# 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

CMStatistics, "Optimal Transport and Statistics"
December 16, 2019

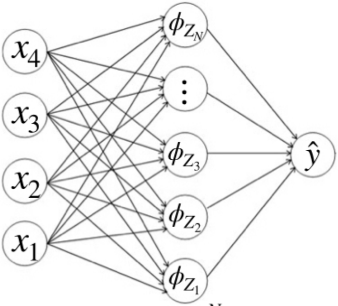
# Outline

- General problem $\rightarrow$ minimization of the MMD
- Wasserstein gradient flow of the MMD
- A Criterion for global convergence
- A noise-injection algorithm for better convergence

# General problem

Finite dimensional non-convex optimization (regression setting):

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



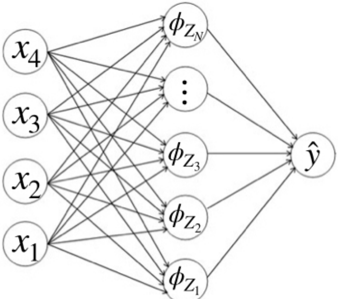

$$\min_{Z_1,\dots,Z_N} \mathbb{E}_{data}[\|y - \frac{1}{N}\sum_{i=1}^{N} \phi_{Z_i}(x)\|^2]$$

# 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:

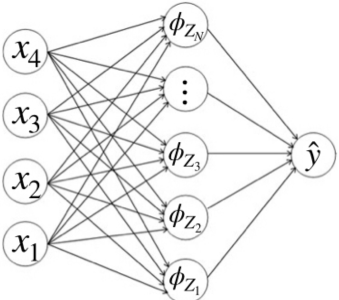
$$Z_i^{t+1} = Z_i^t - \gamma \nabla_{Z_i} \mathcal{L}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i^t}\right)$$

# General problem

Finite dimensional non-convex optimization (regression setting):

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

$(x, y) \sim data$




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

- Optimization using gradient descent GD:

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

- Hard to describe the dynamics of GD!

# General problem

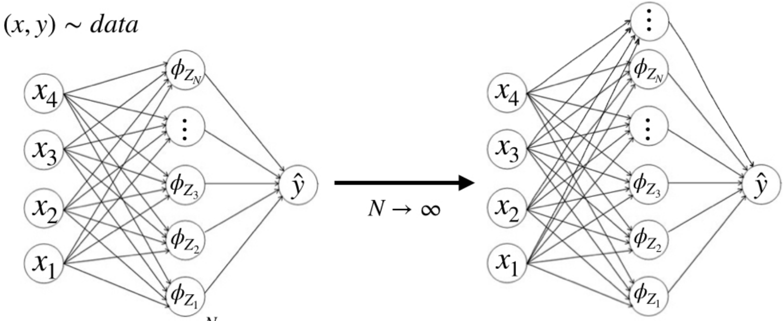
Infinite dimensional non-convex optimization [Chizat and Bach, 2018], [Mei et al., 2018]:

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

# General problem

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

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

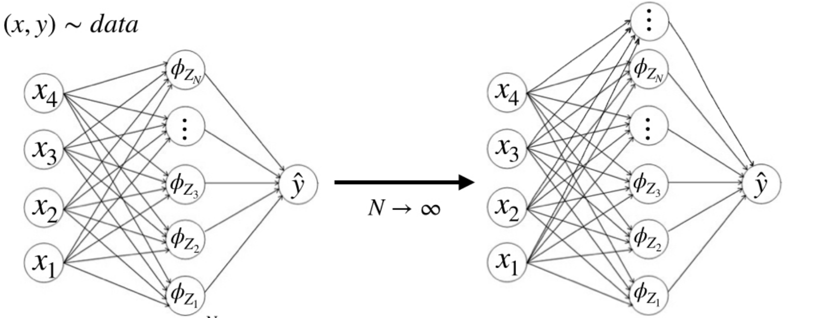



$$\min_{Z_1,\ldots,Z_N} \mathbb{E}_{data}[\|y - \frac{1}{N}\sum_{i=1}^{N}\phi_{Z_i}(x)\|^2] \xrightarrow[N\to\infty]{} \min_{\nu\in\mathscr{P}} \mathbb{E}_{data}[\|y - \mathbb{E}_{Z\sim\nu}[\phi_Z(x)]\|^2]$$

# General problem

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

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




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

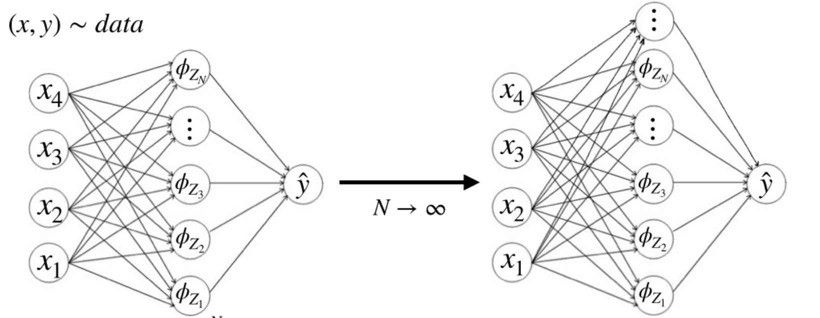
[1] [Rotskoff and Vanden-Eijnden, 2018, Chizat and Bach, 2018]

# General problem

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

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

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

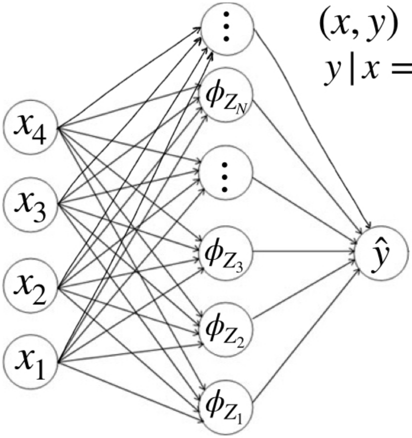



$$\min_{Z_1,\dots,Z_N} \mathbb{E}_{data}[\|y - \frac{1}{N}\sum_{i=1}^{N} \phi_{Z_i}(x)\|^2] \xrightarrow[N \to \infty]{} \min_{\nu \in \mathscr{P}} \mathbb{E}_{data}[\|y - \mathbb{E}_{Z\sim\nu}[\phi_Z(x)]\|^2]$$

