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

June 10, 2020

Overview

- **Problem considered**: Transporting mass from an initial distribution ν_0 to a target distribution ν^*, by finding a continuous path ν_t decreasing a loss $\mathcal{F}(\nu_t)$.

  $\Longrightarrow$ **Wasserstein Gradient flows over the space of distributions**

Overview

- **Problem considered**: Transporting mass from an initial distribution ν_0 to a target distribution ν^*, by finding a continuous path ν_t decreasing a loss $\mathcal{F}(\nu_t)$.

  $\Longrightarrow$ **Wasserstein Gradient flows over the space of distributions**

- **Applications**:
  - Convergence properties of neural networks with infinite width.
  - "Sampling": Data summarization

Overview

- **Problem considered**: Transporting mass from an initial distribution ν_0 to a target distribution ν^*, by finding a continuous path ν_t decreasing a loss $\mathcal{F}(\nu_t)$.

  $\Longrightarrow$ **Wasserstein Gradient flows over the space of distributions**

- **Applications**:
  - Convergence properties of neural networks with infinite width.
  - "Sampling": Data summarization

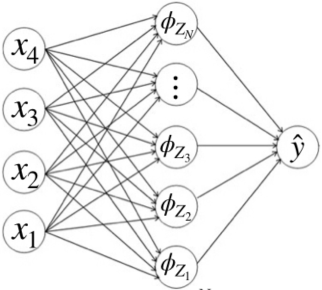
- **This work** :
  - Particular functional $\mathcal{F}(\nu) = MMD^2(\nu, \nu^*)$.
  - Investigate the global convergence of the Wasserstein gradient flow of the MMD.

Outline

- Motivation
- Wasserstein gradient flow of the MMD
- A Criterion for global convergence
- A noise-injection algorithm for better empirical convergence

Motivation: Optimization of neural networks

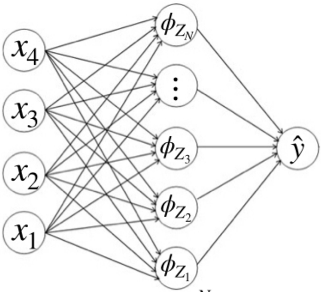
$(x, y) \sim data$



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

Motivation: Optimization of neural networks

$(x, y) \sim data$



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

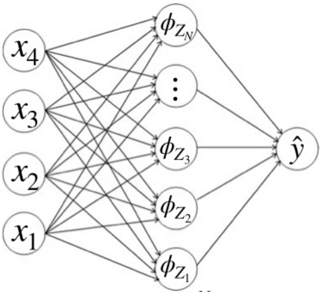
$$\min_{Z_1,\ldots,Z_N \in \mathcal{Z}} \mathcal{L}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i}\right)$$

- Optimization using gradient descent GD:

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

Motivation: Optimization of neural networks

$(x, y) \sim data$



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

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

- Optimization using gradient descent GD:

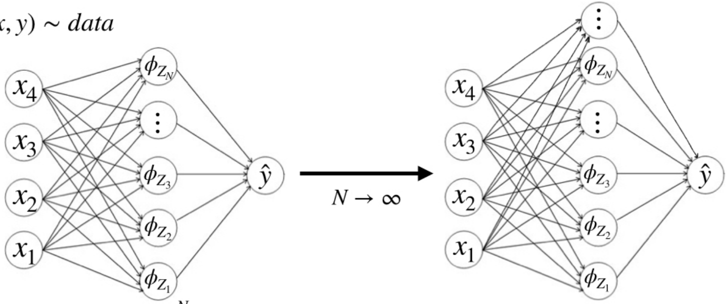
$$Z_i^{t+1} = Z_i^t - \gamma \nabla_{Z_i} \mathcal{L}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i^t}\right)$$

- Hard to describe the dynamics of GD!

Motivation: Optimization of infinite width neural networks

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

$(x, y) \sim data$



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

Motivation: Optimization of infinite width neural networks

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

- Global Convergence of GD when $N \to \infty$ [1] and:

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

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

Motivation: Optimization of infinite width neural networks

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

- Global Convergence of GD when $N \to \infty$ [1] and:

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

- In this work, interested in more general forms for $\phi_Z(x)$.

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

Motivation: Optimization of infinite width neural networks

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

- Global Convergence of GD when $N \to \infty$ [1] and:

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

- In this work, interested in more general forms for $\phi_Z(x)$.
- **Connexion to the MMD** :
  - Well-defined setting: $y = \mathbb{E}_{U\sim\nu^*}[\phi_U(x)]$
  - Random feature formulation:

$$\mathcal{L}(\nu) = \mathbb{E}_x\left[\|\mathbb{E}_{U\sim\nu^*}[\phi_U(x)] - \mathbb{E}_{Z\sim\nu}[\phi_Z(x)]\|^2\right] = MMD^2(\nu, \nu^\star)$$

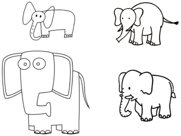
  - MMD with kernel $k(U, Z) = \mathbb{E}_x[\phi_U(x)^\top \phi_Z(x)]$

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

The Maximum Mean Discrepancy [Gretton et al., 2012]

Consider samples from two distributions ν^* and ν_0.

