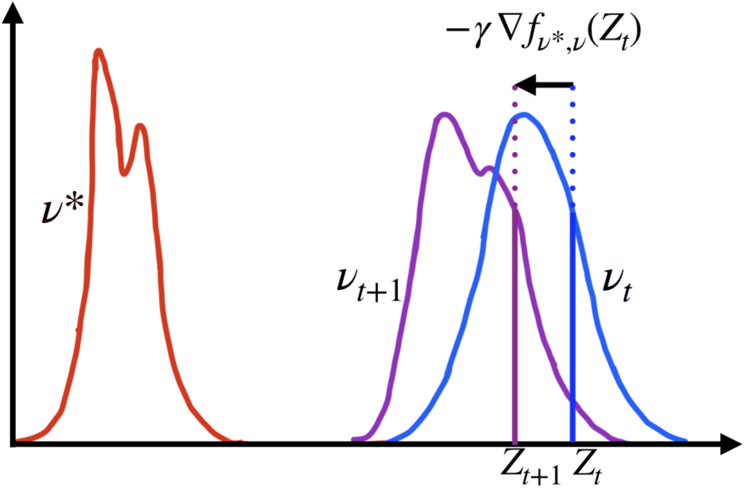
$$f_{\nu^*,\nu_t}(Z_t) = \mathbb{E}_{Z' \sim \nu_t}[k(Z_t, Z')] - \mathbb{E}_{U \sim \nu^*}[k(Z_t, U)]$$

Wasserstein gradient descent



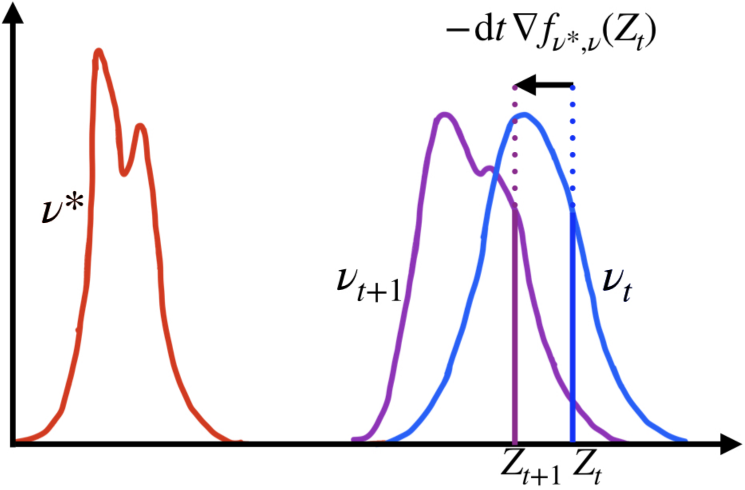
$$f_{\nu^*,\nu_t}(Z_t) = \mathbb{E}_{Z' \sim \nu_t}[k(Z_t, Z')] - \mathbb{E}_{U \sim \nu^*}[k(Z_t, U)]$$

Wasserstein gradient descent



$$Z_{t+1} = Z_t - \gamma \nabla_{Z_t} f_{\nu^*,\nu_t}(Z_t), \qquad Z_t \sim \nu_t$$

Wasserstein gradient descent



$$\mathrm{d}Z_t = -\,\mathrm{d}t\,\nabla_{Z_t} f_{\nu^*,\nu_t}(Z_t), \qquad Z_t \sim \nu_t$$

Wasserstein gradient flow

- Continuous time equation: Mc-Kean Vlasov dynamics [2]

$$\frac{\mathrm{d}Z_t}{\mathrm{d}t} = -\nabla_{Z_t} f_{\nu^*,\nu_t}(Z_t), \qquad Z_t \sim \nu_t$$

[2] [Kac, 1956]

[3] [Otto, 2001, Villani, 2004, Ambrosio et al., 2004]

Wasserstein gradient flow

- Continuous time equation: Mc-Kean Vlasov dynamics [2]

$$\frac{\mathrm{d}Z_t}{\mathrm{d}t} = -\nabla_{Z_t} f_{\nu^*,\nu_t}(Z_t), \qquad Z_t \sim \nu_t$$

- Equivalent to a PDE in ν_t:

$$\partial_t \nu_t = div(\nu_t \nabla f_{\nu^*,\nu_t})$$

[2] [Kac, 1956]

[3] [Otto, 2001, Villani, 2004, Ambrosio et al., 2004]

Wasserstein gradient flow

- Continuous time equation: Mc-Kean Vlasov dynamics [2]

$$\frac{\mathrm{d}Z_t}{\mathrm{d}t} = -\nabla_{Z_t} f_{\nu^*,\nu_t}(Z_t), \qquad Z_t \sim \nu_t$$

