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

October 15, 2019

Outline

- General problem
- Wasserstein gradient of the MMD
- A Criterion for global convergence
- A noise-injection algorithm for better convergence

General problem

Finite dimensional non-convex optimization:

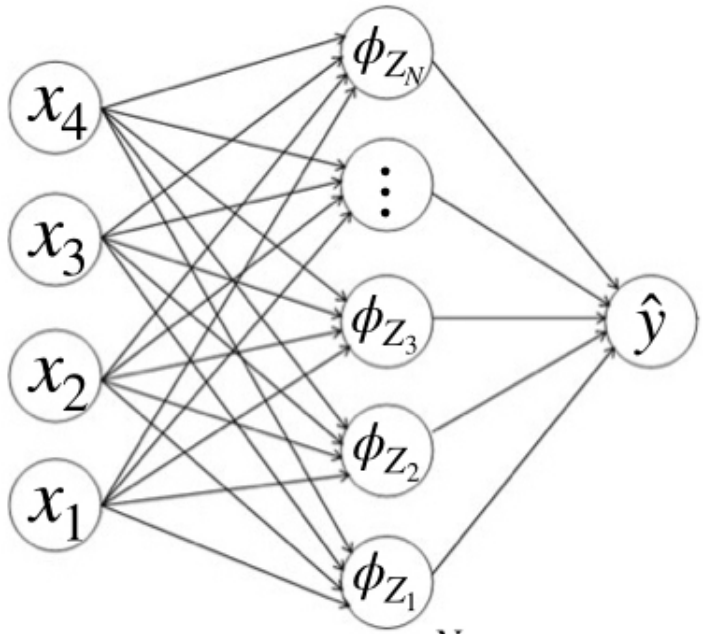
$$\min_{Z^1,\dots,Z^N \in \mathcal{Z}} \mathcal{L}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z^n}\right)$$

General problem

Finite dimensional non-convex optimization:

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

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



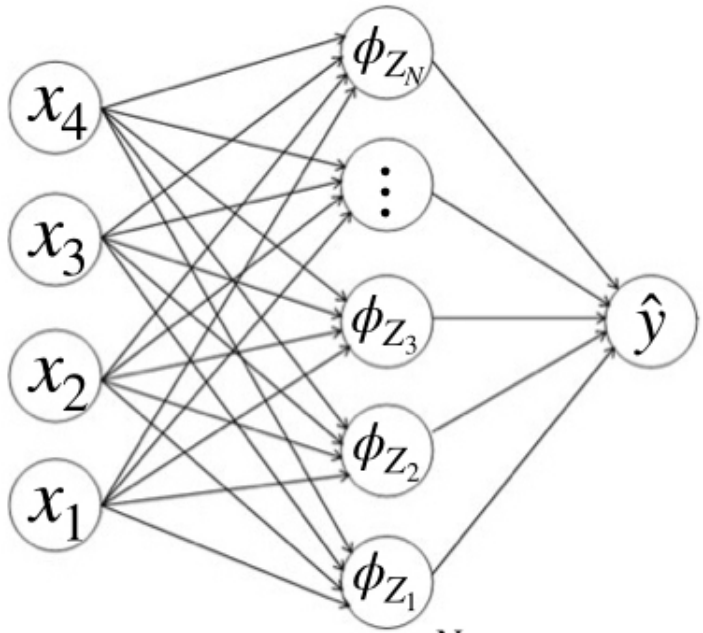
$$\min_{Z_1,\dots,Z_N} \mathbb{E}_{data}[\|y - \frac{1}{N}\sum_{i=1}^{N} \phi_{Z_i}(x)\|^2]$$

General problem

Finite dimensional non-convex optimization:

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

$(x, y) \sim data$



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

- Optimization using gradient descent GD:

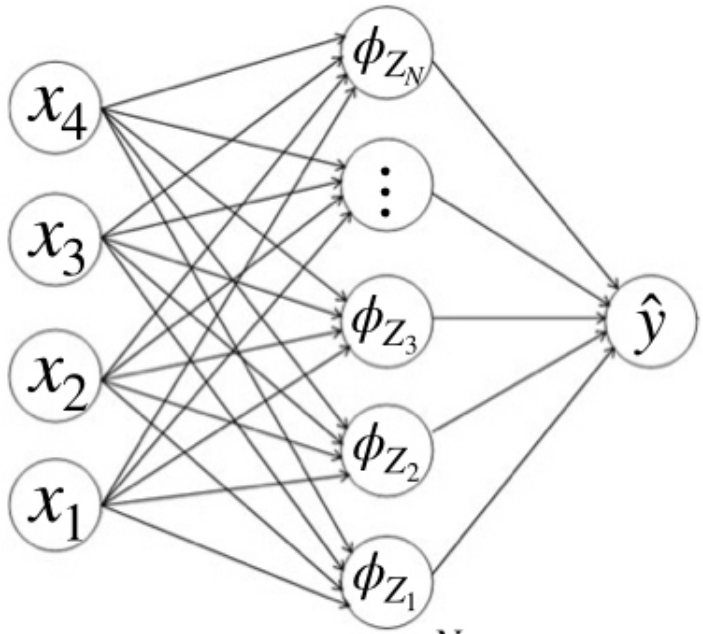
$$Z^n_{t+1} = Z^n_t - \gamma \nabla_{Z^n} \mathcal{L}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z^n_t}\right)$$

General problem

Finite dimensional non-convex optimization:

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

$(x, y) \sim data$



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

- Optimization using gradient descent GD:

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

- Hard to describe the dynamics of GD!

General problem

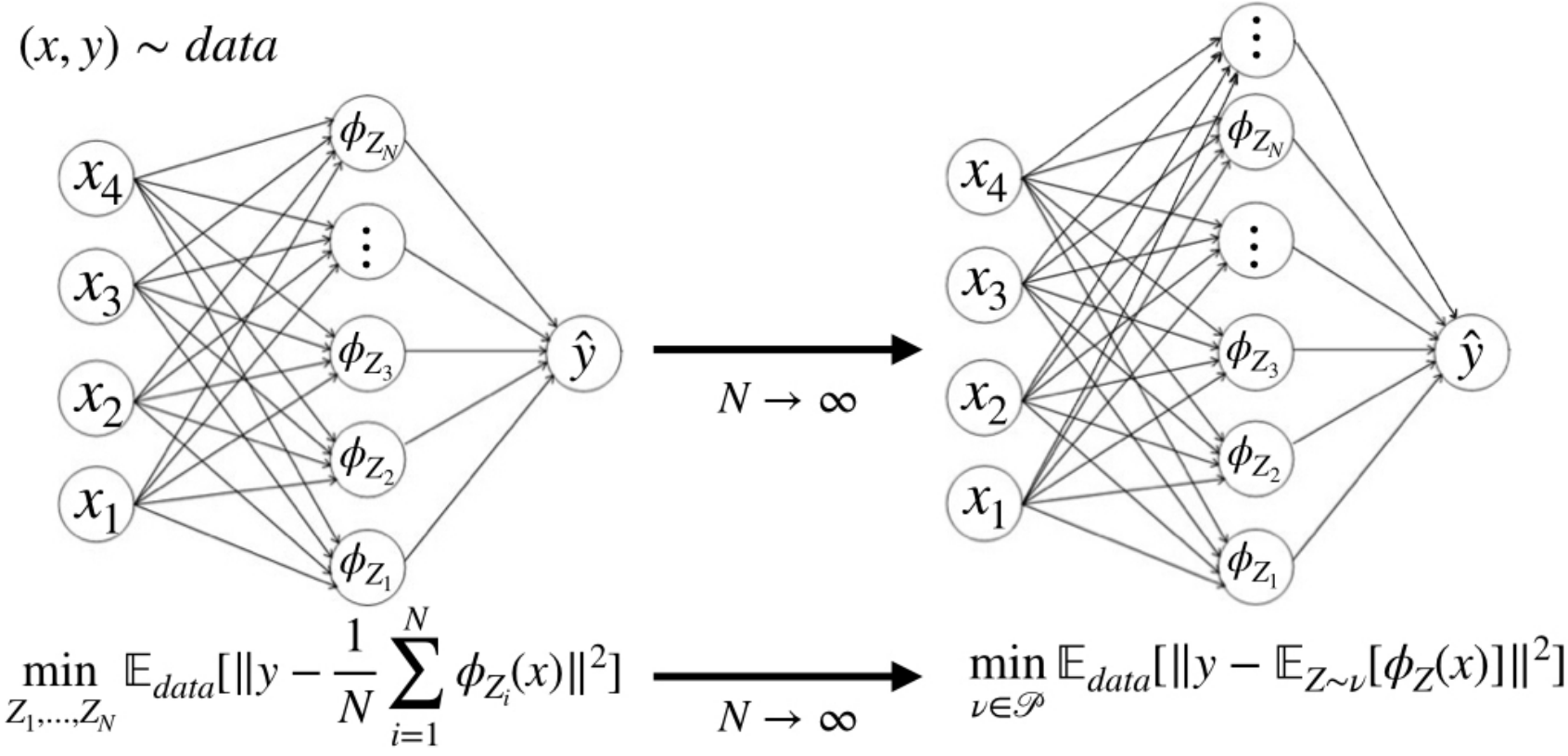
Infinite dimensional non-convex optimization:

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

General problem

Infinite dimensional non-convex optimization:

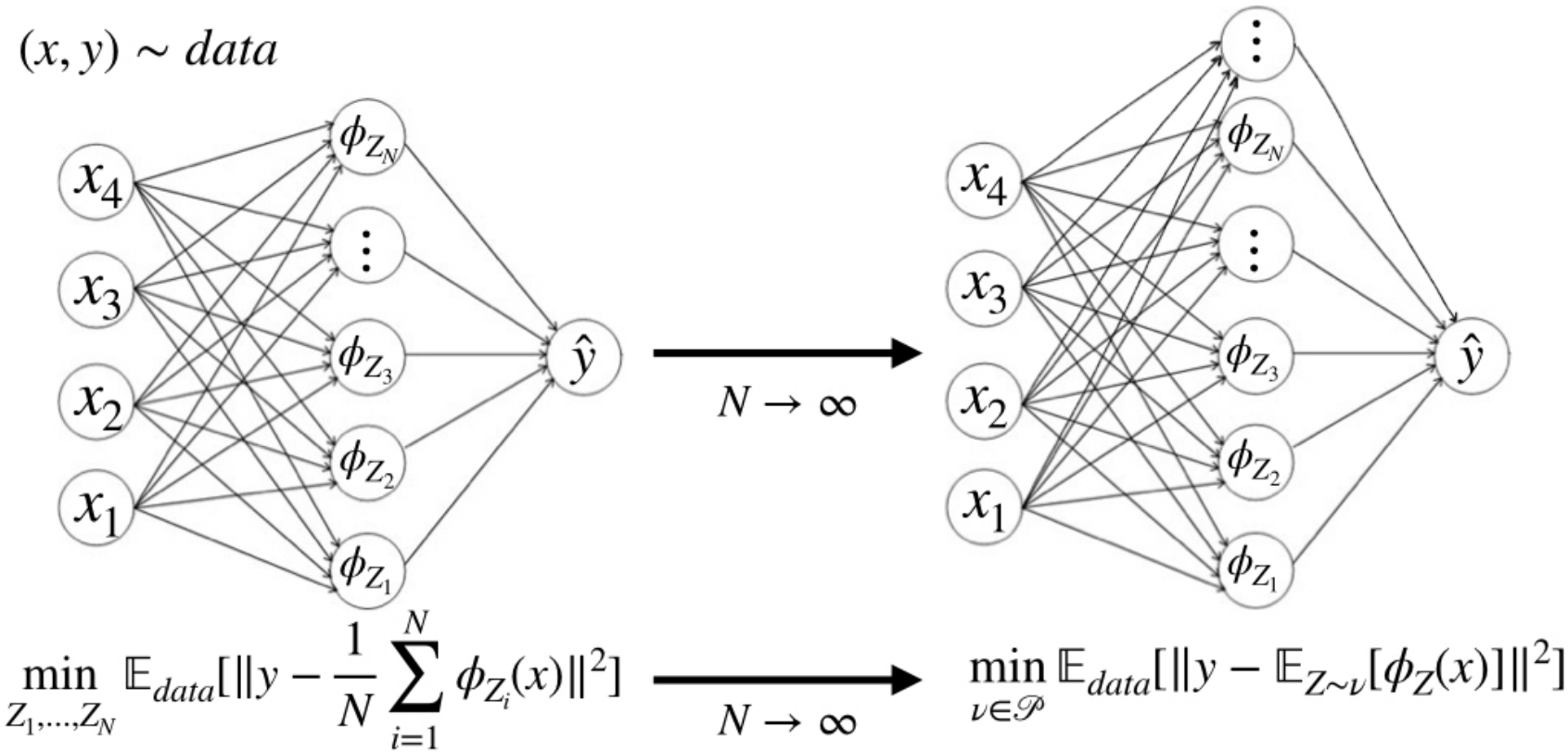
$$\min_{Z^1,\dots,Z^N \in \mathcal{Z}} \mathcal{L}\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{Z^n}\right) \xrightarrow[N\to\infty]{} \min_{\nu\in\mathcal{P}} \mathcal{L}(\nu)$$