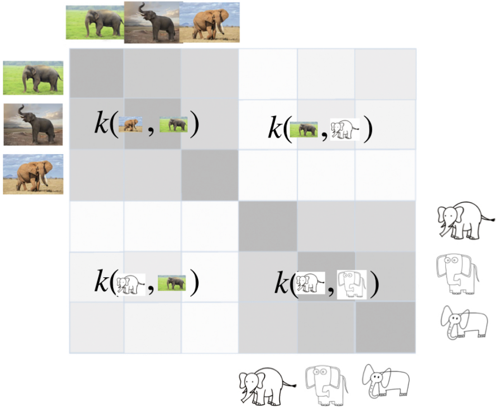
$U^m \sim \nu^*$



$Z^n \sim \nu_0$

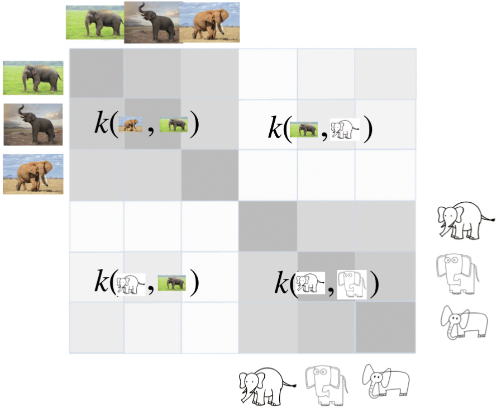
The Maximum Mean Discrepancy [Gretton et al., 2012]

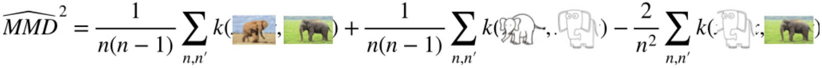
Compute a similarity matrix using a kernel k



The Maximum Mean Discrepancy [Gretton et al., 2012]

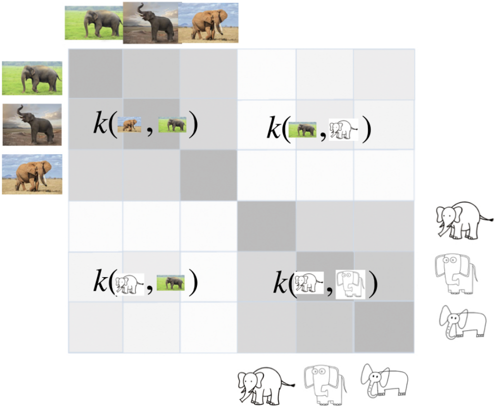
Compute a similarity matrix using a kernel *k*





The Maximum Mean Discrepancy [Gretton et al., 2012]

Compute a similarity matrix using a kernel k



$$MMD^2(\nu^*, \nu_0) = \mathbb{E}_{\substack{U \sim \nu^* \\ U' \sim \nu^*}}[k(U, U')] + \mathbb{E}_{\substack{Z \sim \nu_0 \\ Z' \sim \nu_0}}[k(Z, Z')] - 2\mathbb{E}_{\substack{U \sim \nu^* \\ Z' \sim \nu_0}}[k(U, Z)]$$

Gradient flows - Euclidean setting

- $(Z_t)_{t\geq 0}$ is a gradient flow of a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ if it satisfies:

$$\frac{dZ_t}{dt} = -\nabla F(Z_t), \qquad Z_0 = z_0$$

Gradient flows - Euclidean setting

- $(Z_t)_{t\geq 0}$ is a gradient flow of a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ if it satisfies:

$$\frac{dZ_t}{dt} = -\nabla F(Z_t), \qquad Z_0 = z_0$$

- Given z_0, Z_t is unique and well defined under mild conditions on F (Cauchy-Lipschitz thm).

Gradient flows - Euclidean setting

- $(Z_t)_{t\geq 0}$ is a gradient flow of a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ if it satisfies:

$$\frac{dZ_t}{dt} = -\nabla F(Z_t), \qquad Z_0 = z_0$$

- Given z_0, Z_t is unique and well defined under mild conditions on F (Cauchy-Lipschitz thm).
- The gradient $\nabla F(z)$ is defined w.r.t Euclidean metric:

$$\nabla F(z)^\top u := g_z(\nabla F(z), u) = dF_z(u).$$

Gradient flows - Euclidean setting

- $(Z_t)_{t\geq 0}$ is a gradient flow of a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ if it satisfies:

$$\frac{dZ_t}{dt} = -\nabla F(Z_t), \qquad Z_0 = z_0$$

- Given z_0, Z_t is unique and well defined under mild conditions on F (Cauchy-Lipschitz thm).
- The gradient $\nabla F(z)$ is defined w.r.t Euclidean metric:

$$\nabla F(z)^\top u := g_z(\nabla F(z), u) = dF_z(u).$$

- Euclidean distance as a geodesic distance:

$$\|Z - Z'\|^2 = \inf_{(v_t, z_t)_{0\leq t\leq 1}} \int_0^1 g_{z_t}(v_t, v_t)dt$$

Gradient flows on the space of distributions

- For a functional $\mathcal{F}$ on probability space, a gradient flow formally looks like

$$\frac{d\nu_t}{dt} = -\nabla \mathcal{F}(\nu_t), \qquad \nu_0.$$

- Need a suitable metric to give a meaning for $\nabla \mathcal{F}(\nu_t)$.

