Maximum Mean Discrepancy Gradient Flow

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

[1]Gatsby Computational Neuroscience Unit, University College London [2]KAUST

Overview

General setting:

✓ Non-convex optimization in probability space with the Maximum Mean Discrepancy as a cost function.

✓ Interested in Gradient descent dynamics in the limit of large samples $N \to \infty$.

Goals:

✓ Criterion for global convergence of gradient descent when N approaches infinity.

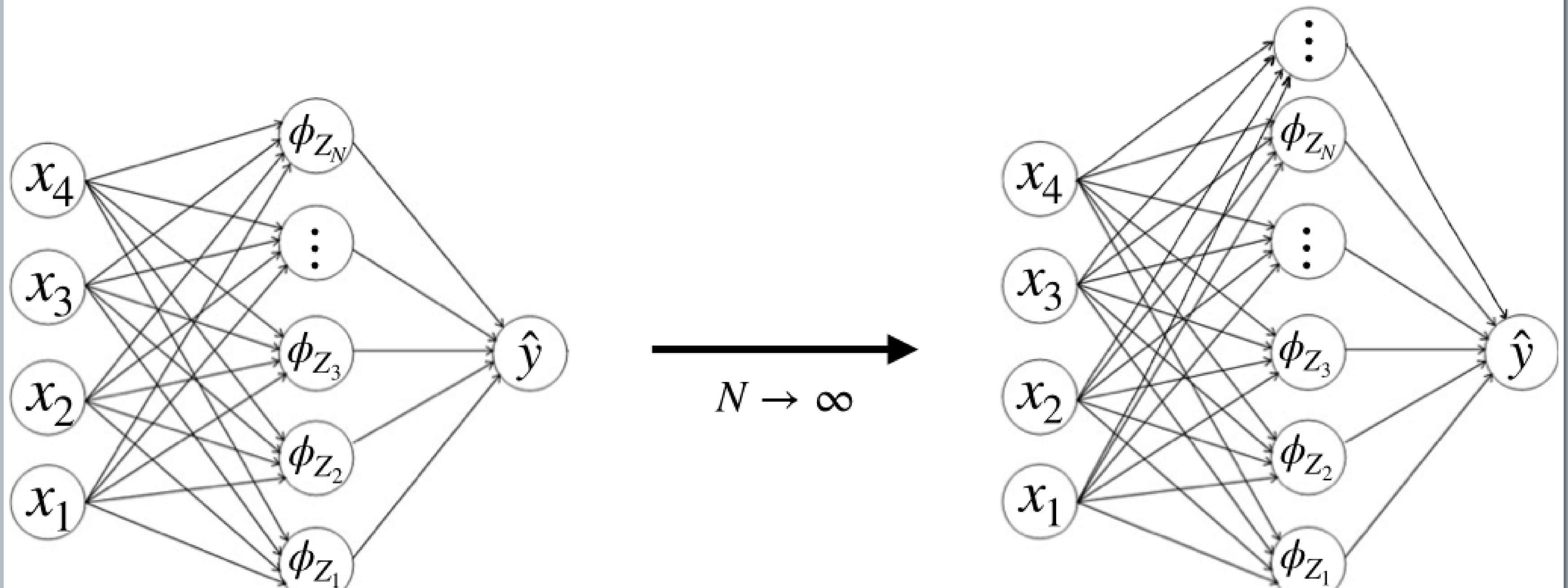
✓ New algorithm based on noise-injection to improve convergence.

✓ Application 1: Optimization of neural networks.

✓ Application 2: Criterion to characterise convergence in Implicit Generative models.

Motivation 1: Gradient Descent dynamics in neural networks

Easier to characterize the gradient descent dynamics in the Mean-field limit [3, 7]