General problem

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

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



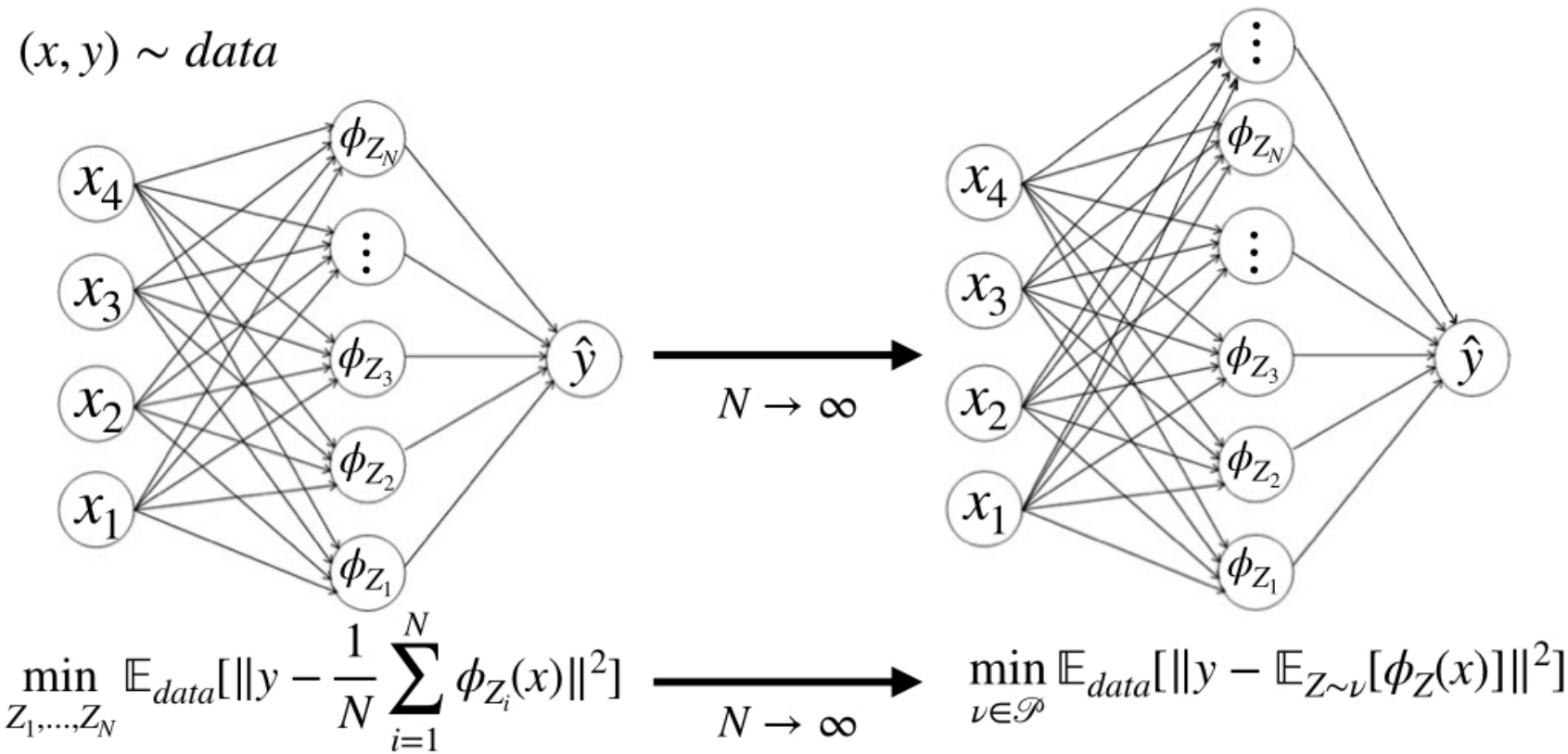
[1] [Rotskoff and Vanden-Eijnden, 2018, Chizat and Bach, 2018]

General problem

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

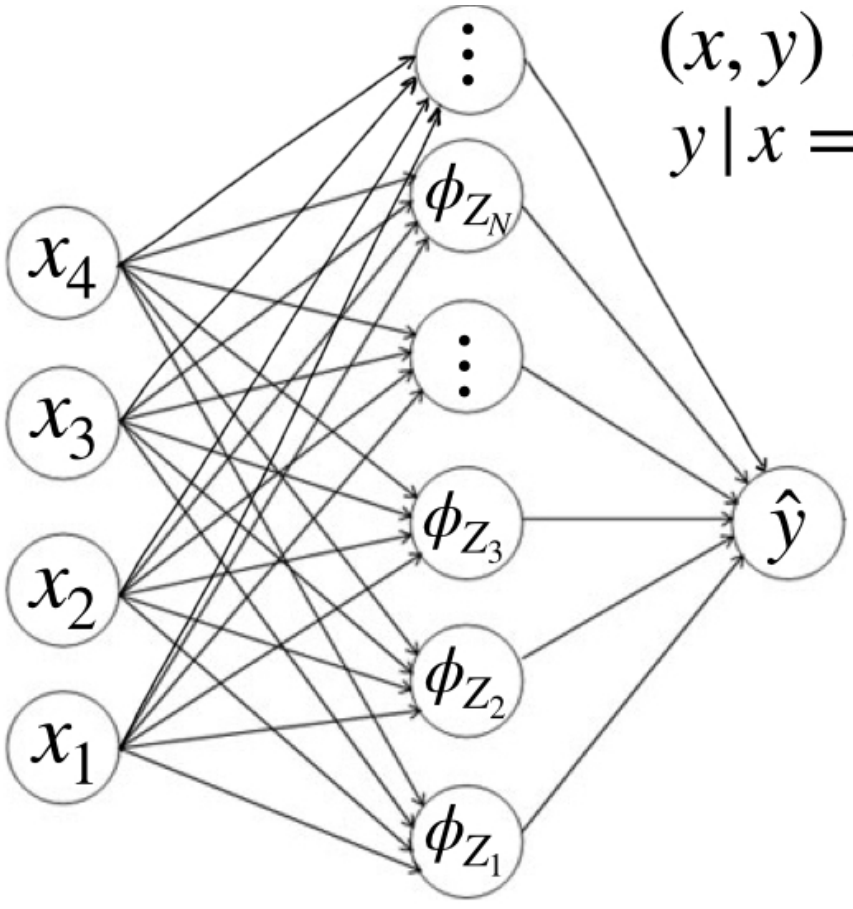
$$\phi_Z(x) = w g_\theta(x), \qquad Z = (w, \theta)$$