---

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

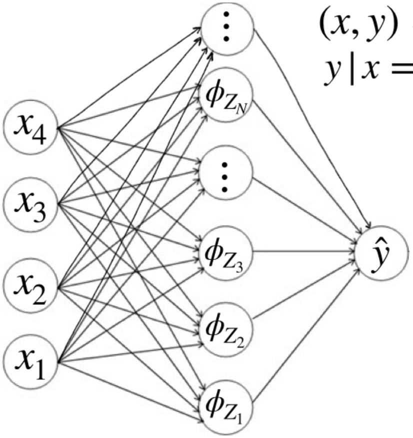
# General problem




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

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

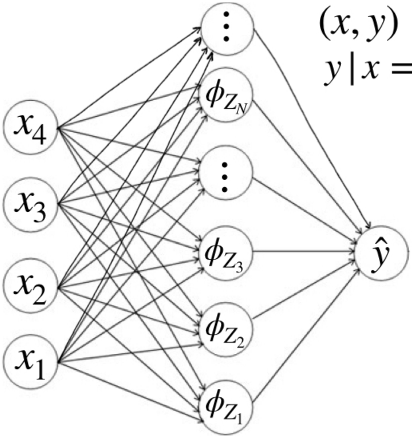
# General problem




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

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

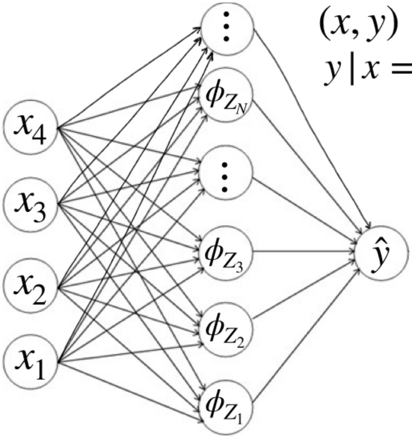
# General problem




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

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

# General problem

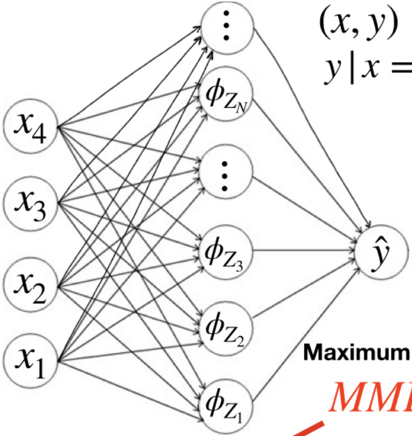



$$(x, y) \sim data$$
$$y \,|\, x = \mathbb{E}_{U \sim \nu^*}[\phi_U(x)]$$

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

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

# General problem




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

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

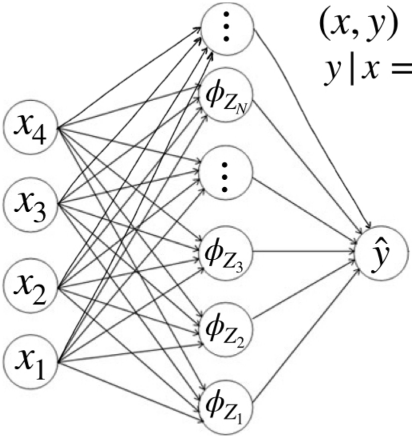
**Maximum Mean Discrepancy**

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

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

[Gretton et al., 2012]

# General problem



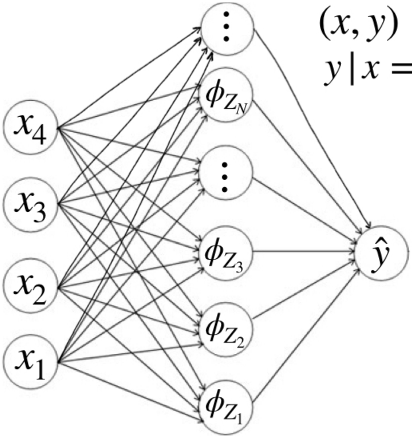

$$(x, y) \sim data$$

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

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

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

# General problem



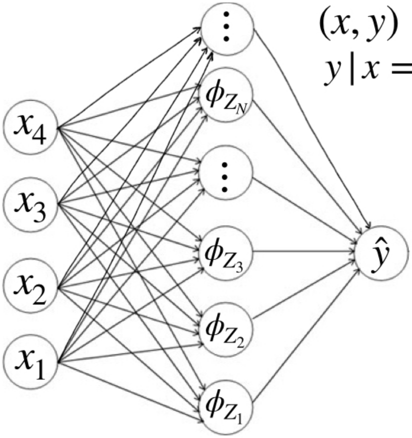

$$(x, y) \sim data$$

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

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

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

# General problem

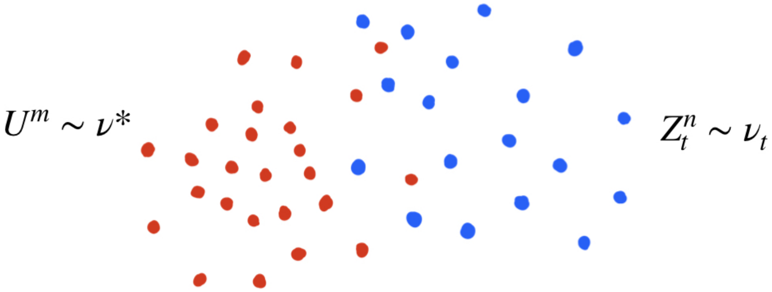



$$(x, y) \sim data$$
$$y \,|\, x = \mathbb{E}_{U \sim \nu^*}[\phi_U(x)]$$

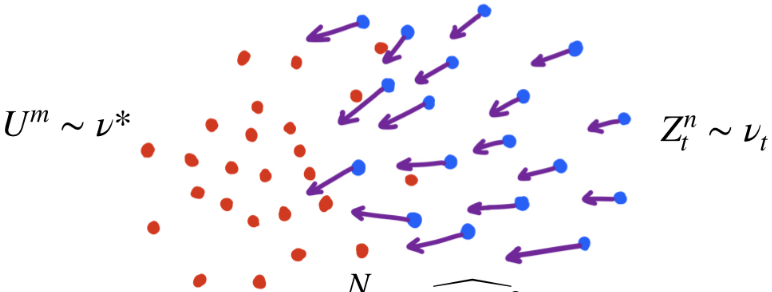
$$\min_{\nu \in \mathscr{P}} MMD^2(\nu^*, \nu)$$