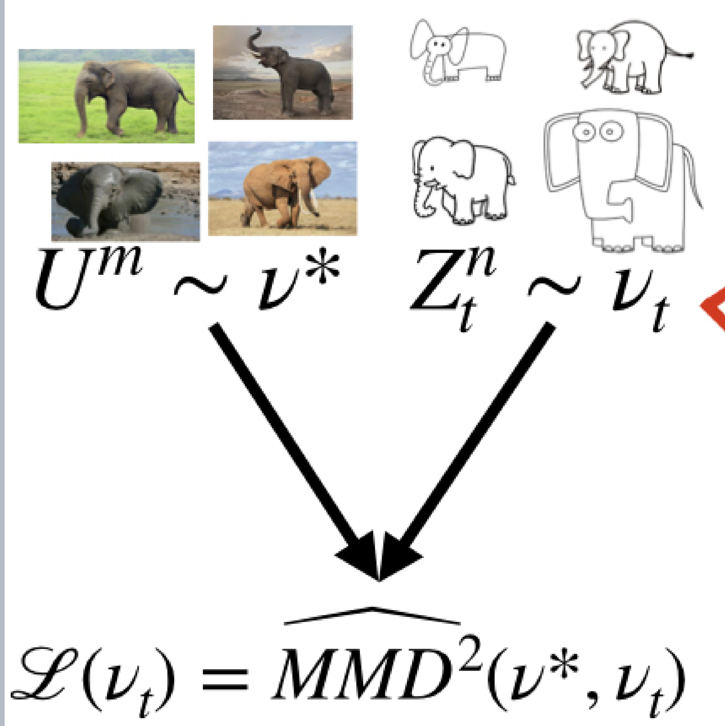
$$\min_{Z_1,\dots,Z_N} \mathbb{E}_{data}[\|y - \frac{1}{N}\sum_{i=1}^{N}\phi_{Z_i}(x)\|^2] \xrightarrow[N\to\infty]{} \min_{\nu\in\mathcal{P}} \mathbb{E}_{data}[\|y - \mathbb{E}_{Z\sim\nu}[\phi_Z(x)]\|^2]$$

In the well-specified case, i.e.: $\mathbb{E}_{data}[y|x] = \mathbb{E}_{U\sim\nu^*}[\phi_U(x)]$, equivalent to minimizing the MMD with a random feature kernel k:

$$\min_{\nu\in\mathcal{P}} MMD_k^2(\nu^*,\nu), \qquad k(Z,Z') = \mathbb{E}_{data}[\phi_Z(x)^\top \phi_{Z'}(x)]$$

Motivation 2: Implicit Generative models

- Good performance for Implicit generative models using the MMD as a loss [5, 2, 1].
- Hard to characterize global convergence.



Parametric

$$\begin{cases} Z_t^n = G_{\psi_t}(s^n) \\ s^n \sim p \end{cases}$$

$$\psi_{t+1} = \psi_t - \gamma\nabla_{\psi}\mathcal{L}(\nu_t)$$

Non Parametric

$$Z_{t+1}^n = Z_t^n - \gamma\nabla_{Z^n}\mathcal{L}(\nu_t)$$

- Easier to analyse the non-parametric setting.

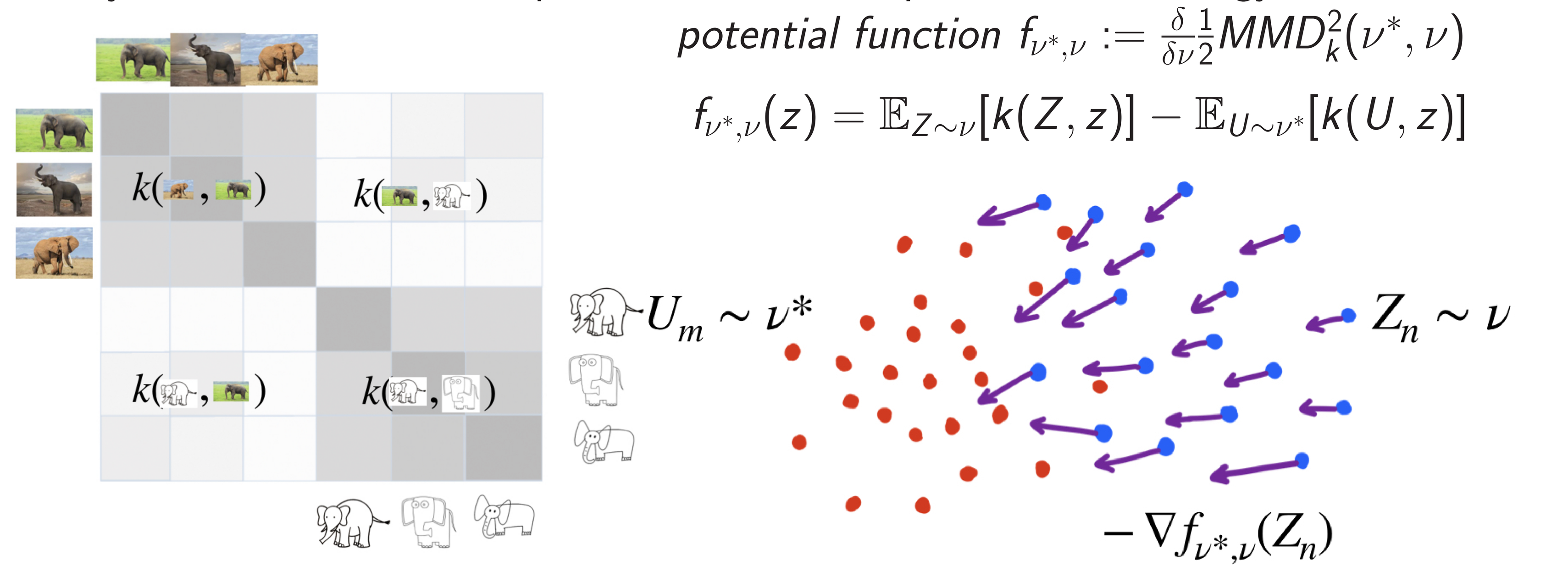
Maximum Mean Discrepancy (MMD)