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



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

General problem

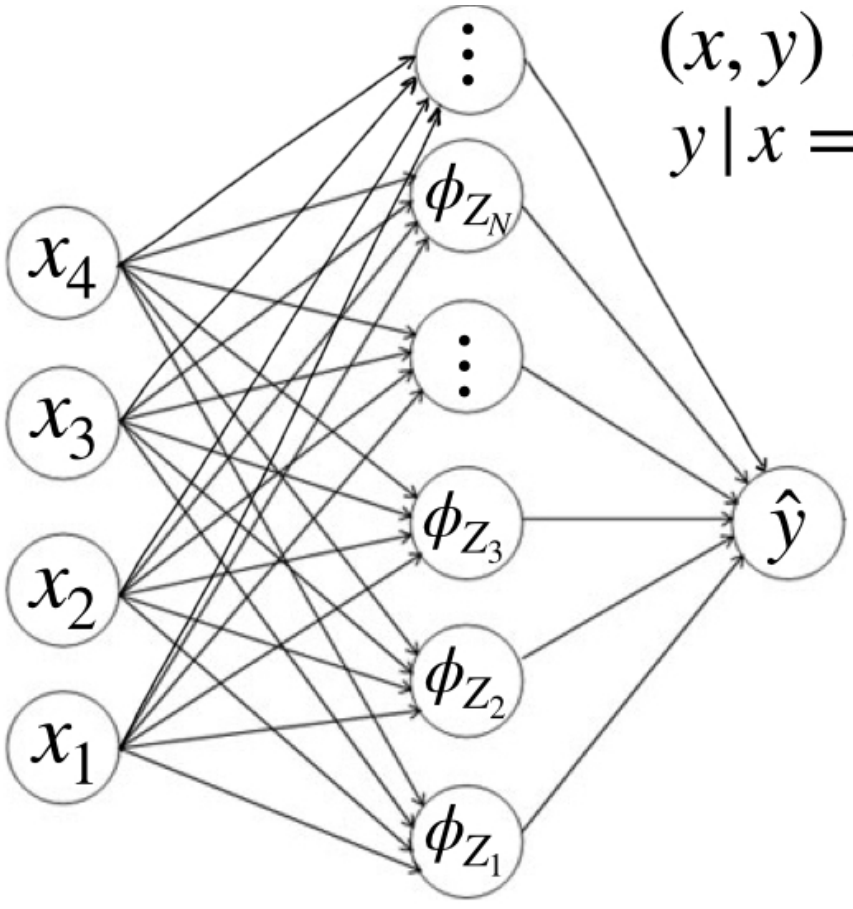


$$(x, y) \sim data$$

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

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

General problem

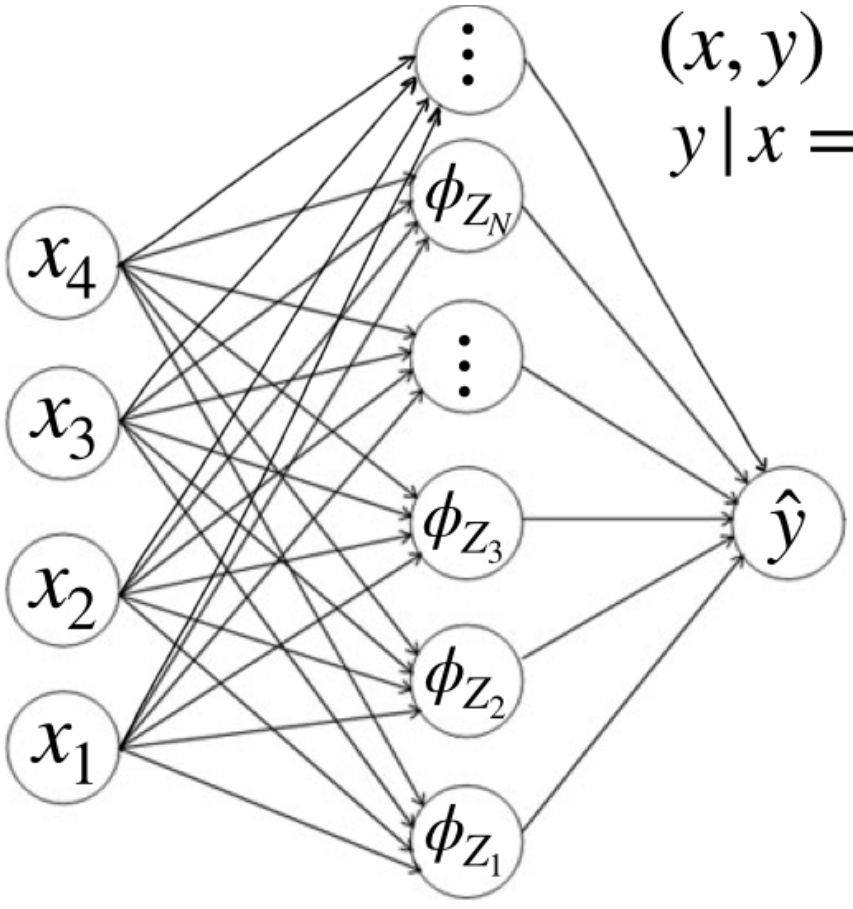


$$(x, y) \sim data$$

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

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

General problem

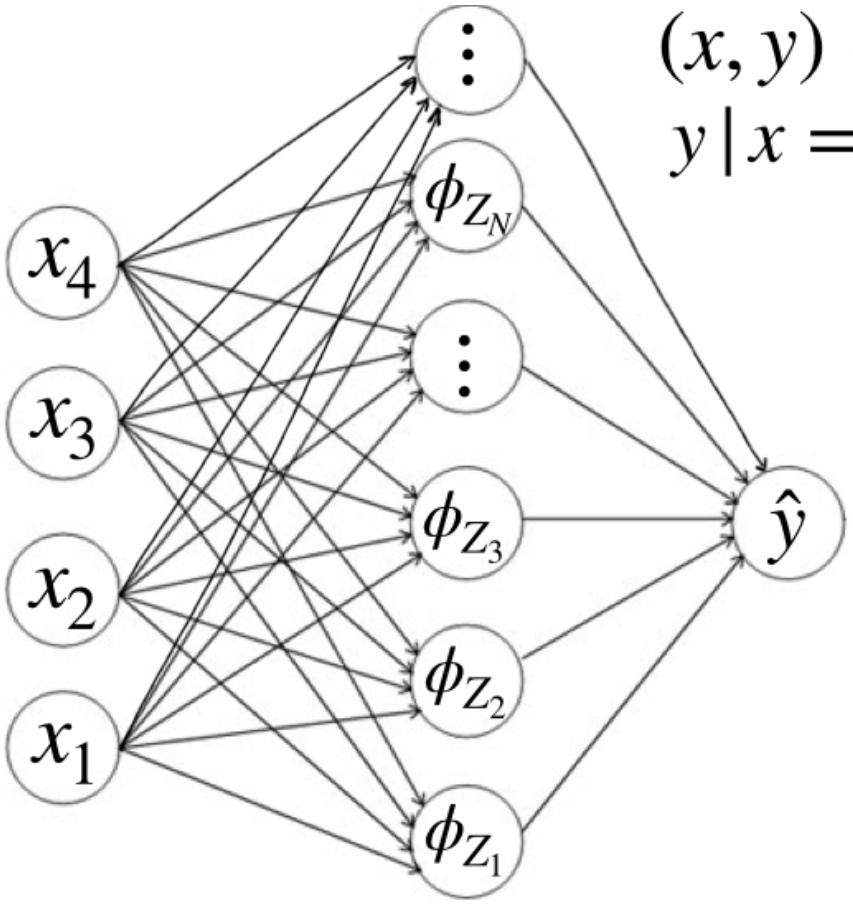


$$(x, y) \sim data$$

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

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

General problem



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

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

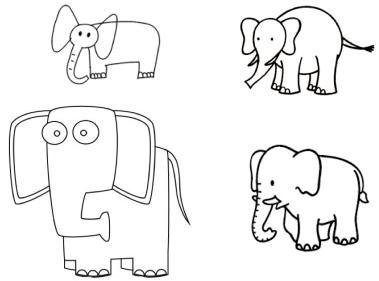
$$\min_{\nu \in \mathcal{P}} MMD^2(\nu^*, \nu)$$

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

The Maximum Mean Discrepancy[2]

Consider samples from two distributions ν^* and ν_0.

$U^m \sim \nu^*$

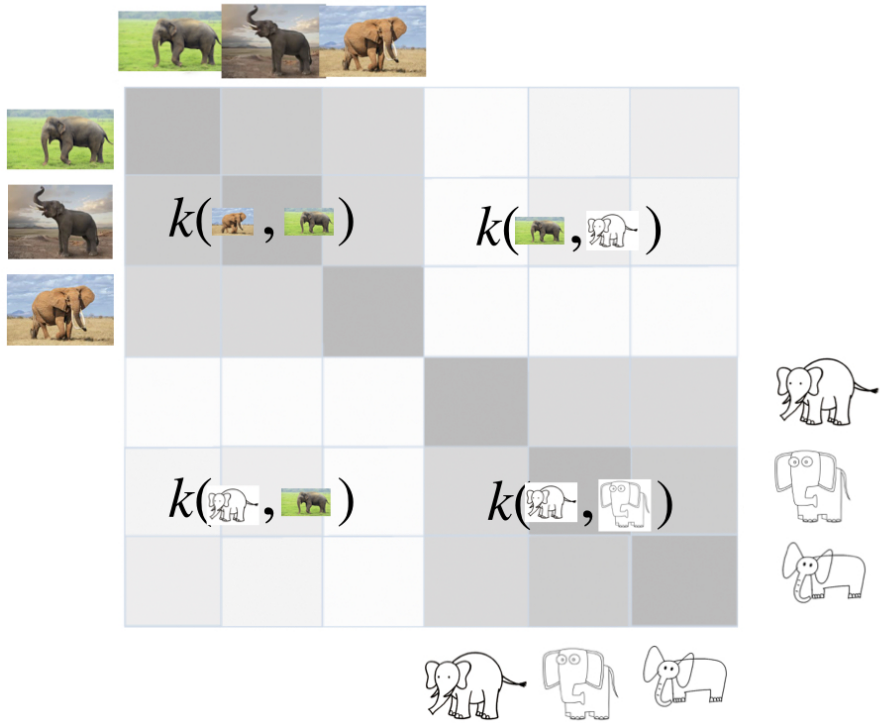


$Z^n \sim \nu_0$

[2] [Gretton et al., 2012]