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

# Gradient descent



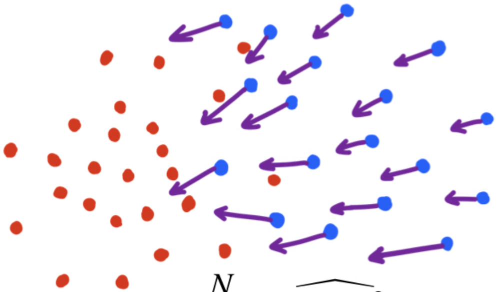
# Gradient descent




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

# Gradient descent

$U^m \sim \nu^*$

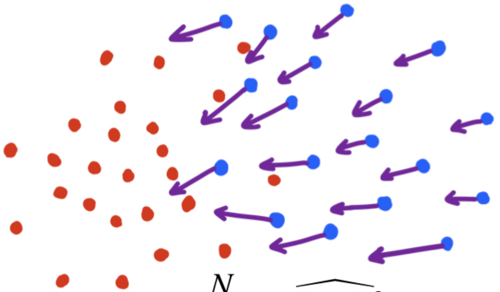


$Z_t^n \sim \nu_t$

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

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

# Gradient descent




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

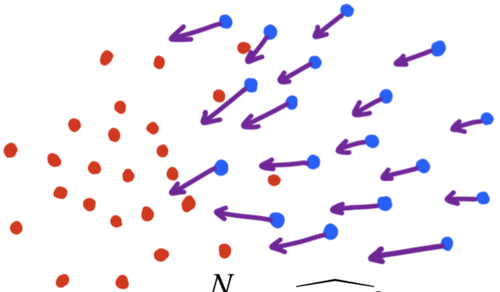
$$\frac{N}{2} \nabla_{Z_t^n} \widehat{MMD}^2(\nu^*, \nu_t) = \frac{1}{N-1} \sum_{n' \neq n} \partial_1 k(Z_t^n, Z_t^{n'}) - \frac{1}{M} \sum_{m=1}^{M} \partial_1 k(Z_t^n, U^m)$$

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

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

# Gradient descent

$U^m \sim \nu^*$



$Z_t^n \sim \nu_t$

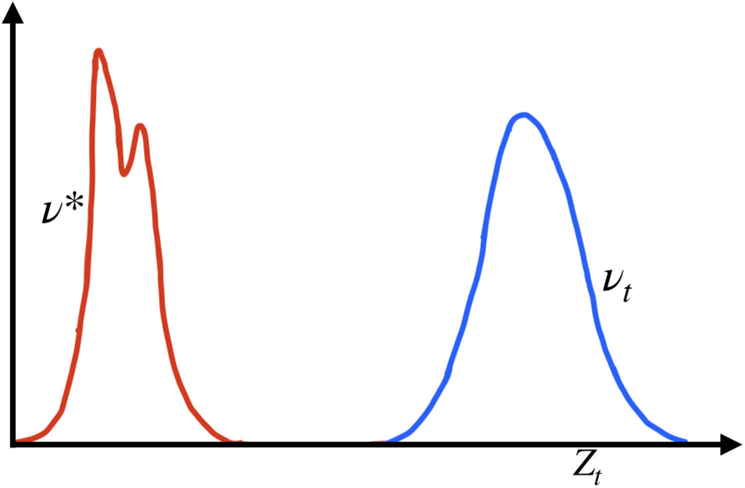
$$Z_{t+1}^n = Z_t^n - \gamma \frac{N}{2} \nabla_{Z_t^n} \widehat{MMD}^2(\nu^*, \nu_t)$$