Wasserstein-2 metric [Benamou and Brenier, 2000, Otto, 2001]

- Wasserstein-2 distance:

$$W_2^2(\nu,\mu) = \inf_{\pi\Pi(\nu,\mu)} \mathbb{E}_{(Z,Z')\sim\pi}[\|Z - Z'\|^2].$$

- The Wasserstein distance as a geodesic distance[2]

$$W_2^2(\nu,\mu) := \inf_{(\rho_t,f_t)} \int_0^1 \int \|\nabla f_t(x)\|^2 \, \mathrm{d}\rho_t(x) dt,$$
$$\partial_t \rho_t + div(\rho_t \nabla f_t) = 0$$

[2] [Benamou and Brenier, 2000]

Wasserstein-2 metric [Benamou and Brenier, 2000, Otto, 2001]

- Wasserstein-2 distance:

$$W_2^2(\nu,\mu) = \inf_{\pi \Pi(\nu,\mu)} \mathbb{E}_{(Z,Z')\sim\pi}[\|Z - Z'\|^2].$$

- The Wasserstein distance as a geodesic distance[2]

$$W_2^2(\nu,\mu) := \inf_{(\rho_t,f_t)} \int_0^1 \int \|\nabla f_t(x)\|^2 \,\mathrm{d}\rho_t(x) dt,$$
$$\partial_t \rho_t + div(\rho_t \nabla f_t) = 0$$

- Wasserstein metric:

$$g_\nu(\delta,\delta) := \int \|\nabla f(x)\|^2 \,\mathrm{d}\nu(x), \quad \delta + div(\nu \nabla f) = 0.$$

---

[2] [Benamou and Brenier, 2000]

Wasserstein-2 gradient [Otto, 2001, Ambrosio et al., 2004]

- Wasserstein metric:

$$g_\nu(\delta, \delta) := \int \|\nabla f(x)\|^2 \, \mathrm{d}\nu(x), \quad \delta + div(\nu \nabla f) = 0.$$

Wasserstein-2 gradient [Otto, 2001, Ambrosio et al., 2004]

- Wasserstein metric:

$$g_\nu(\delta, \delta) := \int \|\nabla f(x)\|^2 \, \mathrm{d}\nu(x), \quad \delta + div(\nu \nabla f) = 0.$$

- First variation of a functional along direction δ:

$$d\mathcal{L}_\nu(\delta) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\mathcal{L}(\nu + \epsilon\delta) - \mathcal{L}(\nu) \right) := \int \frac{\partial \mathcal{L}}{\partial \nu}(\nu)(z) d\delta(z).$$

Wasserstein-2 gradient [Otto, 2001, Ambrosio et al., 2004]

- Wasserstein metric:

$$g_\nu(\delta,\delta) := \int \|\nabla f(x)\|^2 \mathrm{d}\nu(x), \quad \delta + div(\nu \nabla f) = 0.$$

- First variation of a functional along direction δ:

$$d\mathcal{L}_\nu(\delta) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\mathcal{L}(\nu + \epsilon\delta) - \mathcal{L}(\nu) \right) := \int \frac{\partial \mathcal{L}}{\partial \nu}(\nu)(z) d\delta(z).$$

- Under mild condition on ν and δ there exists a vector field ∇f_δ satisfying:

$$\delta + div(\nu \nabla f_\delta) = 0$$

Wasserstein-2 gradient [Otto, 2001, Ambrosio et al., 2004]

- Wasserstein metric:

$$g_\nu(\delta, \delta) := \int \|\nabla f(x)\|^2 \, \mathrm{d}\nu(x), \quad \delta + div(\nu \nabla f) = 0.$$

- First variation of a functional along direction δ:

$$d\mathcal{L}_\nu(\delta) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\mathcal{L}(\nu + \epsilon\delta) - \mathcal{L}(\nu) \right) := \int \frac{\partial \mathcal{L}}{\partial \nu}(\nu)(z) d\delta(z).$$

- Under mild condition on ν and δ there exists a vector field ∇f_δ satisfying:

$$\delta + div(\nu \nabla f_\delta) = 0$$

- Wasserstein-2 gradient of $\mathcal{F}$ obtained by integration by part:

$$d\mathcal{L}_\nu(\delta) = \int \nabla \frac{\partial \mathcal{L}}{\partial \nu}(\nu)^\top \nabla f_\delta d\nu = g_\nu(\nabla^{W_2} \mathcal{L}, \delta)$$

$$\nabla^{W_2} \mathcal{L}(\nu) := -div(\nu \nabla \frac{\partial \mathcal{L}}{\partial \nu}(\nu))$$

Wasserstein-2 gradient flow of the MMD

- First variation of the MMD:

$$\frac{\partial MMD^2}{\partial \nu}(\nu)(z) := f_{\nu^*,\nu}(z) = 2\left(\mathbb{E}_{U\sim\nu^*}[k(U,z)] - \mathbb{E}_{U\sim\nu}[k(U,z)]\right)$$

Wasserstein-2 gradient flow of the MMD

- First variation of the MMD:

$$\frac{\partial MMD^2}{\partial \nu}(\nu)(z) := f_{\nu^*,\nu}(z) = 2\left(\mathbb{E}_{U\sim\nu^*}[k(U,z)] - \mathbb{E}_{U\sim\nu}[k(U,z)]\right)$$

- Gradient flow of the MMD:

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

Wasserstein-2 gradient flow of the MMD

- First variation of the MMD:

$$\frac{\partial MMD^2}{\partial \nu}(\nu)(z) := f_{\nu^*,\nu}(z) = 2\left(\mathbb{E}_{U\sim\nu^*}[k(U,z)] - \mathbb{E}_{U\sim\nu}[k(U,z)]\right)$$

- Gradient flow of the MMD:

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

- Equivalent to a Stochastic Differential Equation: Mc-Kean Vlasov dynamics

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

Wasserstein-2 gradient flow of the MMD

- First variation of the MMD:

$$\frac{\partial MMD^2}{\partial \nu}(\nu)(z) := f_{\nu^*,\nu}(z) = 2\left(\mathbb{E}_{U\sim\nu^*}[k(U,z)] - \mathbb{E}_{U\sim\nu}[k(U,z)]\right)$$