- Equivalent to a PDE in ν_t:

$$\partial_t \nu_t = div(\nu_t \nabla f_{\nu^*,\nu_t})$$

- Interpretation as a gradient flow in probability space[3], a curve $\nu : [0, \infty] \to \mathcal{P}$:

$$\partial_t \nu_t = -\nabla_{\nu_t} \mathcal{L}(\nu_t) \qquad \mathcal{L}(\nu) := \frac{1}{2} MMD^2(\nu^*, \nu)$$

can be obtained as the limit when $\tau \to 0$ of:

$$\nu_{t+1} \in \arg\min_{\nu \in \mathcal{P}} \mathcal{L}(\nu) + \frac{1}{2\tau} W_2^2(\nu, \nu_t).$$

---

[2] [Kac, 1956]

[3] [Otto, 2001, Villani, 2004, Ambrosio et al., 2004]

Outline

Convergence of gradient flows and convexity

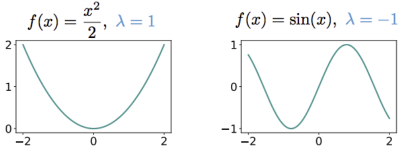
In a **euclidean** setting, a gradient flow of a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is a curve $x : [0, \infty] \to \mathbb{R}^d$:

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

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

A function f defined on $\mathbb{R}^d$ is λ-convex if for any $x, y \in \mathbb{R}^d$ and $t \in [0, 1]$:

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



When $\lambda > 0$: any gradient flow x_t converges to a unique $x^\star = \text{argmin}_x f(x)$ when $t \to \infty$.

Convergence of the MMD (W2-)gradient flow

In the **W2** setting, a gradient flow of a functional $\mathcal{L} : \mathcal{P} \to \mathbb{R}$ is a curve $\nu : [0, \infty] \to \mathcal{P}$ that satisfies:

$$\frac{d\nu_t}{dt} = -\nabla_{\nu_t}\mathcal{L}(\nu_t)$$

Existence, uniqueness results on gradient flows on $\mathcal{P}$ rely on the notion of **convexity**, wrt W2 geodesic curves.

Convergence of the MMD (W2-)gradient flow

In the **W2** setting, a gradient flow of a functional $\mathcal{L} : \mathcal{P} \to \mathbb{R}$ is a curve $\nu : [0, \infty] \to \mathcal{P}$ that satisfies:

$$\frac{d\nu_t}{dt} = -\nabla_{\nu_t}\mathcal{L}(\nu_t)$$

Existence, uniqueness results on gradient flows on $\mathcal{P}$ rely on the notion of **convexity**, wrt W2 geodesic curves.

A functional $\mathcal{L}$ is (λ)-geodesically convex if for any $t \in [0, 1]$:

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

where $d(\rho(0), \rho(1))^2 = W_2^2(\rho(0), \rho(1))$.

Convergence of the MMD (W2-)gradient flow

In the **W2** setting, a gradient flow of a functional $\mathcal{L} : \mathcal{P} \to \mathbb{R}$ is a curve $\nu : [0, \infty] \to \mathcal{P}$ that satisfies:

$$\frac{d\nu_t}{dt} = -\nabla_{\nu_t}\mathcal{L}(\nu_t)$$

Existence, uniqueness results on gradient flows on $\mathcal{P}$ rely on the notion of **convexity**, wrt W2 geodesic curves.

A functional $\mathcal{L}$ is (λ)-geodesically convex if for any $t \in [0, 1]$:

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

where $d(\rho(0), \rho(1))^2 = W_2^2(\rho(0), \rho(1))$.

Convergence of the MMD (W2-)gradient flow

In the **W2** setting, a gradient flow of a functional $\mathcal{L} : \mathcal{P} \to \mathbb{R}$ is a curve $\nu : [0, \infty] \to \mathcal{P}$ that satisfies:

$$\frac{d\nu_t}{dt} = -\nabla_{\nu_t}\mathcal{L}(\nu_t)$$

Existence, uniqueness results on gradient flows on $\mathcal{P}$ rely on the notion of **convexity**, wrt W2 geodesic curves.

A functional $\mathcal{L}$ is (λ)-geodesically convex if for any $t \in [0, 1]$:

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

where $d(\rho(0), \rho(1))^2 = W_2^2(\rho(0), \rho(1))$.