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

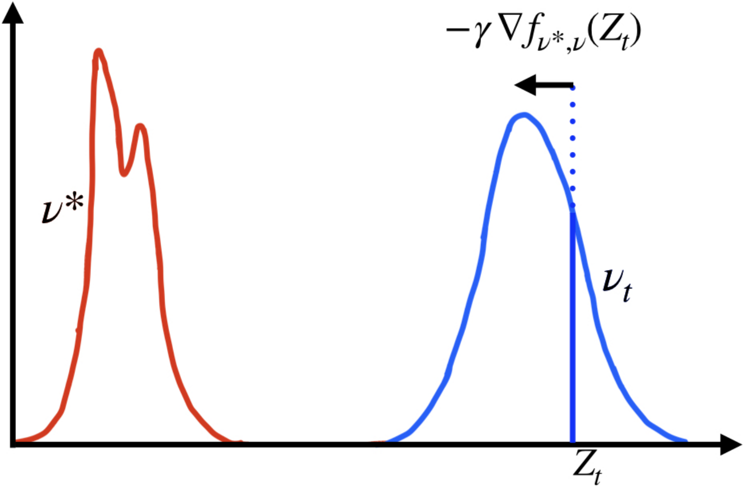
$$\Big\downarrow \; N, M \to \infty$$

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

# Wasserstein gradient descent

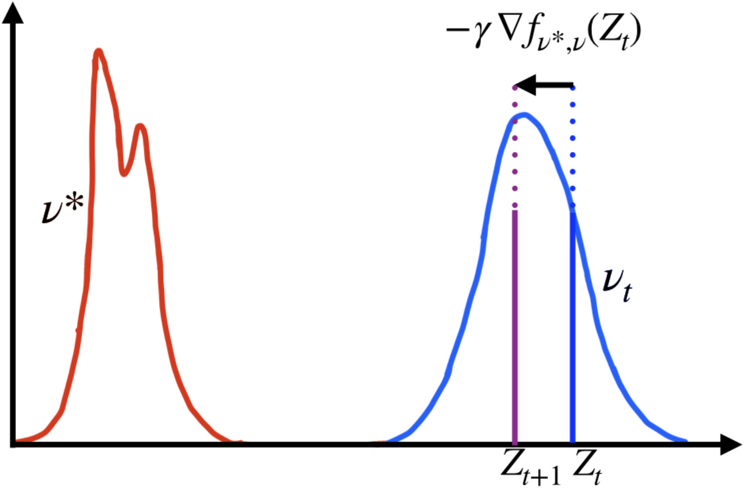


# Wasserstein gradient descent



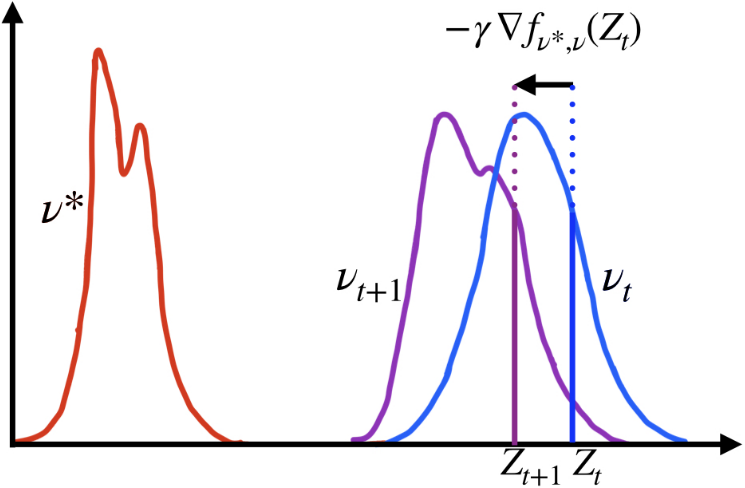

$$f_{\nu^*,\nu_t}(Z_t) = \mathbb{E}_{Z'\sim\nu_t}[k(Z_t, Z')] - \mathbb{E}_{U\sim\nu^*}[k(Z_t, U)]$$

# Wasserstein gradient descent



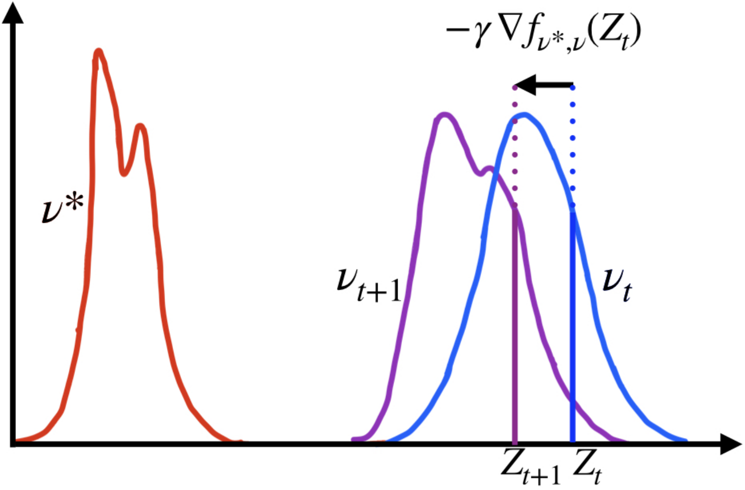

$$f_{\nu^*,\nu_t}(Z_t) = \mathbb{E}_{Z' \sim \nu_t}[k(Z_t, Z')] - \mathbb{E}_{U \sim \nu^*}[k(Z_t, U)]$$

# Wasserstein gradient descent



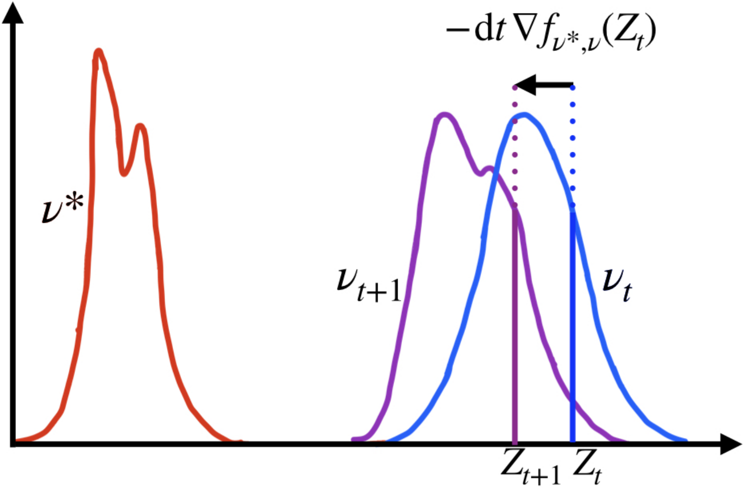

$$f_{\nu^*,\nu_t}(Z_t) = \mathbb{E}_{Z' \sim \nu_t}[k(Z_t, Z')] - \mathbb{E}_{U \sim \nu^*}[k(Z_t, U)]$$

# 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

# Wasserstein gradient flow

- Continuous time equation: Mc-Kean Vlasov dynamics

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

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

# Wasserstein gradient flow

- Continuous time equation: Mc-Kean Vlasov dynamics

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

- Equivalent to a PDE in $\nu_t$:

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

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

# Wasserstein gradient flow

- Continuous time equation: Mc-Kean Vlasov dynamics

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

- Equivalent to a PDE in $\nu_t$:

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

- Interpretation as a gradient flow in probability space [2]:

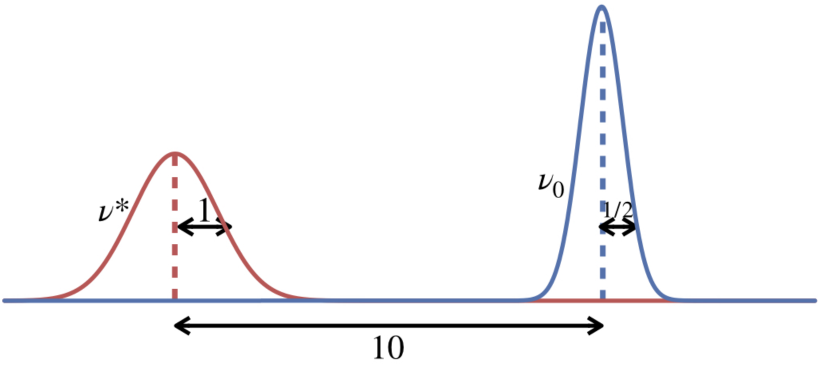
$$\partial_t \nu_t = -\nabla_{\nu_t} \mathcal{L}(\nu_t) \qquad \mathcal{L}(\nu) := \frac{1}{2} MMD^2(\nu^*, \nu)$$

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

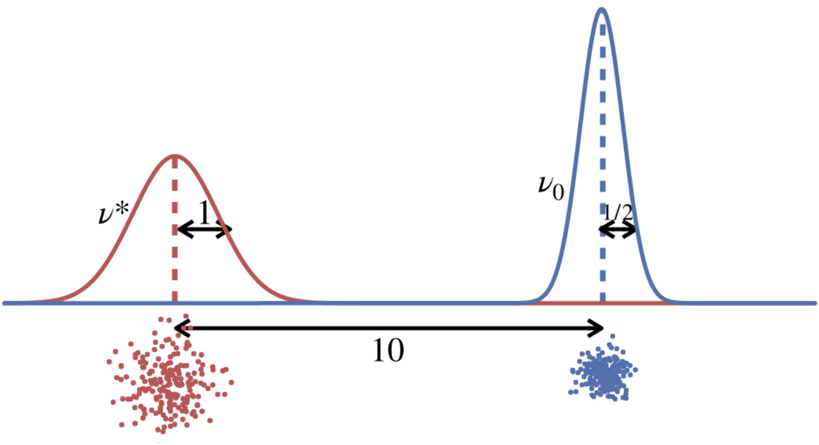
$$\nu_{t+1} \in \arg\min_{\nu} \mathcal{L}(\nu) + \frac{1}{2\tau} W_2^2(\nu, \nu_t).$$

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

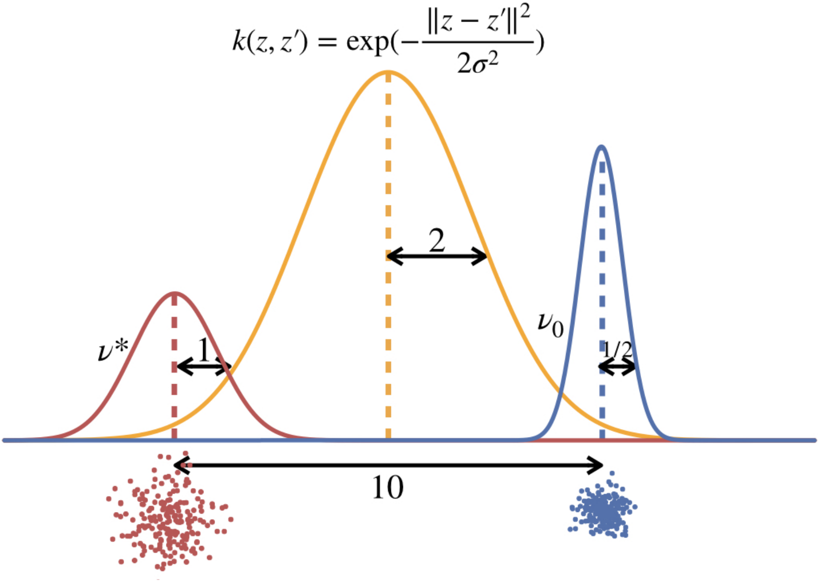
# Convergence: Failure case



# Convergence: Failure case



# Convergence: Failure case



# Convergence: Failure case

# A criterion for convergence

- Define the Negative Sobolev distance:

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

# A criterion for convergence

- Define the Negative Sobolev distance:

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

- Assume that $S(\nu^*|\nu_t) \leq C$ for all $t$, then for $\gamma$ small enough

$$MMD^2(\nu^*, \nu_t) \leq \frac{1}{MMD^2(\nu^*, \nu_0) + 8\gamma C^{-1} t}$$

# A criterion for convergence

- Define the Negative Sobolev distance:

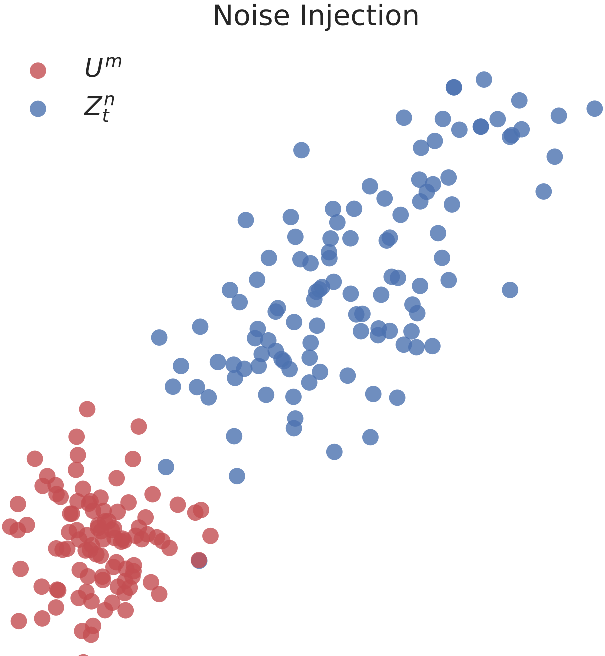
$$S(\nu^*|\nu_t) = \sup_{g, \mathbb{E}_{Z\sim\nu_t}[\|\nabla g(Z)\|^2]\leq 1} |\mathbb{E}_{Z\sim\nu_t}[g(Z)] - \mathbb{E}_{U\sim\nu^*}[g(U)]|$$

- Assume that $S(\nu^*|\nu_t) \leq C$ for all $t$, then for $\gamma$ small enough