The Maximum Mean Discrepancy[2]

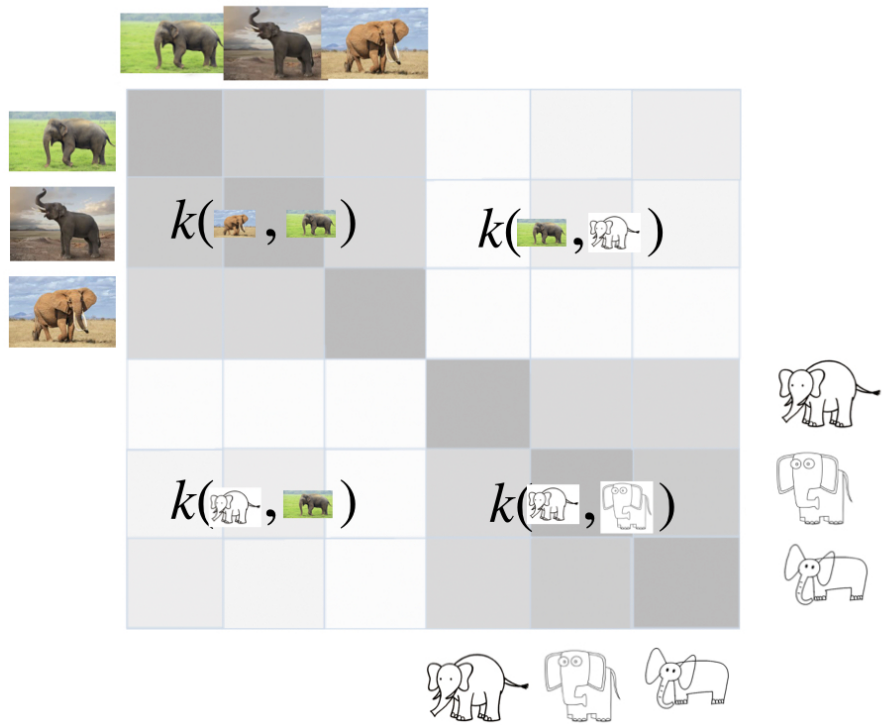
Compute a similarity matrix using a kernel k



[2] [Gretton et al., 2012]

The Maximum Mean Discrepancy[2]

Compute a similarity matrix using a kernel *k*

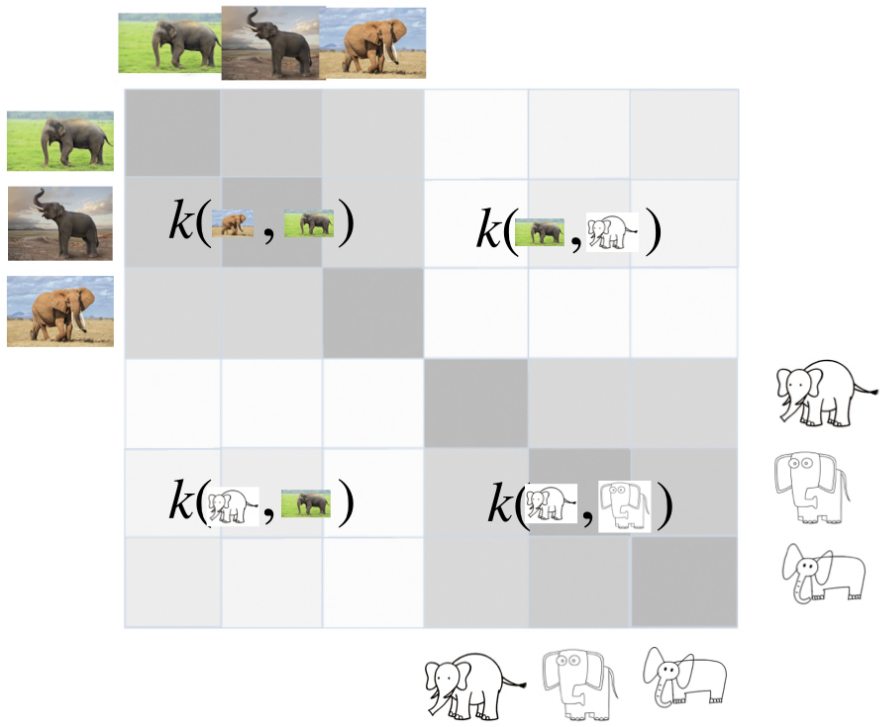


$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{n,n'} k(\,,\,) + \frac{1}{n(n-1)} \sum_{n,n'} k(\,,\,) - \frac{2}{n^2} \sum_{n,n'} k(\,,\,)$$

[2] [Gretton et al., 2012]

The Maximum Mean Discrepancy[2]

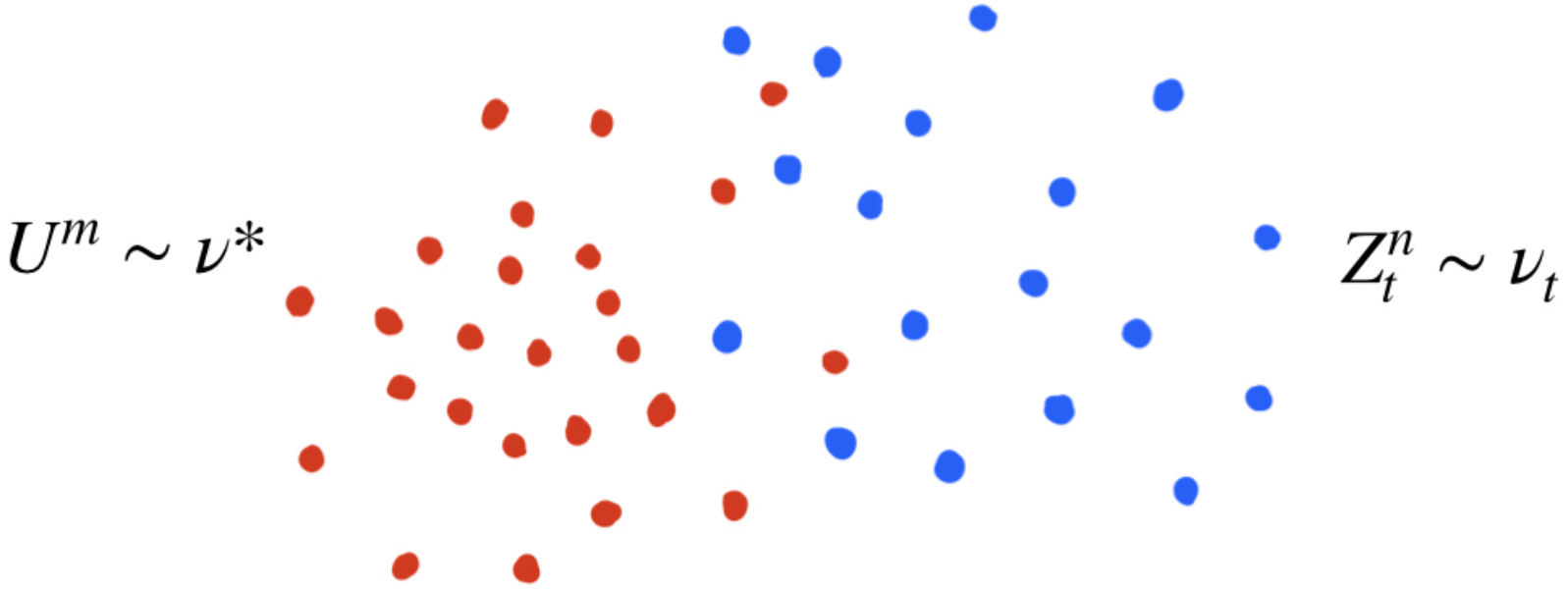
Compute a similarity matrix using a kernel *k*