- Gradient flow of the MMD:

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

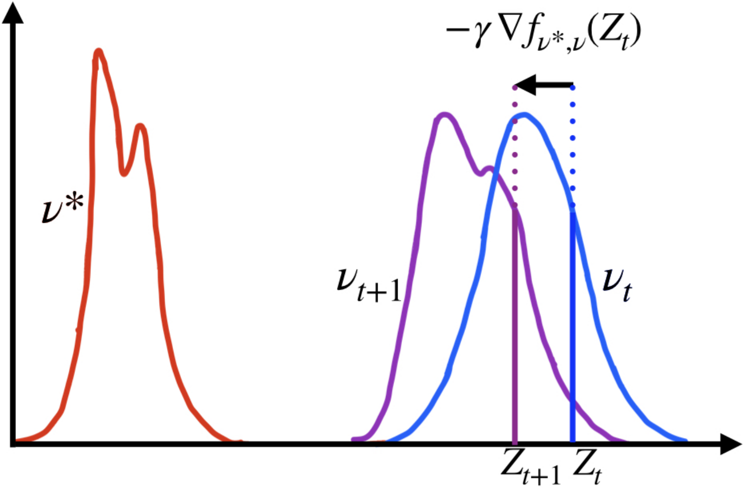
- Equivalent to a Stochastic Differential Equation: Mc-Kean Vlasov dynamics

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

- Discrete-time version:

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

Wasserstein-2 gradient flow of the MMD



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

Global convergence: First strategy

Displacement convexity:

- A geodesic ρ_t between ρ_0 and ρ_1 is given by optimal coupling $\pi^\star$:

$$X_t \sim \rho_t \iff X_t = (1_t)X_0 + tX_1 \qquad (X_0, X_1) \sim \pi^\star$$

- A functional $\mathcal{F}$ is displacement convex if:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\rho_0) + t\mathcal{F}(\rho_1)$$

- Unfortunately the MMD is not displacement convex in general.

Global convergence: Second Strategy

Dissipation inequalities:

- Rate by which $\mathcal{F}$ decreases along the gradient flow:

$$\frac{d\mathcal{F}(\nu_t)}{dt} = -\mathbb{E}_{\nu_t}[\|\nabla f_{\nu^\star,\nu_t}\|^2]$$

Global convergence: Second Strategy

Dissipation inequalities:

- Rate by which $\mathcal{F}$ decreases along the gradient flow:

$$\frac{d\mathcal{F}(\nu_t)}{dt} = -\mathbb{E}_{\nu_t}[\|\nabla f_{\nu^\star,\nu_t}\|^2]$$

- Assumption: Controlling the dissipation rate: (general Lojasiewicz inequality)

$$\mathcal{F}(\nu) \leq C\mathbb{E}_{\nu}[\|\nabla f_{\nu^\star,\nu}\|^2]$$

Global convergence: Second Strategy

Dissipation inequalities:

- Rate by which $\mathcal{F}$ decreases along the gradient flow:

$$\frac{d\mathcal{F}(\nu_t)}{dt} = -\mathbb{E}_{\nu_t}[\|\nabla f_{\nu^\star,\nu_t}\|^2]$$

- Assumption: Controlling the dissipation rate: (general Lojasiewicz inequality)

$$\mathcal{F}(\nu) \leq C\mathbb{E}_{\nu}[\|\nabla f_{\nu^\star,\nu}\|^2]$$

- Combining both equations and using Gronwall lemma:

$$\mathcal{F}(\nu_t) \leq \frac{1}{\mathcal{F}(\nu_0)^{-1} + 2C^{-1}t}$$

Global convergence: Second Strategy

Dissipation inequalities:

- Rate by which $\mathcal{F}$ decreases along the gradient flow:

$$\frac{d\mathcal{F}(\nu_t)}{dt} = -\mathbb{E}_{\nu_t}[\|\nabla f_{\nu^\star,\nu_t}\|^2]$$

- Assumption: Controlling the dissipation rate: (general Lojasiewicz inequality)

$$\mathcal{F}(\nu) \leq C\mathbb{E}_{\nu}[\|\nabla f_{\nu^\star,\nu}\|^2]$$

- Combining both equations and using Gronwall lemma:

$$\mathcal{F}(\nu_t) \leq \frac{1}{\mathcal{F}(\nu_0)^{-1} + 2C^{-1}t}$$

- Does the Lojasiewicz inequality hold for the MMD?

Lojasiewicz-type inequality for the MMD

- Find $C > 0$ such that:

$$\mathcal{F}(\nu) \leq C\mathbb{E}_{\nu}[\|\nabla f_{\nu^{\star},\nu}\|^2]$$

Lojasiewicz-type inequality for the MMD

- Find $C > 0$ such that:

$$\mathcal{F}(\nu) \leq C\mathbb{E}_{\nu}[\|\nabla f_{\nu^{\star},\nu}\|^2]$$

- By Cauchy-Schwartz inequality in the RKHS space:

$$\mathrm{MMD}^2(\nu_t, \nu^{\star}) \leq S(\nu^{*}|\nu_t)\mathbb{E}_{\nu}[\|\nabla f_{\nu^{\star},\nu}\|^2]$$

Lojasiewicz-type inequality for the MMD

- Find $C > 0$ such that:

$$\mathcal{F}(\nu) \leq C\mathbb{E}_{\nu}[\|\nabla f_{\nu^\star,\nu}\|^2]$$

- By Cauchy-Schwartz inequality in the RKHS space:

$$\mathrm{MMD}^2(\nu_t, \nu^\star) \leq S(\nu^*|\nu_t)\mathbb{E}_{\nu}[\|\nabla f_{\nu^\star,\nu}\|^2]$$