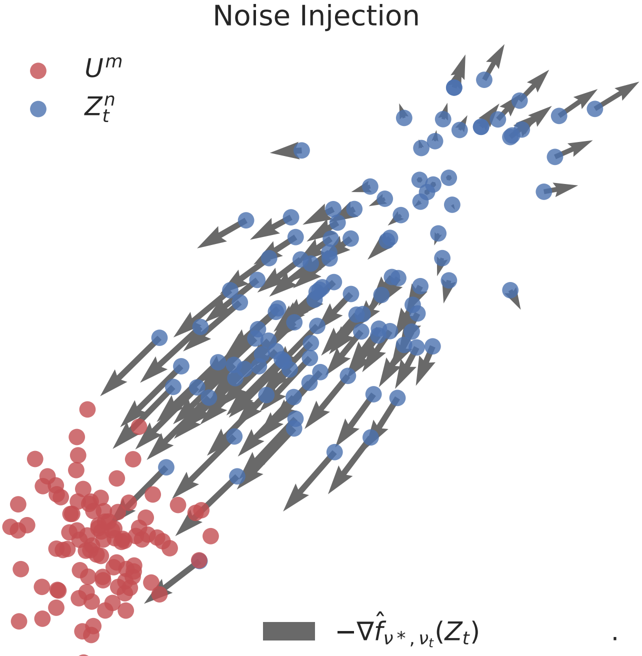
$$MMD^2(\nu^*, \nu_t) \leq \frac{1}{MMD^2(\nu^*, \nu_0) + 8\gamma C^{-1} t}$$

- Depends on the whole sequence $\nu_t$: Hard to verify in general, can only be checked for simple examples
- We've seen failure cases in practice.