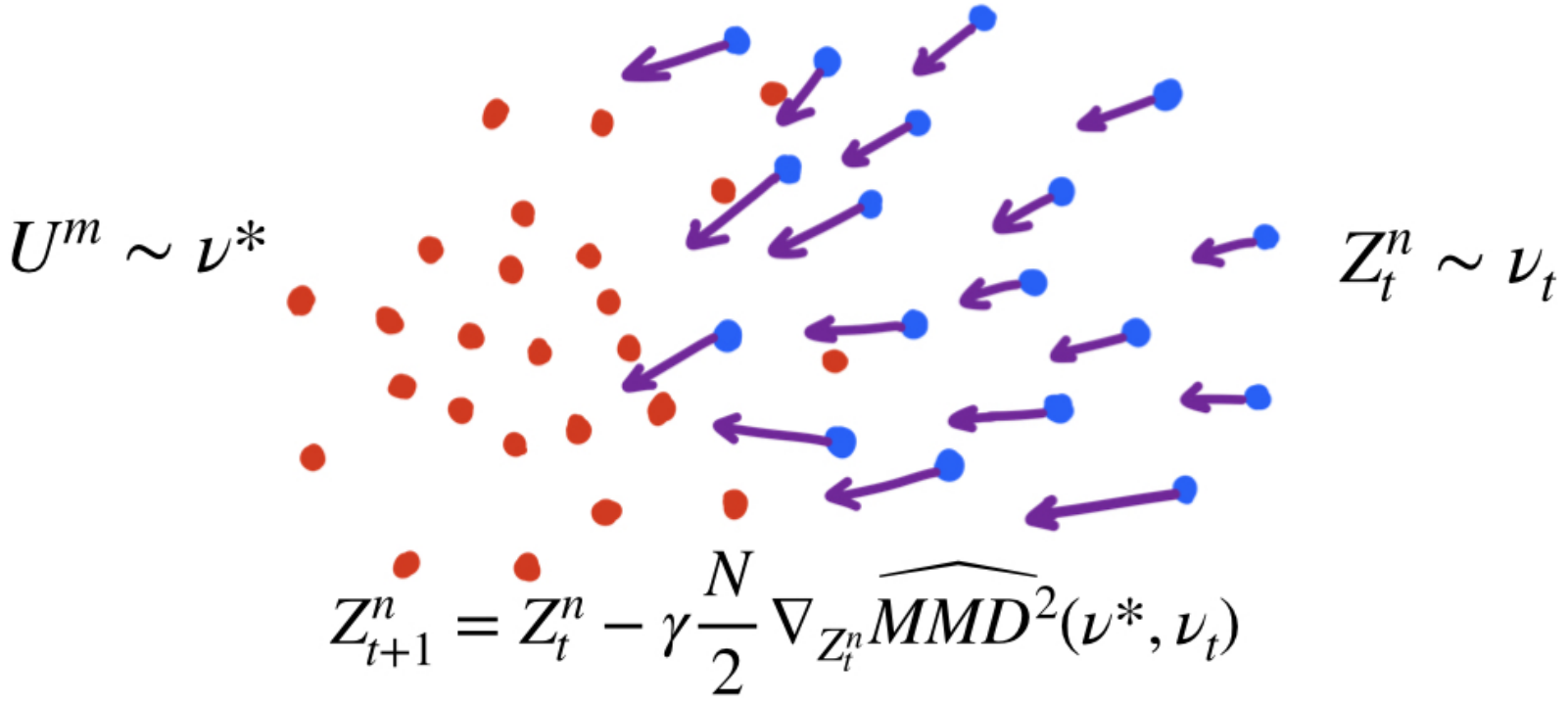
$$MMD^2(\nu^*, \nu_0) = \mathbb{E}_{\substack{U \sim \nu^* \\ U' \sim \nu^*}}[k(U, U')] + \mathbb{E}_{\substack{Z \sim \nu_0 \\ Z' \sim \nu_0}}[k(Z, Z')] - 2\mathbb{E}_{\substack{U \sim \nu^* \\ Z' \sim \nu_0}}[k(U, Z)]$$

[2] [Gretton et al., 2012]