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

Too bad... $\lambda > 0$ would have guaranteed that all gradient flows of $\mathcal{L}$ would converge the **unique** minimizer of $\mathcal{L}$ [Carrillo et al., 2006]

Second strategy - Obtain a Lojasiewicz inequality

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

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

Second strategy - Obtain a Lojasiewicz inequality

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

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

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

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

Second strategy - Obtain a Lojasiewicz inequality

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

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

Second strategy - Obtain a Lojasiewicz inequality

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

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

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

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

Since:

$$\partial_t \nu_t = div(\nu_t \nabla f_{\nu^*,\nu_t})$$

A criterion for convergence

Define the weighted Negative Sobolev norm:

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

A criterion for convergence

Define the weighted Negative Sobolev norm:

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

It can be shown that:

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

A criterion for convergence

Define the weighted Negative Sobolev norm:

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

It can be shown that:

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

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

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

A criterion for convergence

Define the weighted Negative Sobolev norm:

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

It can be shown that:

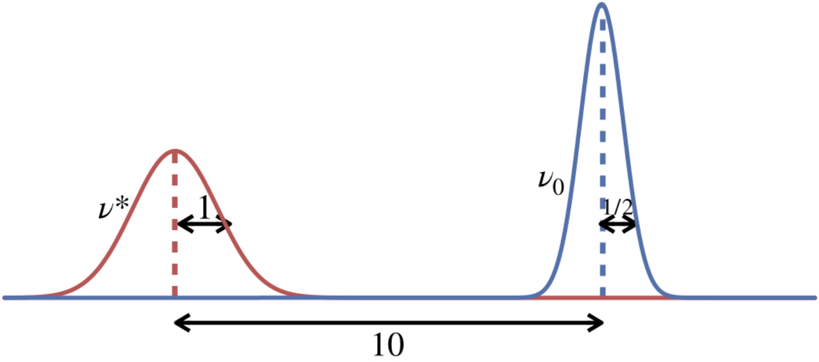
$$\|f_{\nu^*,\nu_t}\|^2_{\mathcal{H}} \leq \|f_{\nu^*,\nu_t}\|_{\dot{H}(\nu_t)}\|\nu^* - \nu_t\|_{\dot{H}^{-1}(\nu_t)}$$

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

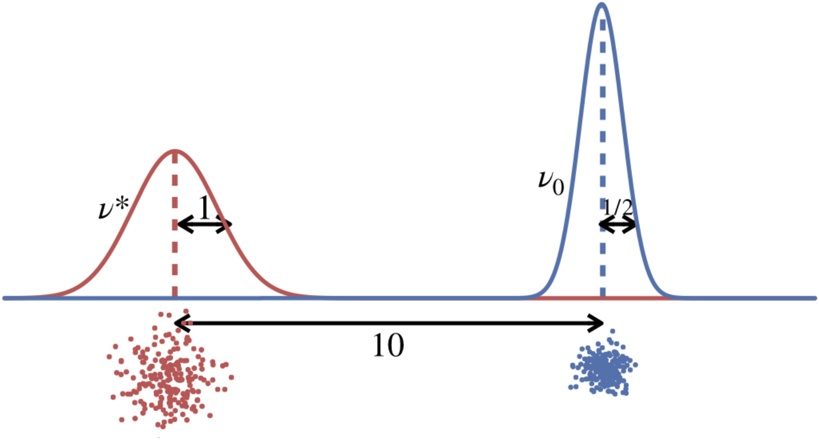
$$MMD^2(\nu^*, \nu_t) \leq \frac{1}{MMD^2(\nu^*, \nu_0) + 4C^{-1}t}$$