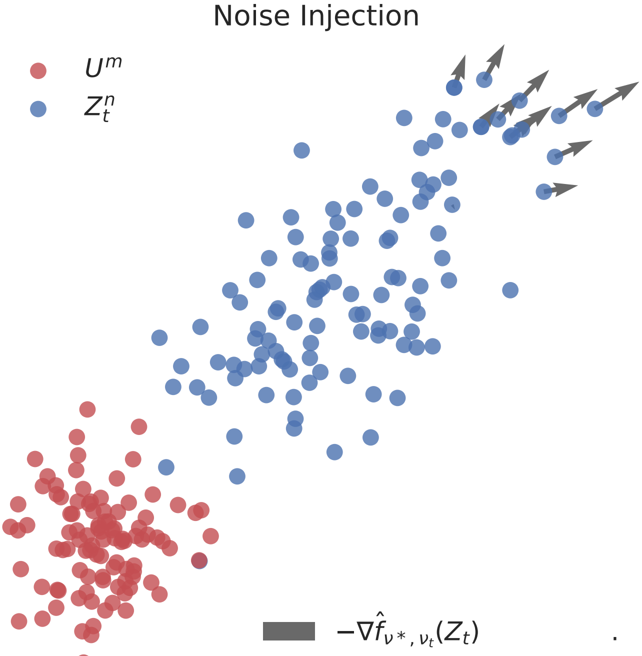
# Noise Injection



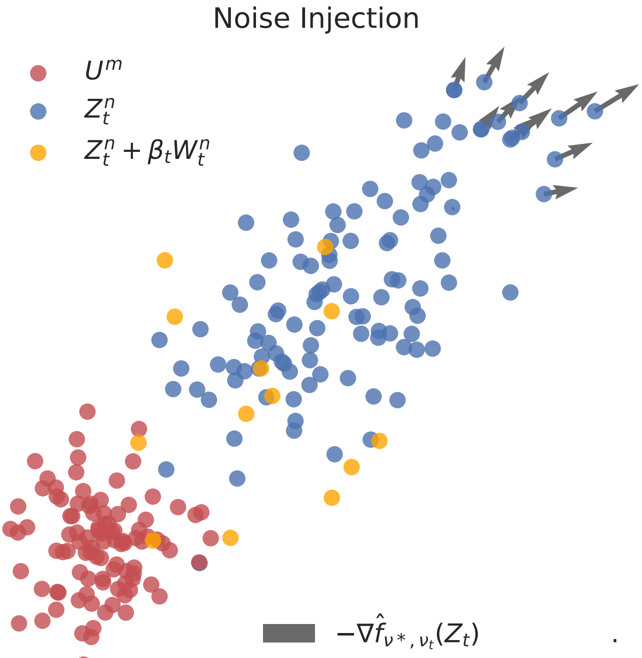
# Noise Injection



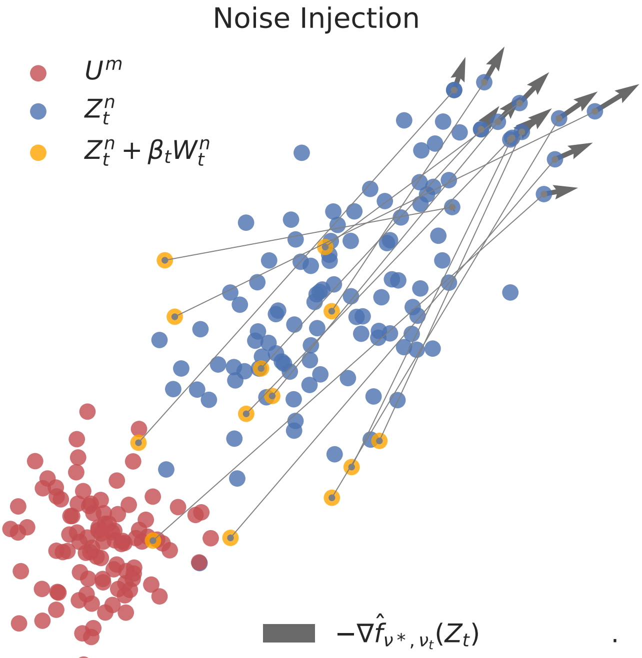
# Noise Injection



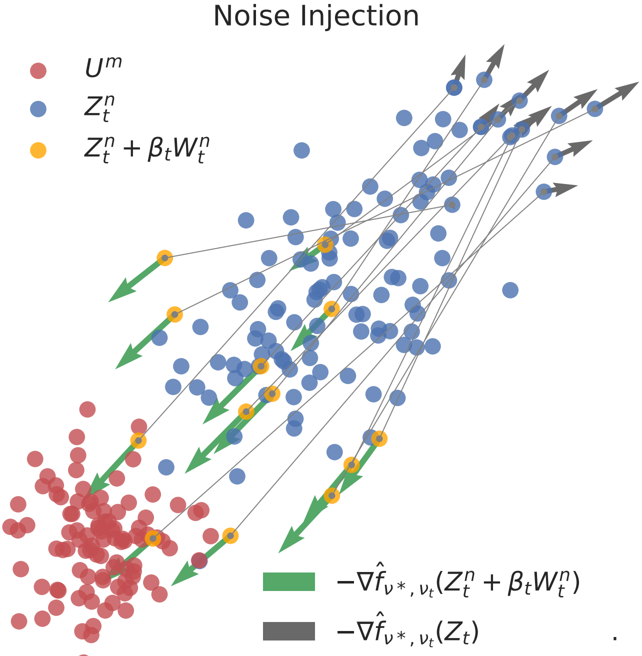
# Noise Injection



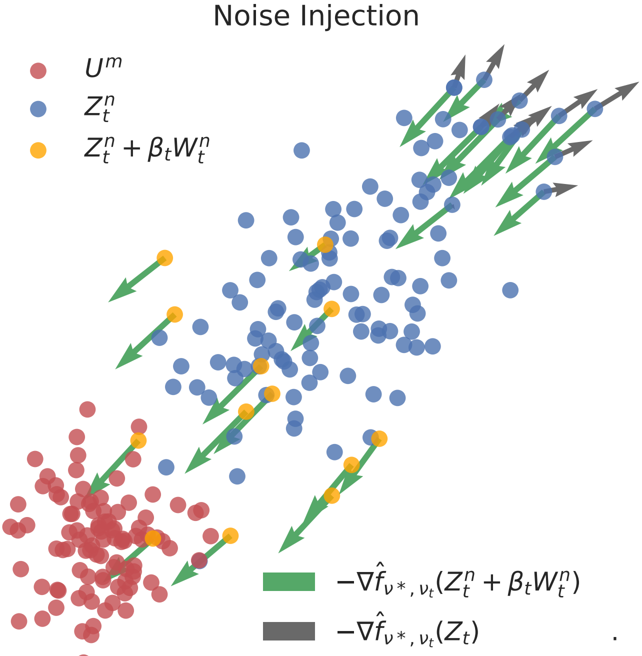
# Noise Injection



# Noise Injection

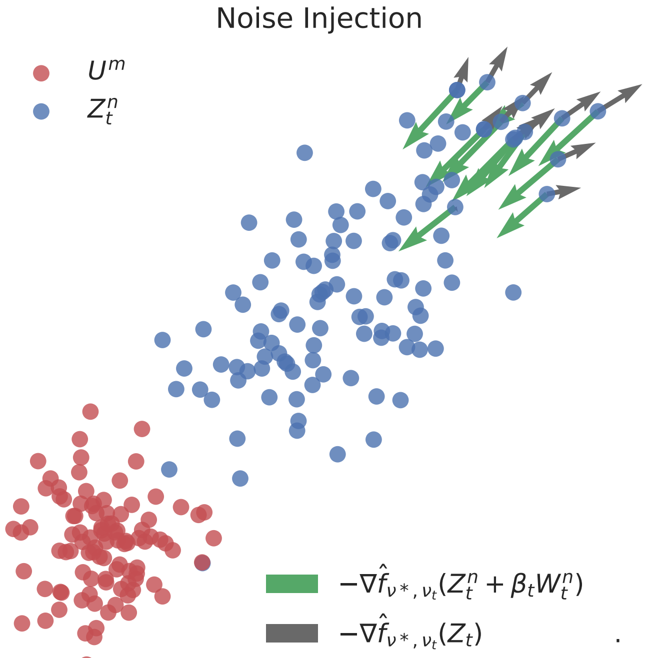


# Noise Injection

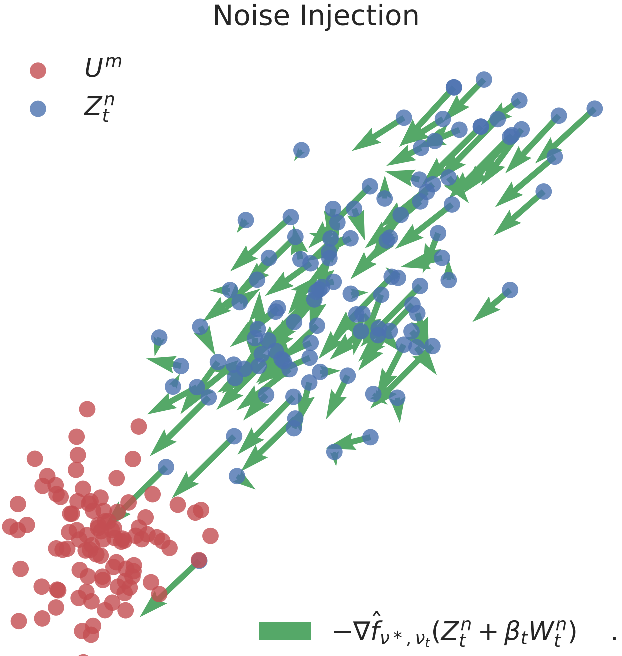



.

# Noise Injection



# Noise Injection




.

# Noise Injection

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

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

[4] [Duchi et al., 2012]

[5] [Mei et al., 2018]

# Noise Injection

- Idea: Evaluate $\nabla f_{\nu^*,\nu_t}$ outside of the support of $\nu_t$ to get a better signal!
- Sample $u_t \sim \mathcal{N}(0,1)$ and $\beta_t$ 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] [Duchi et al., 2012]

[5] [Mei et al., 2018]

# Noise Injection

- Idea: Evaluate $\nabla f_{\nu^*,\nu_t}$ outside of the support of $\nu_t$ to get a better signal!
- Sample $u_t \sim \mathcal{N}(0,1)$ and $\beta_t$ 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] or *randomized smoothing*[4] , but extended to interacting particles.