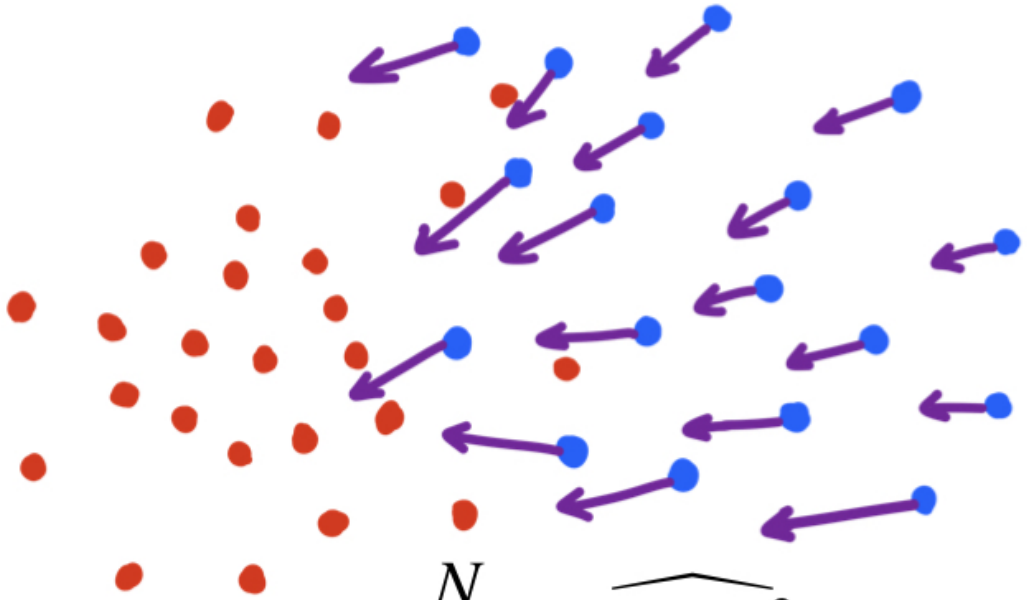
Gradient descent



Gradient descent



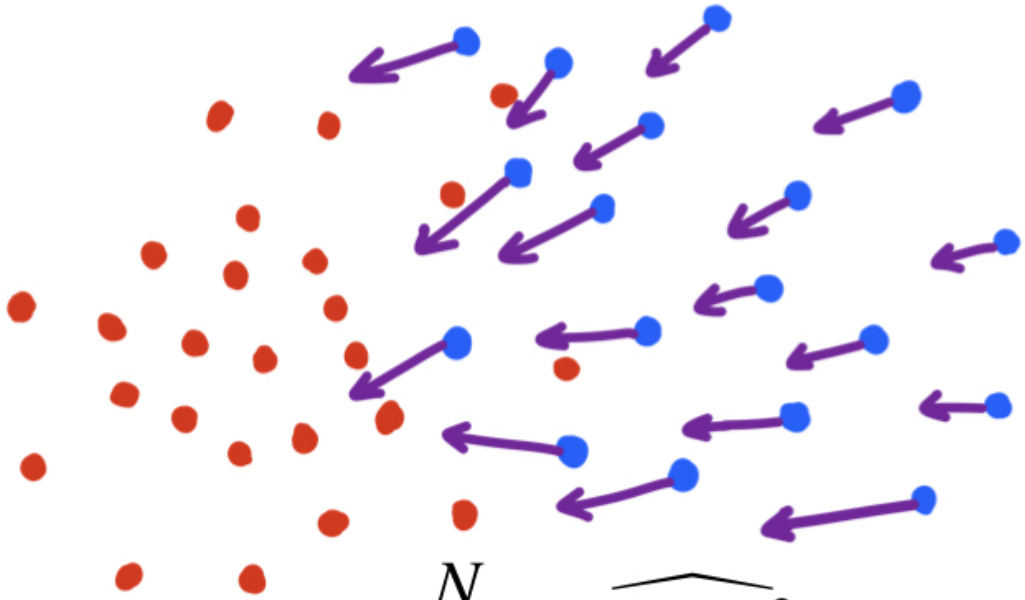
Gradient descent



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

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

Gradient descent



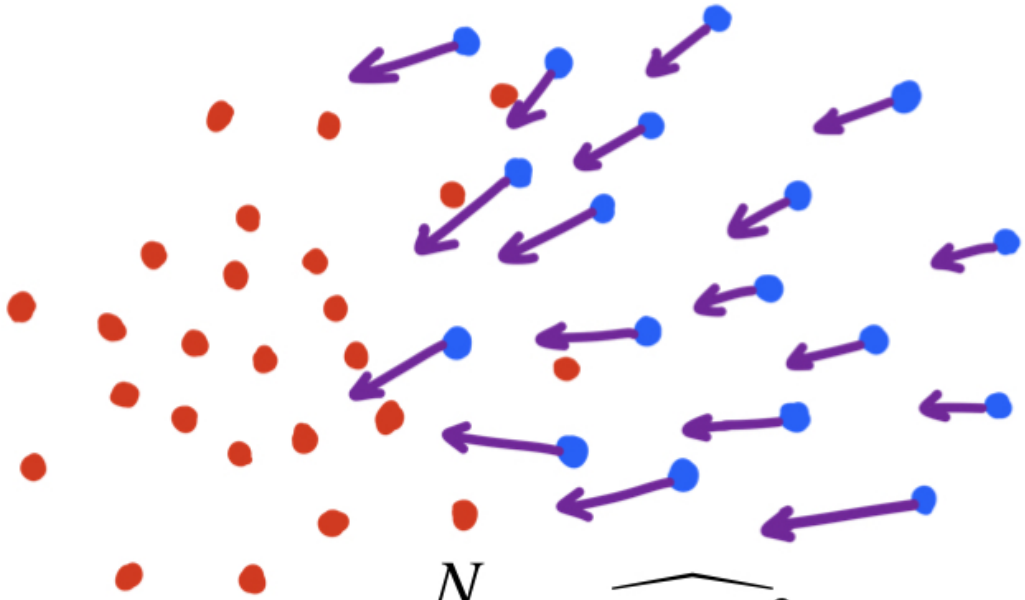
$$Z_{t+1}^n = Z_t^n - \gamma \frac{N}{2} \nabla_{Z_t^n} \widehat{MMD}^2(\nu^*, \nu_t)$$

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

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

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

Gradient descent



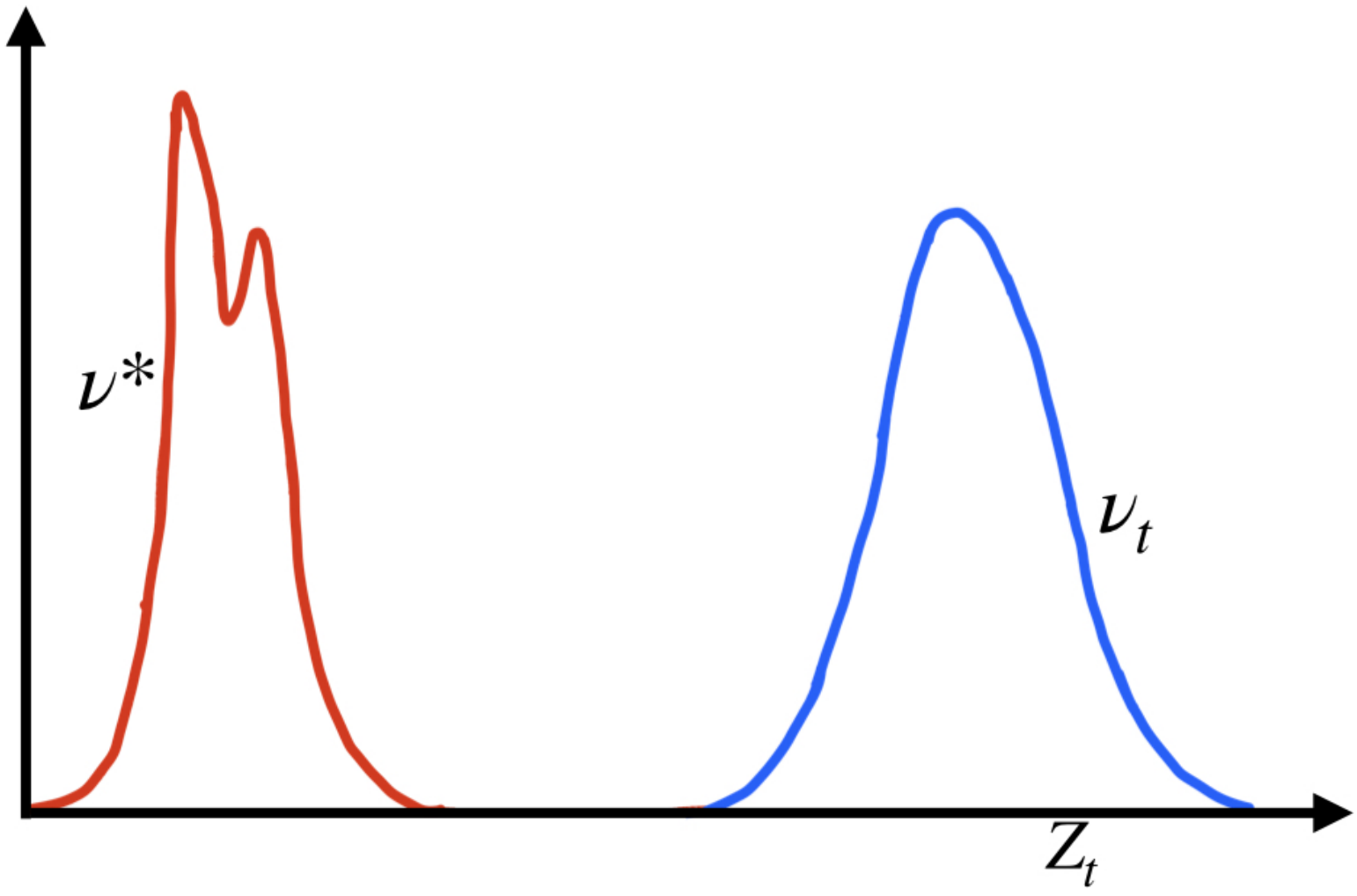
$$Z_{t+1}^n = Z_t^n - \gamma \frac{N}{2} \nabla_{Z_t^n} \widehat{MMD}^2(\nu^*, \nu_t)$$