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

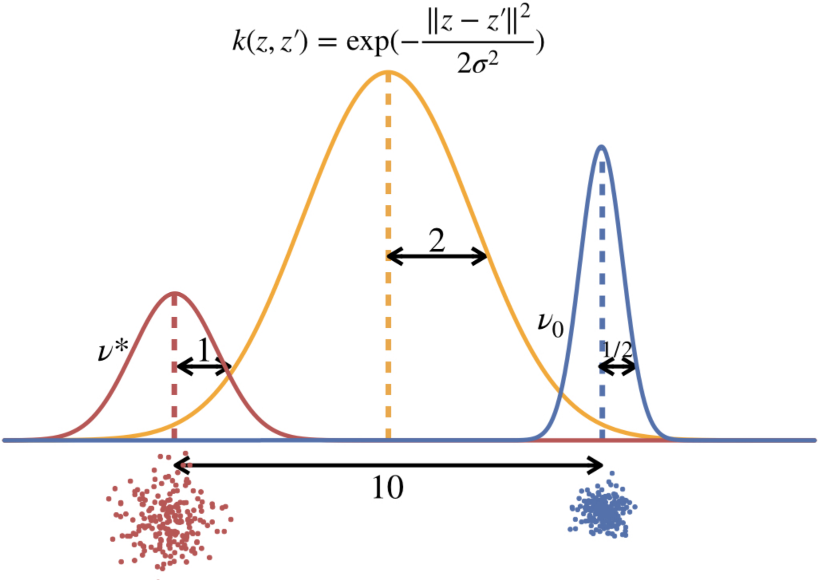
Convergence: Failure case



Convergence: Failure case



Convergence: Failure case



Convergence: Failure case

Outline

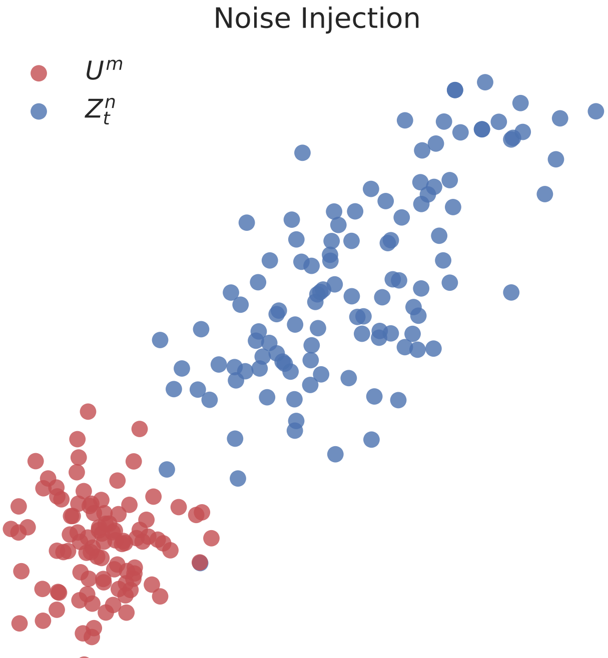
Background and motivation

Wasserstein gradient flow of the MMD

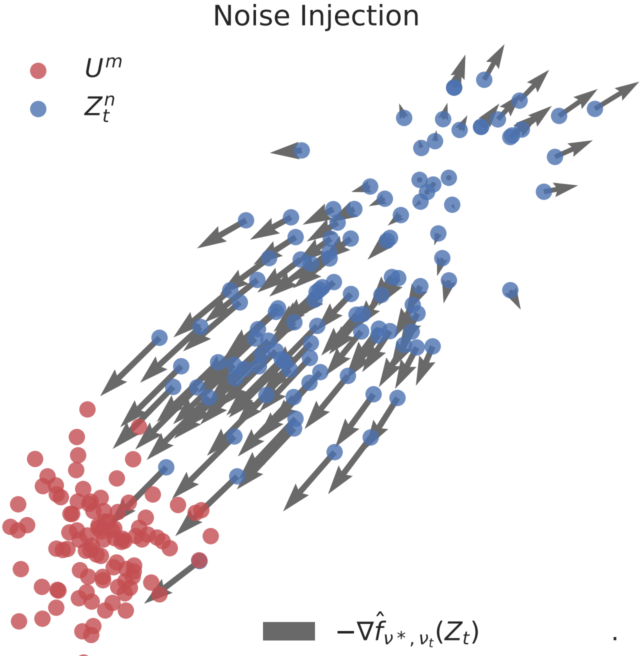
Convergence properties of the MMD gradient flow