The MMD is a distance between probability distributions defined using a positive semi-definite kernel k [4]:

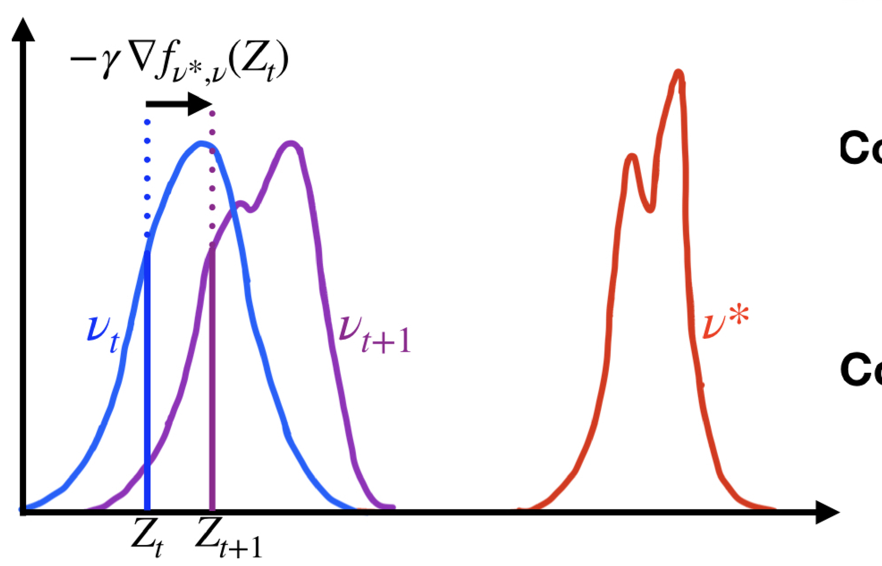
$$\frac{1}{2}MMD_k^2(\nu^*,\nu) = \frac{1}{2}\mathbb{E}_{Z,Z'\sim\nu}[k(Z,Z')] - \mathbb{E}_{\substack{Z\sim\nu \\ U\sim\nu^*}}[k(Z,U)] + \frac{1}{2}\mathbb{E}_{U,U'\sim\nu^*}[k(U,U')]$$

✓ Easy to estimate from samples: ✓ Can be interpreted as the energy relative to a *potential function* $f_{\nu^*,\nu} := \frac{\delta}{\delta\nu}\frac{1}{2}MMD_k^2(\nu^*,\nu)$

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



Gradient Flow of the MMD



Discrete time:

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

$\downarrow \gamma \to 0$

Continuous time:

$$dZ_t = -\nabla f_{\nu^*,\nu_t}(Z_t), \quad Z_t \sim \nu_t$$

Mc-Kean Vlasov Dynamics

$\Updownarrow$

Continuity equation:

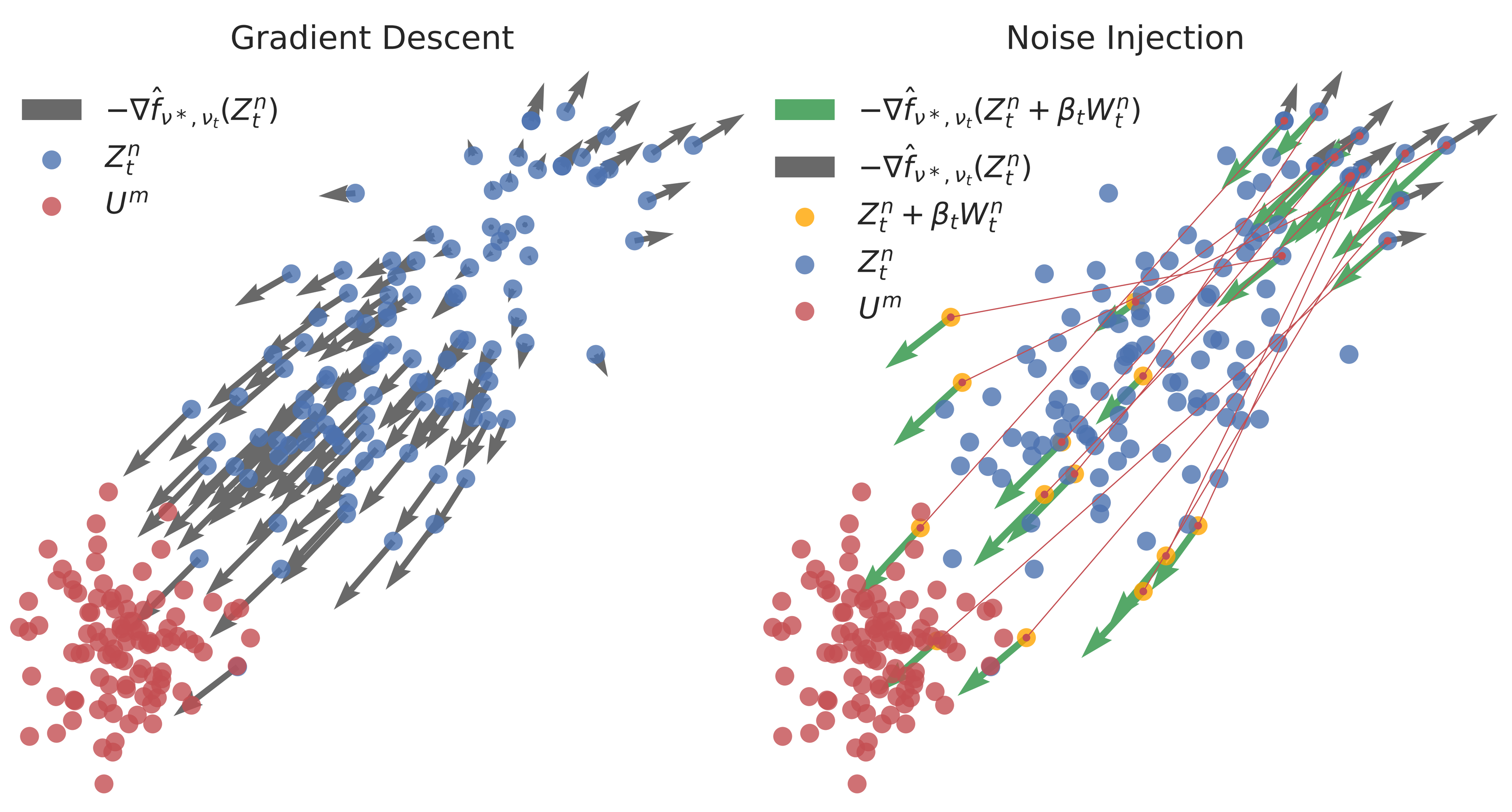
$$\partial_t \nu_t = \underbrace{div(\nu_t \nabla f_{\nu^*,\nu_t})}_{-\nabla_{\nu_t}\frac{1}{2}MMD^2(\nu^*,\nu_t)}$$

- Equivalent to Gradient descent when ν_t restricted to the form: $= \frac{1}{N}\sum_{n=1}^{N}\delta_{Z_t^n}$.
- Stationarity distribution ν_∞ satisfies: $\mathbb{E}_{Z\sim\nu_\infty}[\|\nabla f_{\nu^*,\nu_\infty}(Z)\|^2] = 0$

Noise injection (NI)

- Evaluate $\nabla f_{\nu^*,\nu_t}$ outside of the support of ν_t

$$Z_{t+1} = Z_t - \gamma\nabla f_{\nu^*,\nu_t}(Z_t + \beta_t W_t), \quad Z_t \sim \nu_t, \; W_t \sim \mathcal{N}(0,1)$$