- $S(\nu^*|\nu_t)$ is the Negative Sobolev divergence:

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

Lojasiewicz-type inequality for the MMD

- Find $C > 0$ such that:

$$\mathcal{F}(\nu) \leq C\mathbb{E}_{\nu}[\|\nabla f_{\nu^\star,\nu}\|^2]$$

- By Cauchy-Schwartz inequality in the RKHS space:

$$\mathrm{MMD}^2(\nu_t, \nu^\star) \leq S(\nu^*|\nu_t)\mathbb{E}_{\nu}[\|\nabla f_{\nu^\star,\nu}\|^2]$$

- $S(\nu^*|\nu_t)$ is the Negative Sobolev divergence:

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

- Lojasiewicz inequality holds when $S(\nu^*|\nu_t)$ remains bounded by $C > 0$

Lojasiewicz-type inequality for the MMD

- Find $C > 0$ such that:

$$\mathcal{F}(\nu) \leq C\mathbb{E}_{\nu}[\|\nabla f_{\nu^\star,\nu}\|^2]$$

- By Cauchy-Schwartz inequality in the RKHS space:

$$\mathrm{MMD}^2(\nu_t, \nu^\star) \leq S(\nu^*|\nu_t)\mathbb{E}_{\nu}[\|\nabla f_{\nu^\star,\nu}\|^2]$$

- $S(\nu^*|\nu_t)$ is the Negative Sobolev divergence:

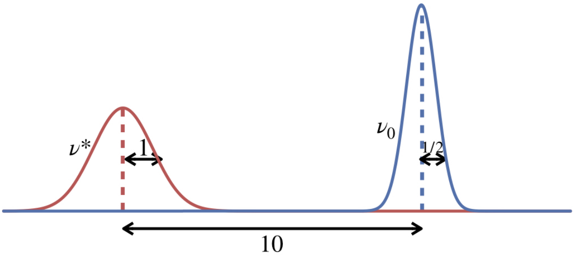
$$S(\nu^*|\nu_t) = \sup_{g, \mathbb{E}_{Z\sim\nu_t}[\|\nabla g(Z)\|^2] \leq 1} |\mathbb{E}_{Z\sim\nu_t}[g(Z)] - \mathbb{E}_{U\sim\nu^*}[g(U)]|$$

- Lojasiewicz inequality holds when $S(\nu^*|\nu_t)$ remains bounded by $C > 0$
- Depends on the whole sequence ν_t: Hard to verify in general

Convergence: Failure case

See animation at

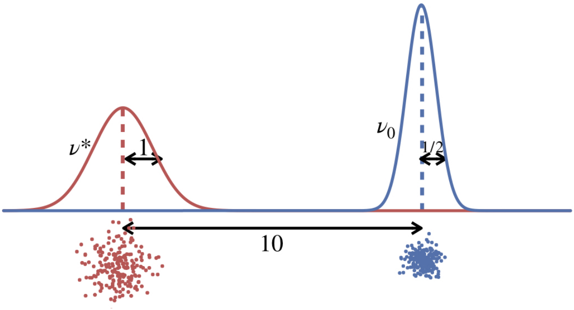
`https://michaelarbel.github.io/MMD_flow.html`



Convergence: Failure case

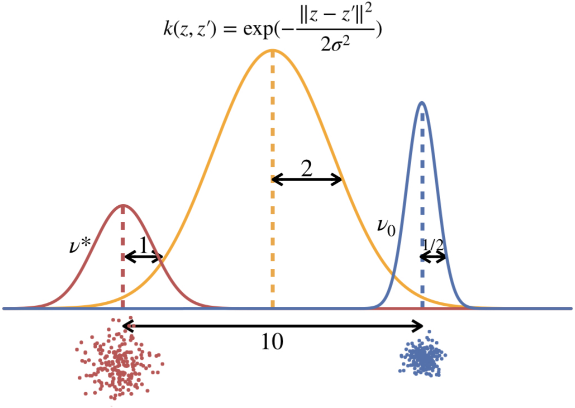
See animation at

`https://michaelarbel.github.io/MMD_flow.html`



Convergence: Failure case

See animation at
https://michaelarbel.github.io/MMD_flow.html



Convergence: Failure case

Some observations:

- Almost all (blue) particles tend to collapse on 1 point at the center of mass m of the target $\nu^\star$, i.e.: $(\nu_t \simeq \delta_m)$
- Some (blue) particles seem to escape towards infinity.
- However, the loss stops decreasing: $\nabla f_{\nu^\star,\nu_t}(z) \simeq 0$ for z on the support of ν_t (which is tiny $\nu_t \approx \delta_m$!!)
- However, in general, $\nabla f_{\nu^\star,\nu_t}(z) \neq 0$ outside the support of ν_t. Can this fact be used somehow to improve convergence ?

Improving empirical convergence: Noise Injection

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

[3] [Chaudhari et al., 2017, Hazan et al., 2016]
[4] [Mei et al., 2018]

Improving empirical convergence: Noise Injection

- Idea: Evaluate $\nabla f_{\nu^*,\nu_t}$ outside of the support of ν_t to get a better signal!
- Sample $u_t \sim \mathcal{N}(0,1)$ and β_t is the noise level:

$$Z_{t+1} = Z_t - \gamma \nabla f_t(Z_t + \beta_t u_t); \qquad Z_t \sim \nu_t$$

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

[4] [Mei et al., 2018]

Improving empirical convergence: Noise Injection

- Idea: Evaluate $\nabla f_{\nu^*,\nu_t}$ outside of the support of ν_t to get a better signal!
- Sample $u_t \sim \mathcal{N}(0, 1)$ and β_t is the noise level:

$$Z_{t+1} = Z_t - \gamma \nabla f_t(Z_t + \beta_t u_t); \qquad Z_t \sim \nu_t$$

- Similar to *continuation methods* [3], but extended to interacting particles.