A noise-injection algorithm for better convergence

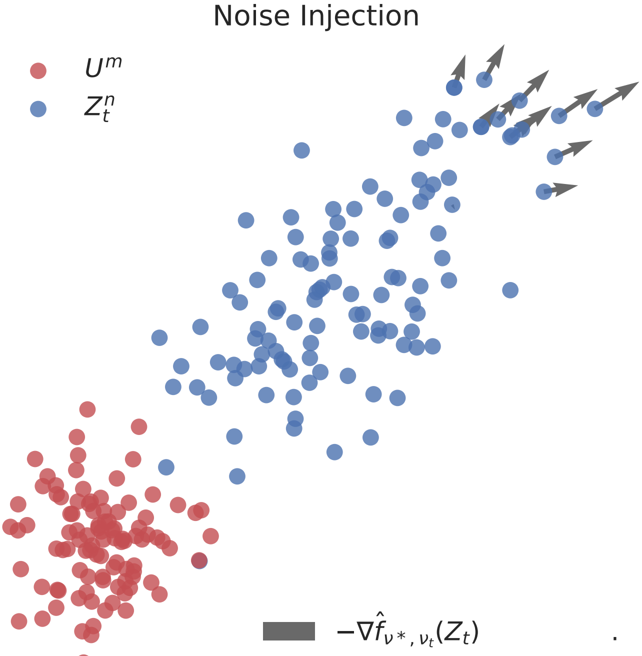
Noise Injection



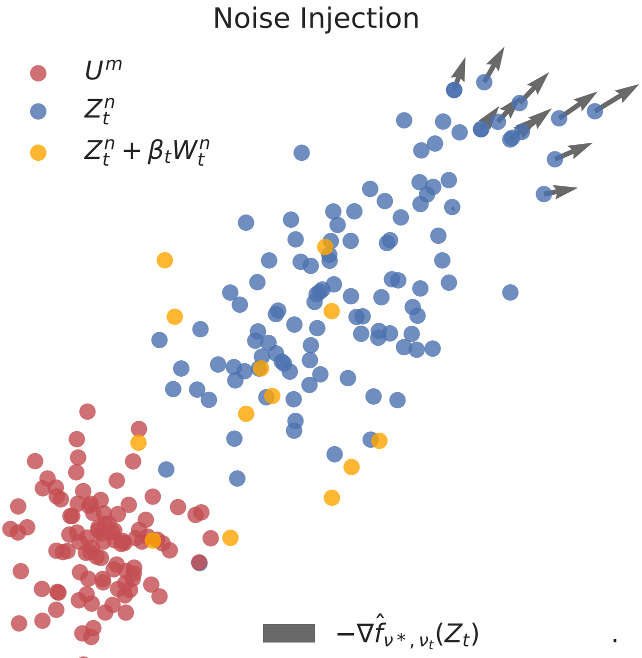
Noise Injection



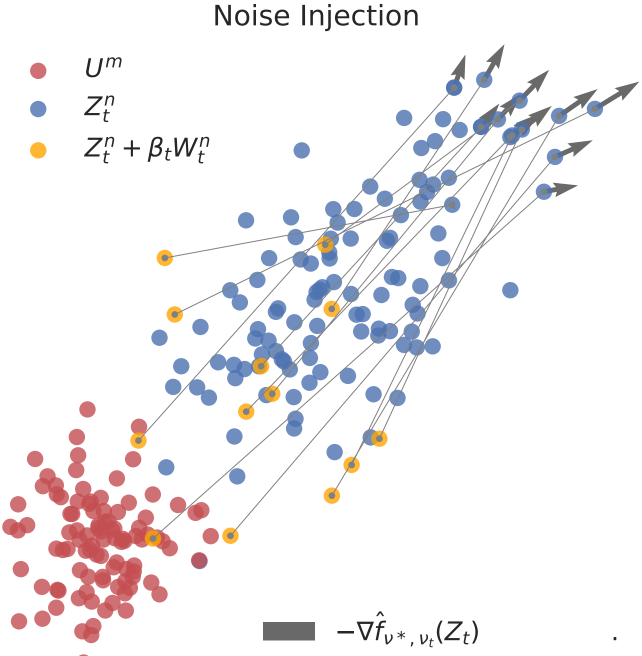
Noise Injection



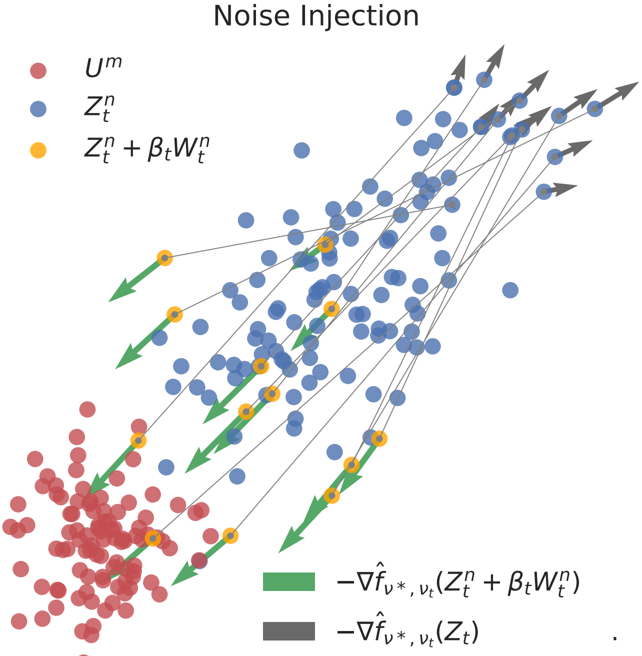
Noise Injection



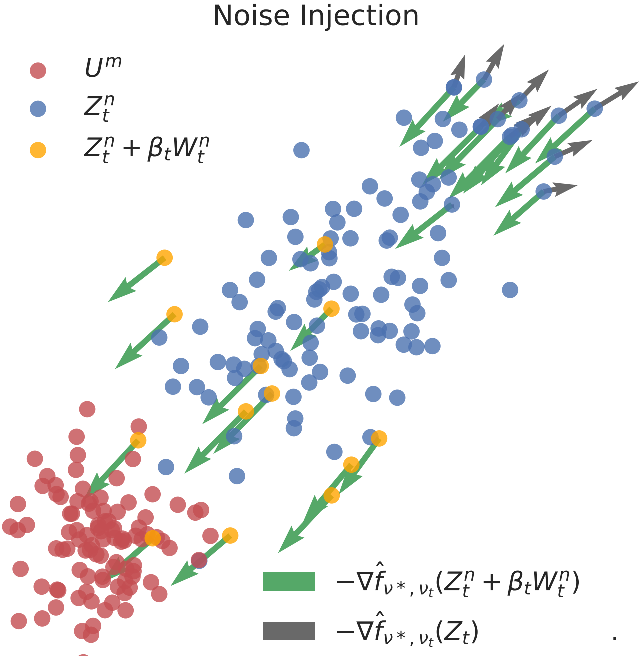
Noise Injection



Noise Injection

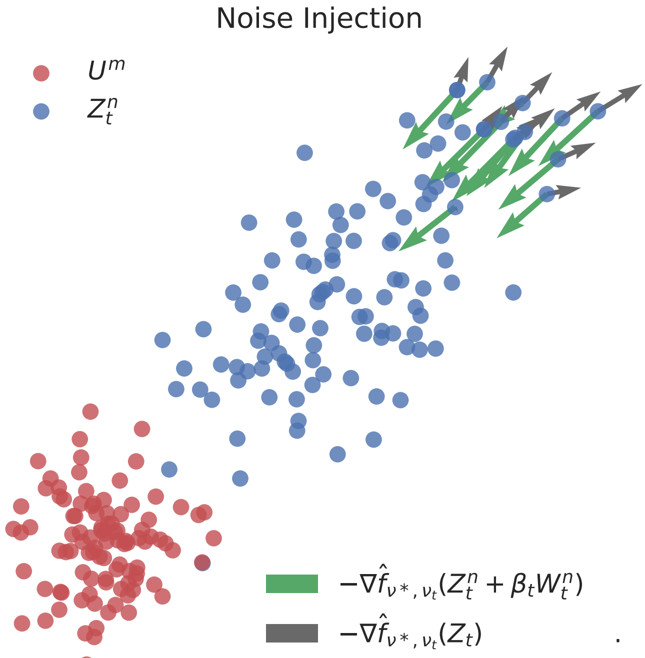


Noise Injection

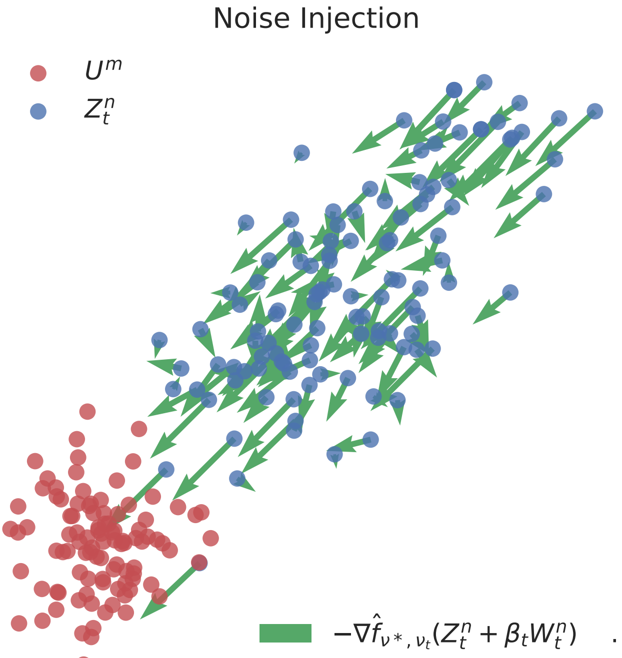


.

Noise Injection



Noise Injection



Noise Injection: Experiments

Noise Injection

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

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

  If ν^∞ positive everywhere this implies $f_{\nu^*,\nu^\infty} = cte = 0$ as soon as $\mathcal{H}$ does not contain non-zero constant functions. But ν^∞ might be singular...

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

Noise Injection

- Sample $u_t \sim \mathcal{N}(0, 1)$ and β_t is the noise level:

$$Z_{t+1} = Z_t - \gamma \nabla f_{\nu_t, \nu^*}(Z_t + \beta_t u_t); \qquad Z_t \sim \nu_t$$

[4] [Chaudhari et al., 2017, Hazan et al., 2016]
[5] [Duchi et al., 2012]
[6] [Mei et al., 2018]

Noise Injection

- Sample $u_t \sim \mathcal{N}(0,1)$ and β_t is the noise level:

$$Z_{t+1} = Z_t - \gamma \nabla f_{\nu_t,\nu^*}(Z_t + \beta_t u_t); \qquad Z_t \sim \nu_t$$

- Similar to *continuation methods*[4] or *randomized smoothing*[5] , but extended to interacting particles.

[4] [Chaudhari et al., 2017, Hazan et al., 2016]
[5] [Duchi et al., 2012]
[6] [Mei et al., 2018]

Noise Injection

- Sample $u_t \sim \mathcal{N}(0,1)$ and β_t is the noise level:

$$Z_{t+1} = Z_t - \gamma \nabla f_{\nu_t,\nu^*}(Z_t + \beta_t u_t); \qquad Z_t \sim \nu_t$$

- Similar to *continuation methods*[4] or *randomized smoothing*[5] , but extended to interacting particles.
- Different from adding noise outside ("diffusion")

$$Z_{t+1} = Z_t - \gamma \nabla f_{\nu^*,\nu_t}(Z_t) + \beta_t u_t$$

which corresponds to an entropic regularization of the original loss [6].