Theory: Global convergence

1. Criterion for convergence of the gradient flow: *Negative Sobolev Distance*:

$$S(\nu^*|\nu_t) := \sup_{\substack{g\in\mathcal{L}_{\in}(\nu_{\sqcup}) \\ \|\nabla g\|_{L_2(\nu_t)}\leq 1}} \mathbb{E}_{Z\sim\nu^*}[g(Z)] - \mathbb{E}_{Z\sim\nu_t}[g(Z)]$$

Assume that $S(\nu^*|\nu_t) \leq C$ for all t, and that k is a characteristic kernel, then ν_t converges weakly towards ν^*. Moreover:

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

2. Criterion for convergence of the noise injection algorithm:
   - Decreasing direction:

$$4\gamma^2\beta_t^2 MMD^2(\nu^*,\nu_t) \leq \mathbb{E}_{\substack{Z\sim\nu_t \\ W\sim\mathcal{N}(0,1)}}[\|\nabla f_{\nu^*,\nu_t}(Z+\beta_t W)\|^2]$$

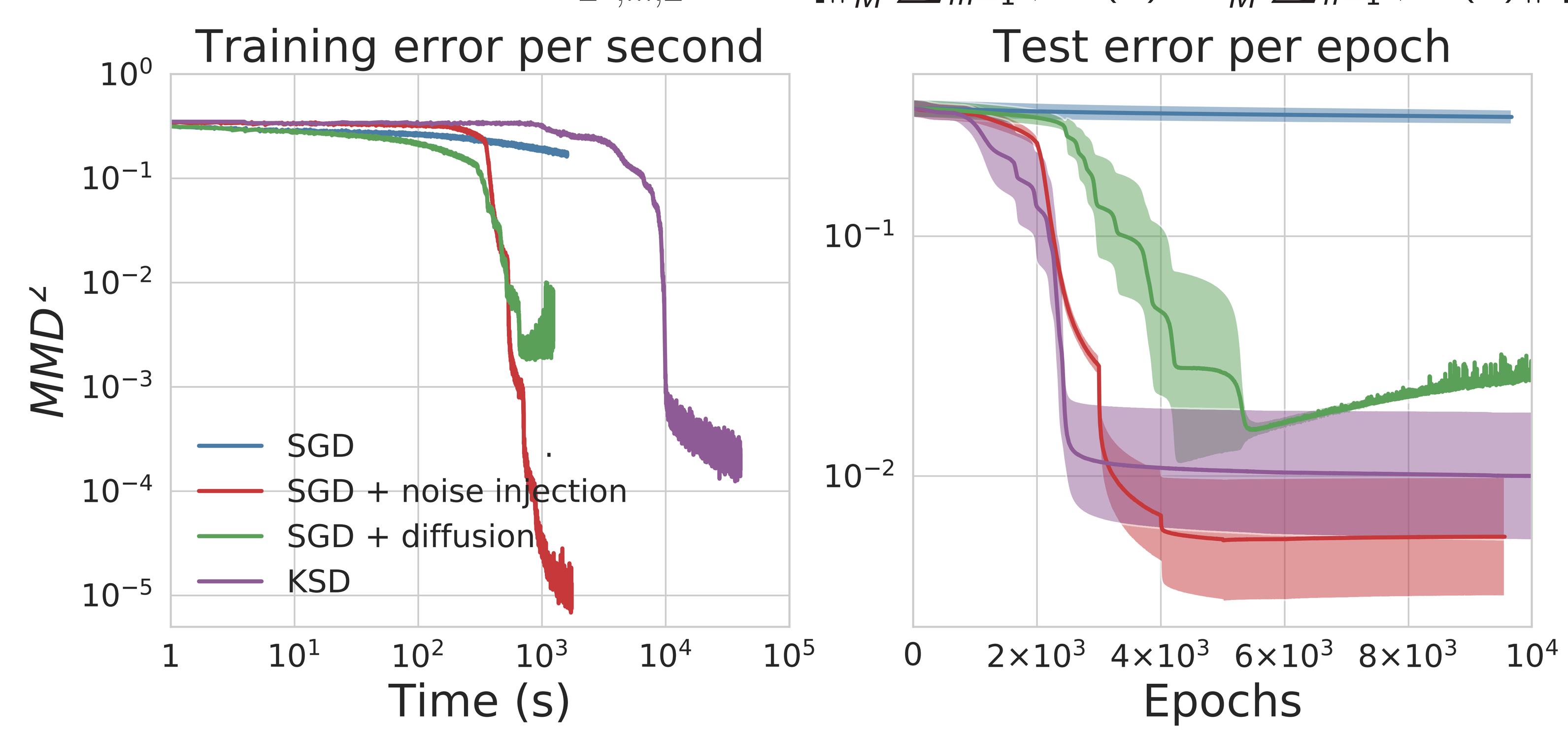
   - Large noise: $\sum_{i=0}^{t}\beta_i^2 \to \infty$