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

[4] [Mei et al., 2018]

Improving empirical convergence: Noise Injection

- Idea: Evaluate $\nabla f_{\nu^*,\nu_t}$ outside of the support of ν_t to get a better signal!
- Sample $u_t \sim \mathcal{N}(0,1)$ and β_t is the noise level:

$$Z_{t+1} = Z_t - \gamma \nabla f_t(Z_t + \beta_t u_t); \qquad Z_t \sim \nu_t$$

- Similar to *continuation methods* [3], but extended to interacting particles.
- Different from entropic regularization[4]

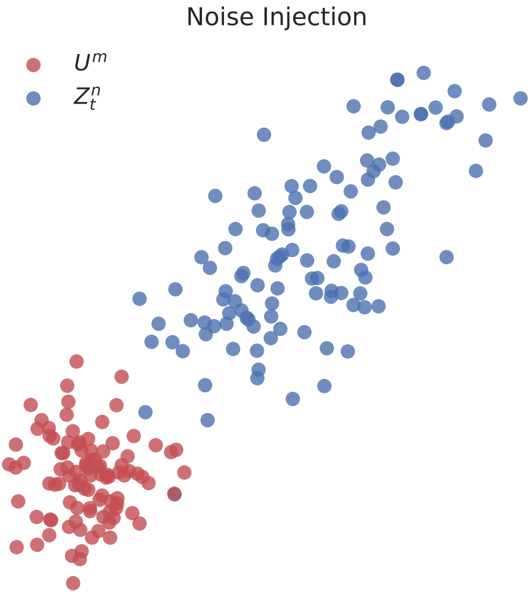
$$Z_{t+1} = Z_t - \gamma \nabla f_{\nu^*,\nu_t}(Z_t) + \beta_t u_t$$

---

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

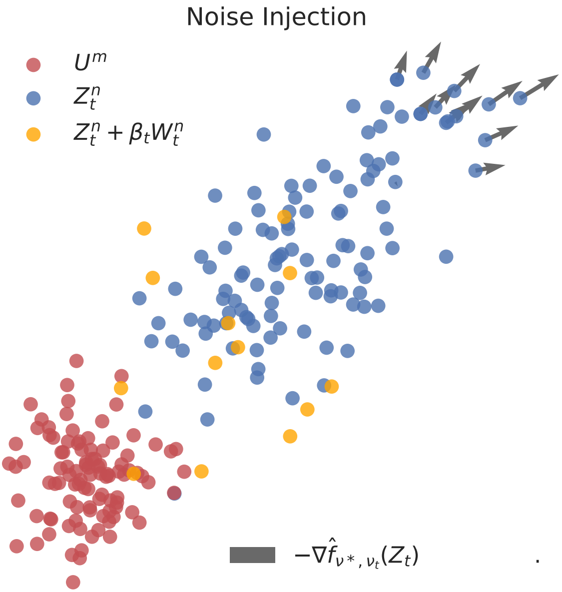
[4] [Mei et al., 2018]

Noise Injection[5]



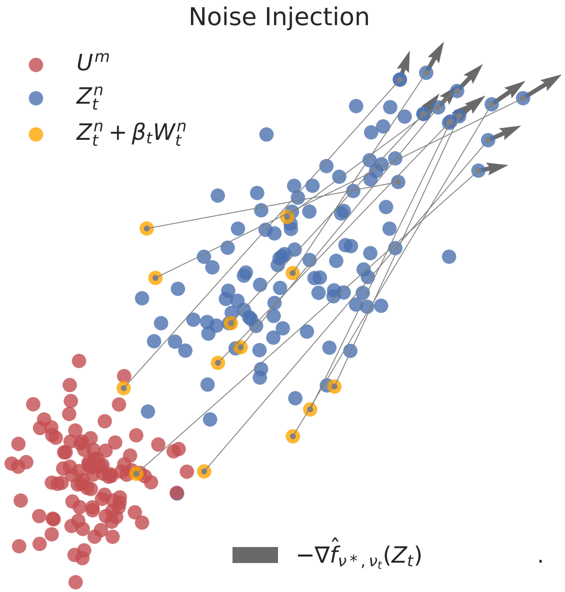
[5]See `https://michaelarbel.github.io/MMD_flow.html`

Noise Injection[5]



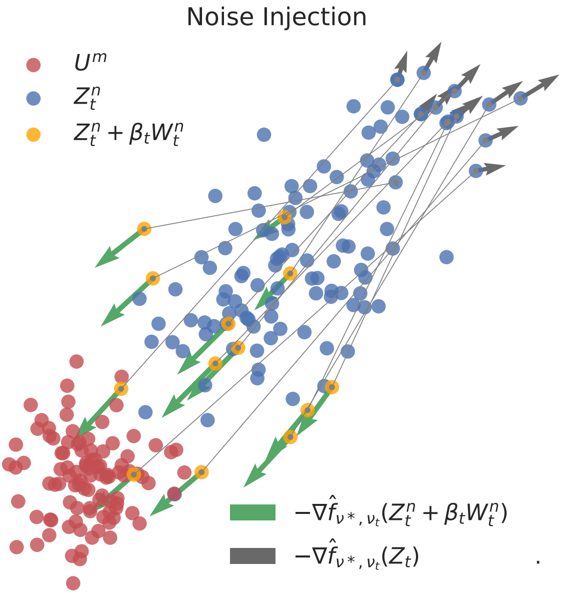
[5]See https://michaelarbel.github.io/MMD_flow.html