[4] [Chaudhari et al., 2017, Hazan et al., 2016]

[5] [Duchi et al., 2012]

[6] [Mei et al., 2018]

Noise Injection: Theory (discrete time)

Tradeoff for β_t

- Large β_t: ν_{t+1} not a descent direction anymore: $MMD^2(\nu^*, \nu_{t+1}) > MMD^2(\nu^*, \nu_t)$

Noise Injection: Theory (discrete time)

Tradeoff for β_t

- Large β_t: ν_{t+1} not a descent direction anymore: $MMD^2(\nu^*, \nu_{t+1}) > MMD^2(\nu^*, \nu_t)$
- Small β_t: Back to the failure mode: $\nabla f_{\nu_t,\nu^*}(X_t + \beta_t u_t) \simeq 0$.

Noise Injection: Theory (discrete time)

Tradeoff for β_t

- Large β_t: ν_{t+1} not a descent direction anymore: $MMD^2(\nu^*, \nu_{t+1}) > MMD^2(\nu^*, \nu_t)$
- Small β_t: Back to the failure mode: $\nabla f_{\nu_t,\nu^*}(X_t + \beta_t u_t) \simeq 0$.

Need β_t such that:

$$\beta_t^2 MMD^2(\nu_t) \leq C_k \mathbb{E}_{\substack{X_t \sim \nu_t \\ U_t \sim \mathcal{N}(0,1)}} [\|\nabla f_t(X_t + \beta_t U_t)\|^2] \qquad (1)$$

and:

$$\sum_{t=1}^{T} \beta_t^2 \to \infty$$

Noise Injection: Theory (discrete time)

Tradeoff for β_t

- Large β_t: ν_{t+1} not a descent direction anymore: $MMD^2(\nu^*, \nu_{t+1}) > MMD^2(\nu^*, \nu_t)$
- Small β_t: Back to the failure mode: $\nabla f_{\nu_t,\nu^*}(X_t + \beta_t u_t) \simeq 0$.

Need β_t such that:

$$\beta_t^2 MMD^2(\nu_t) \leq C_k \mathbb{E}_{\substack{X_t \sim \nu_t \\ U_t \sim \mathcal{N}(0,1)}} [\|\nabla f_t(X_t + \beta_t U_t)\|^2] \qquad (1)$$

and:

$$\sum_{t=1}^{T} \beta_t^2 \to \infty$$

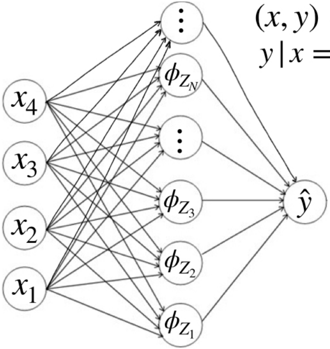
Then

$$MMD^2(\nu^*, \nu_T) \leq MMD^2(\nu^*, \nu_0) e^{-C_k \gamma (1 - \gamma C_k') \sum_{t=1}^{T} \beta_t^2}$$

Noise Injection: Student-Teacher network

Recall the supervised learning problem in the well specified case:

$$(x, y) \sim data$$
$$y \,|\, x = \mathbb{E}_{U \sim \nu^*}[\phi_U(x)]$$

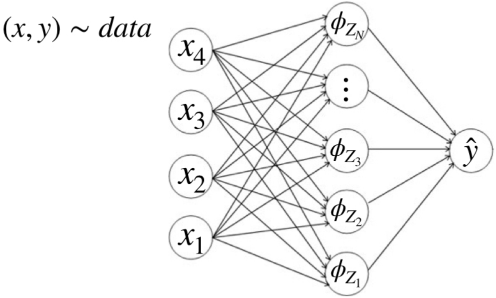


$$\min_{\nu \in \mathscr{P}} \mathbb{E}_{data}[\|y - \mathbb{E}_{Z \sim \nu}[\phi_Z(x)]\|^2]$$

Noise Injection: Student-Teacher network

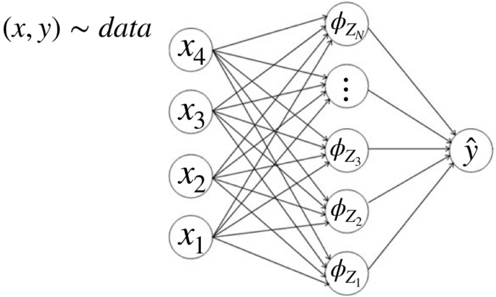
Example of the Student-Teacher network:

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



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

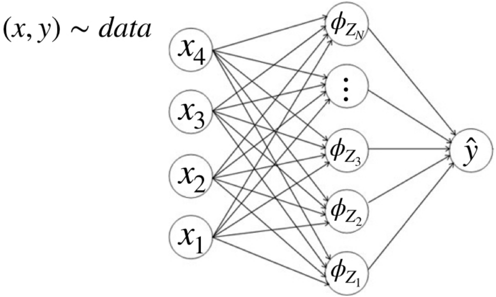
Noise Injection: Student-Teacher network



$$\min_{Z_1,\ldots,Z_N} MMD^2(\nu^*, \frac{1}{N}\sum_{n=1}^{N} \delta_{Z^n})$$

$$k(Z, Z') = \mathbb{E}_{data}[\phi_Z(x)\phi_{Z'}(x)]$$

Noise Injection: Student-Teacher network



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

$$\hat{k}(Z, Z') = \frac{1}{B}\sum_{b=1}^{B}\phi_Z(x_b)\phi_{Z'}(x_b)$$

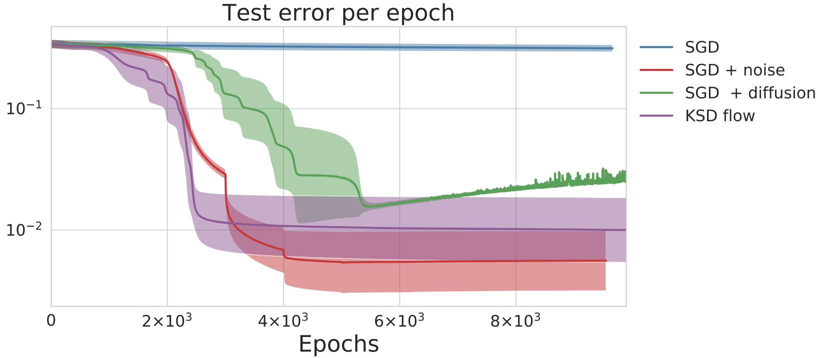
Noise Injection: Experiments

Methods:

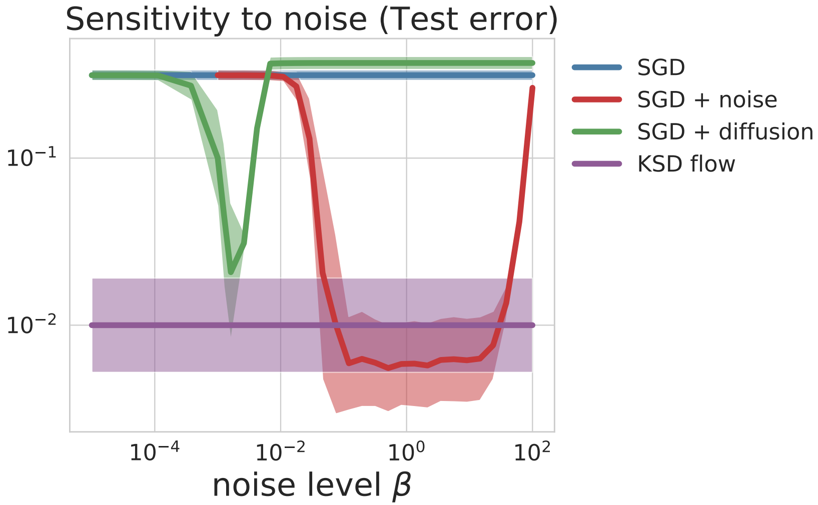
- SGD
- SGD + Noise injection
- SGD + diffusion
- KSD [7]: SGD using the Negative Sobolev distance $\nu \mapsto \|\nu - \nu^*\|_{\dot{H}^{-1}(\nu)}$ as a loss function: also decreases the MMD.

[7] [Mroueh et al., 2019]

Noise Injection: Experiments



Noise Injection: Experiments



Conclusion

Contributions:

- Provided a convergence criterion for the Wasserstein gradient flow of the MMD.
- Studying the gradient flow and its time/space discretizations helps understand the convergence when training big neural networks
- Proposed a pertubation of the dynamics with a noise injection and showed it effectiveness on simple examples.

Openings:

- A criterion for convergence that is independent from the whole optimization trajectory.
- Stronger guarantees for the convergence for the noise injection algorithm.

Thank you!

Ambrosio, L., Gigli, N., and Savaré, G. (2004).
Gradient flows with metric and differentiable structures, and applications to the Wasserstein space.
Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti Lincei. Matematica e Applicazioni, 15(3-4):327–343.

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.

Chaudhari, P., Oberman, A., Osher, S., Soatto, S., and Carlier, G. (2017).
Deep Relaxation: partial differential equations for optimizing deep neural networks.
arXiv:1704.04932 [cs, math].

Chizat, L. and Bach, F. (2018).
On the global convergence of gradient descent for

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 $\nu_n \Longrightarrow$ 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**

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

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.