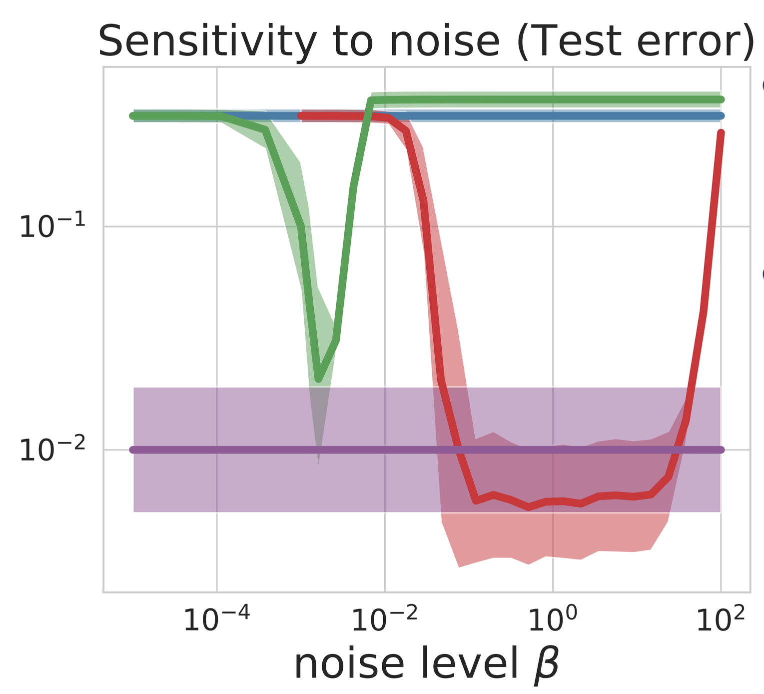
Then for some constant L:

$$MMD^2(\nu^*,\nu_t) \leq MMD(\nu^*,\nu_0)e^{-4\gamma^2(1-3\gamma L)\sum_{i=0}^{t}\beta_i^2}$$

Experimental Comparison

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





- Kernel estimated on mini-batch of data: $\hat{k}(Z,Z') = \frac{1}{B}\sum_{b=1}\phi_Z(x^b)^\top\phi_{Z'}(x^b)$.
- Update rules for $\Delta Z_t^n := Z_{t+1}^n - Z_t^n$:

$$\begin{cases} \text{SGD:} & -\nabla\hat{f}_{\nu^*,\nu_t}(Z_t^n) \\ \text{SGD+NI:} & -\nabla\hat{f}_{\nu^*,\nu_t}(Z_t^n + \beta_t W_t^n) \\ \text{SGD+Diffusion:} & -\nabla\hat{f}_{\nu^*,\nu_t}(Z_t^n) + \beta_t W_t^n \\ \text{KSD [6]:} & -\hat{C}_t^{-1}\nabla\hat{f}_{\nu^*,\nu_t}(Z_t^n) \end{cases}$$

Bibliography

M. Arbel, D. J. Sutherland, M. Bińkowski, and A. Gretton. "On gradient regularizers for MMD GANs". *NIPS* (2018).

M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. "Demystifying MMD GANs". *ICLR*. 2018.

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

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

Y. Li, K. Swersky, and R. Zemel. "Generative Moment Matching Networks". *ICML*. 2015.

Y. Mroueh, T. Sercu, and A. Raj. *Sobolev Descent: Variational Transport of Distributions via Advection*. Private communication. Apr. 2018.

G. M. Rotskoff and E. Vanden-Eijnden. "Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error". *arXiv preprint arXiv:1805.00915* (2018).

This work was funded by the Gatsby Charitable Foundation.

Contact: michael.n.arbel@gmail.com