Noise Injection[5]



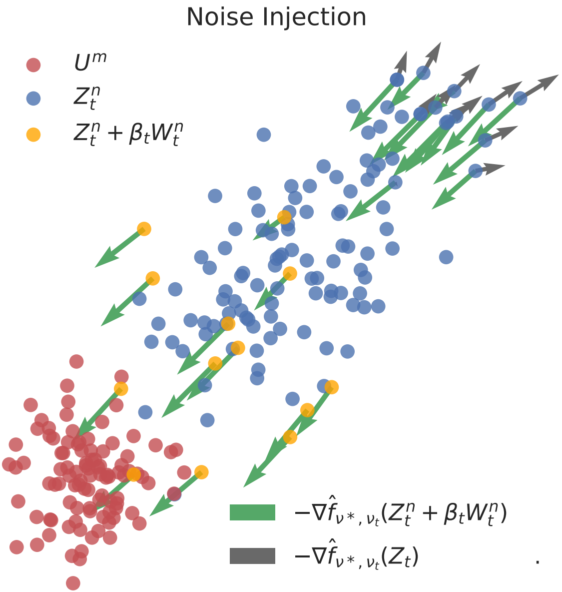
[5]See https://michaelarbel.github.io/MMD_flow.html

Noise Injection[5]



[5]See `https://michaelarbel.github.io/MMD_flow.html`

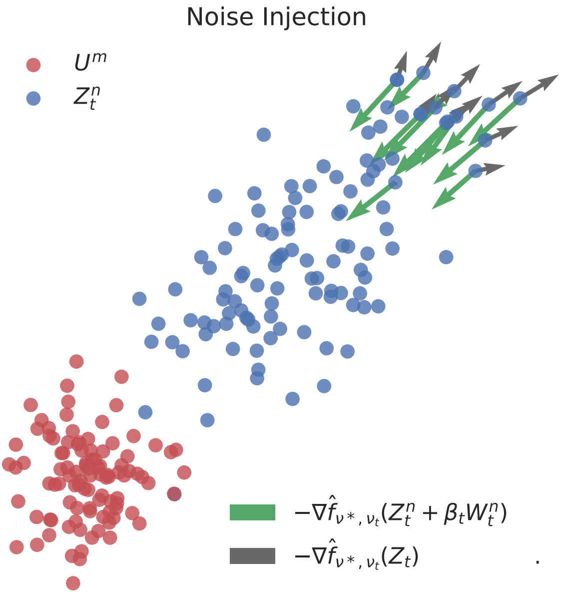
Noise Injection[5]



.

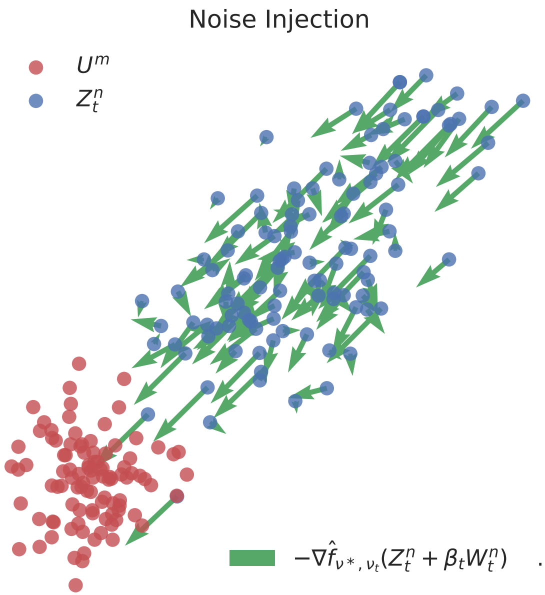
[5]See `https://michaelarbel.github.io/MMD_flow.html`

Noise Injection[5]



[5] See `https://michaelarbel.github.io/MMD_flow.html`

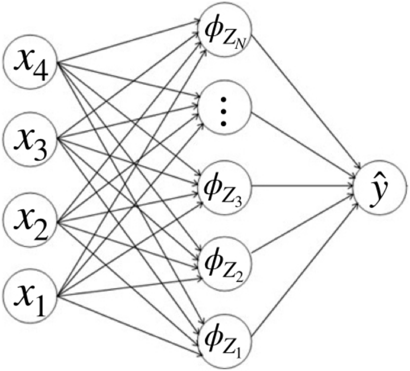
Noise Injection[5]