---

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

[4] [Duchi et al., 2012]

[5] [Mei et al., 2018]

# Noise Injection

- Idea: Evaluate $\nabla f_{\nu^*,\nu_t}$ outside of the support of $\nu_t$ to get a better signal!
- Sample $u_t \sim \mathcal{N}(0,1)$ and $\beta_t$ 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] or *randomized smoothing*[4] , but extended to interacting particles.
- Different from adding noise outside

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

which corresponds to an entropic regularization of the original loss [5].

---

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

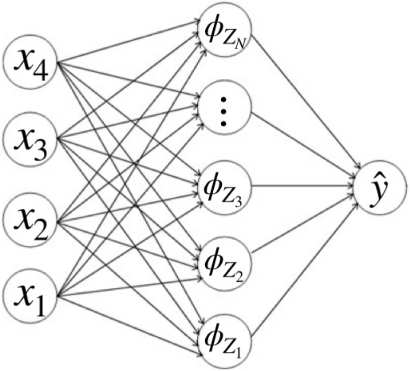
[4] [Duchi et al., 2012]

[5] [Mei et al., 2018]

# Noise Injection: Experiments

# Noise Injection: Student-Teacher network

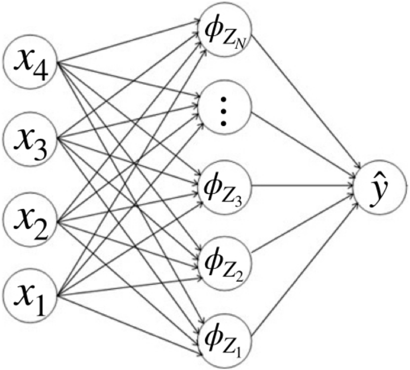
$(x, y) \sim data$




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

# Noise Injection: Student-Teacher network

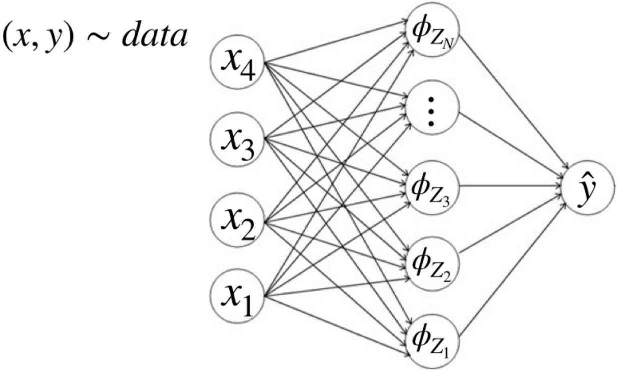
$(x, y) \sim data$




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

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

# Noise Injection: Student-Teacher network




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

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

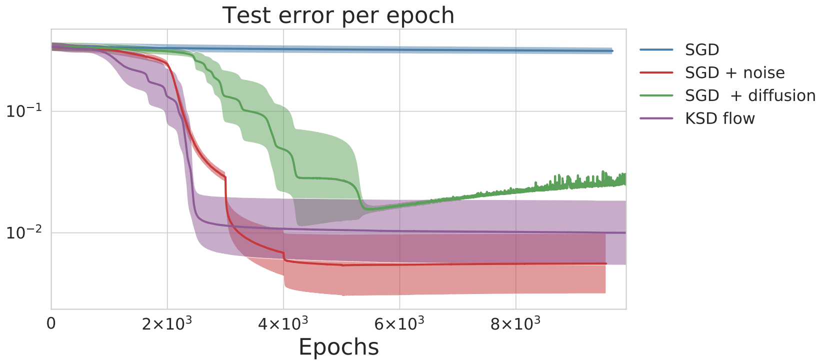
# Noise Injection: Experiments

Methods:

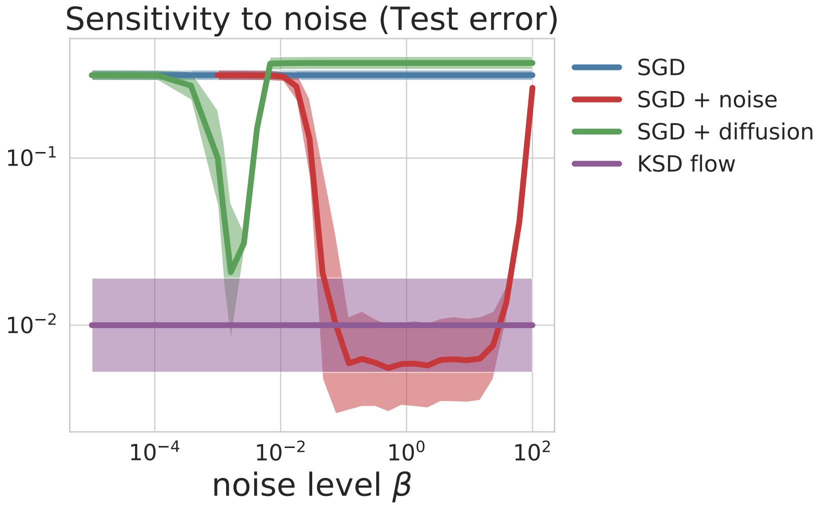
- SGD
- SGD + Noise injection
- SGD + diffusion
- KSD [6]: SGD using the Negative Sobolev distance $\nu \mapsto S(\nu^*|\nu)$ as a loss function: also minimizes the MMD.

[6] [Mroueh et al., 2019]

# Noise Injection: Experiments



# Noise Injection: Experiments



# 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.

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.
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.

# Noise Injection: Theory

Tradeoff for $\beta_t$

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

# Noise Injection: Theory

Tradeoff for $\beta_t$

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

# Noise Injection: Theory

Tradeoff for $\beta_t$

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

Need $\beta_t$ 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:

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

# Noise Injection: Theory

Tradeoff for $\beta_t$

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

Need $\beta_t$ 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:

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

Then

$$MMD^2(\nu^*, \nu_T) \leq MMD^2(\nu^*, \nu_0) e^{-C\gamma \sum_{t=1}^{T} \beta_t^2}$$