[5] See https://michaelarbel.github.io/MMD_flow.html

Noise Injection: Student-Teacher setting

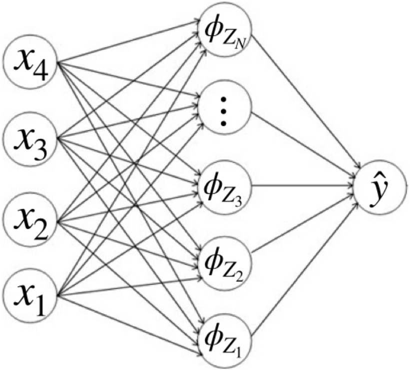
$(x, y) \sim data$



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

Noise Injection: Student-Teacher setting

$(x, y) \sim data$



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

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

Noise Injection: Experiments

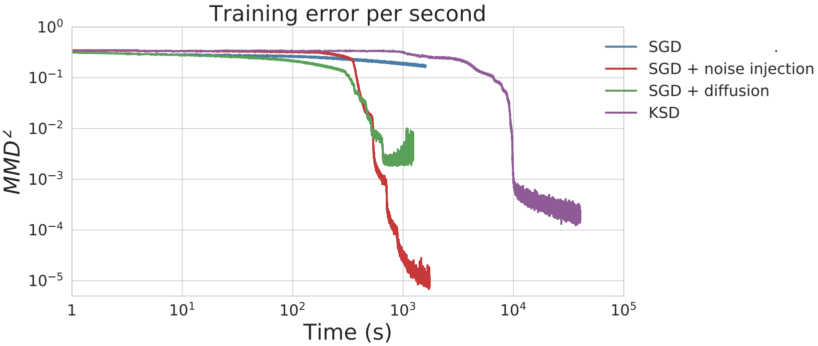
Methods:

- SGD (Approximates the MMD flow)
- SGD + Noise injection
- SGD + diffusion
- KSD [6]: SGD using the Negative Sobolev distance $\nu \mapsto S(\nu^*|\nu)$ as a loss function: also minimizes the MMD.

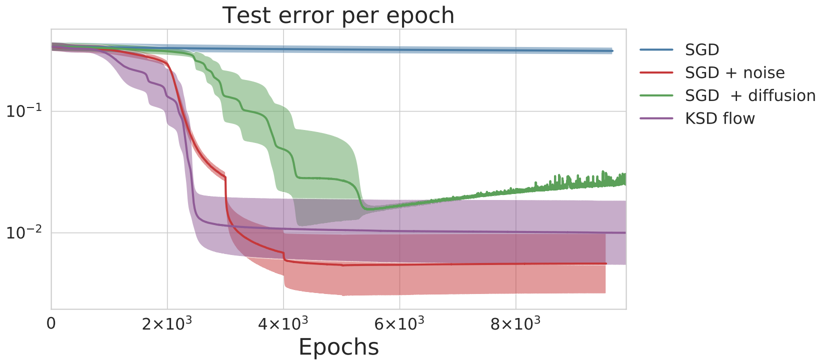
---

[6] [Mroueh et al., 2019]

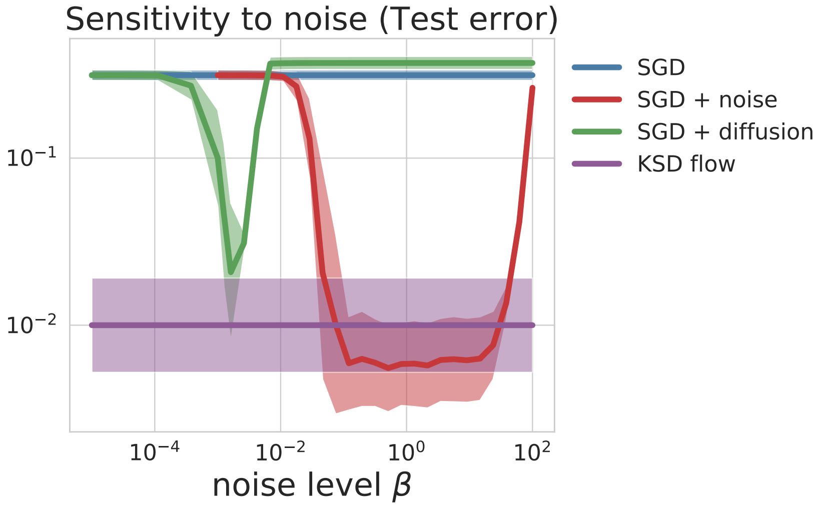
Noise Injection: Experiments



Noise Injection: Experiments



Noise Injection: Experiments



Conclusion

Contributions:

- Provided a convergence criterion for the Wasserstein gradient descent.
- Proposed an extension to the noise injection algorithm for interacting particles and showed it effectiveness on simple examples.

Future work:

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

Benamou, J.-D. and Brenier, Y. (2000).
A computational fluid mechanics solution to the monge-kantorovich mass transfer problem.
Numerische Mathematik, 84(3):375–393.

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 over-parameterized models using optimal transport.

Noise Injection: Theory

Tradeoff for β_t

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

Noise Injection: Theory

Tradeoff for β_t

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

Noise Injection: Theory

Tradeoff for β_t

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

Need β_t such that:

$$MMD^2(\nu^*, \mu_{t+1}) - MMD^2(\nu^*, \mu_t) \leq C\gamma \mathbb{E}_{\substack{X_t \sim \mu_t \\ U_t \sim \mathcal{N}(0,1)}} [\|\nabla f_t(X_t + \beta_t U_t)\|^2]$$

and:

$$\sum_i^t \beta_i^2 \to \infty$$

Noise Injection: Theory

Tradeoff for β_t

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

Need β_t such that:

$$MMD^2(\nu^*, \mu_{t+1}) - MMD^2(\nu^*, \mu_t) \leq C\gamma \mathbb{E}_{\substack{X_t \sim \mu_t \\ U_t \sim \mathcal{N}(0,1)}} [\|\nabla f_t(X_t + \beta_t U_t)\|^2]$$

and:

$$\sum_i^t \beta_i^2 \to \infty$$

Then

$$MMD^2(\nu^*, \nu_t) \leq MMD^2(\nu^*, \nu_0) e^{-C\gamma \sum_i^t \beta_i^2}$$