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

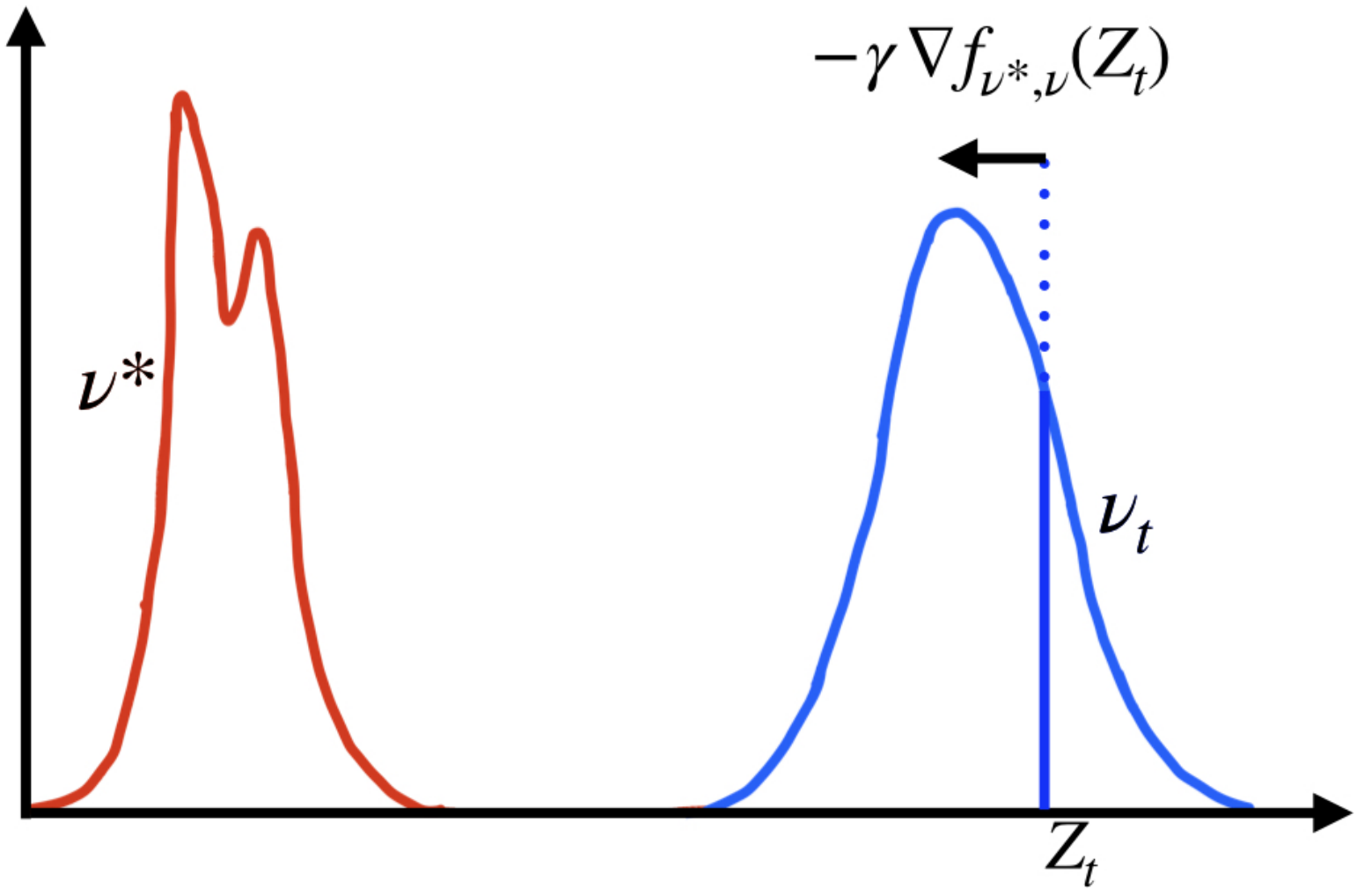
$$\downarrow N, M \to \infty$$

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

Wasserstein gradient descent

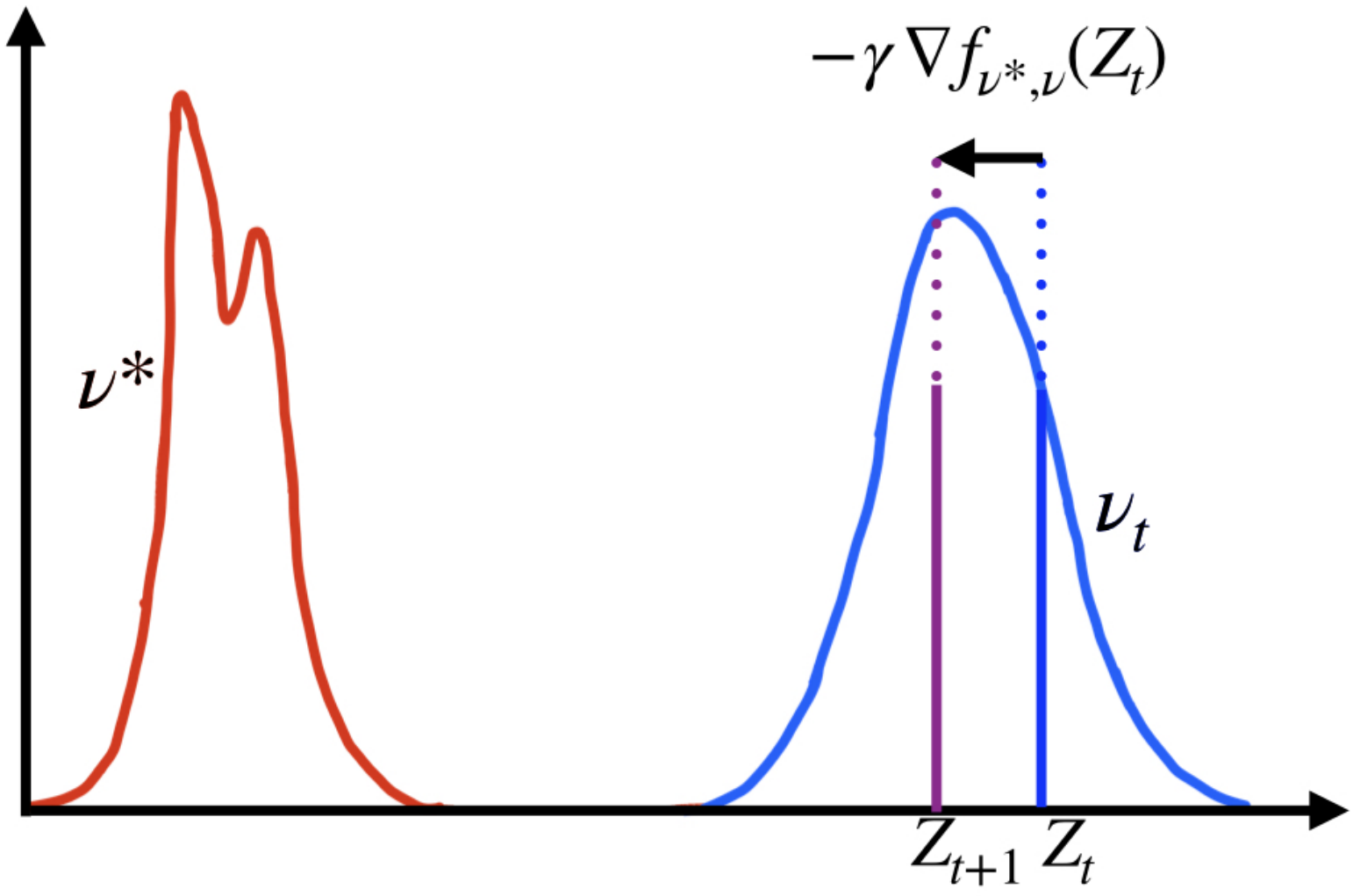


Wasserstein gradient descent



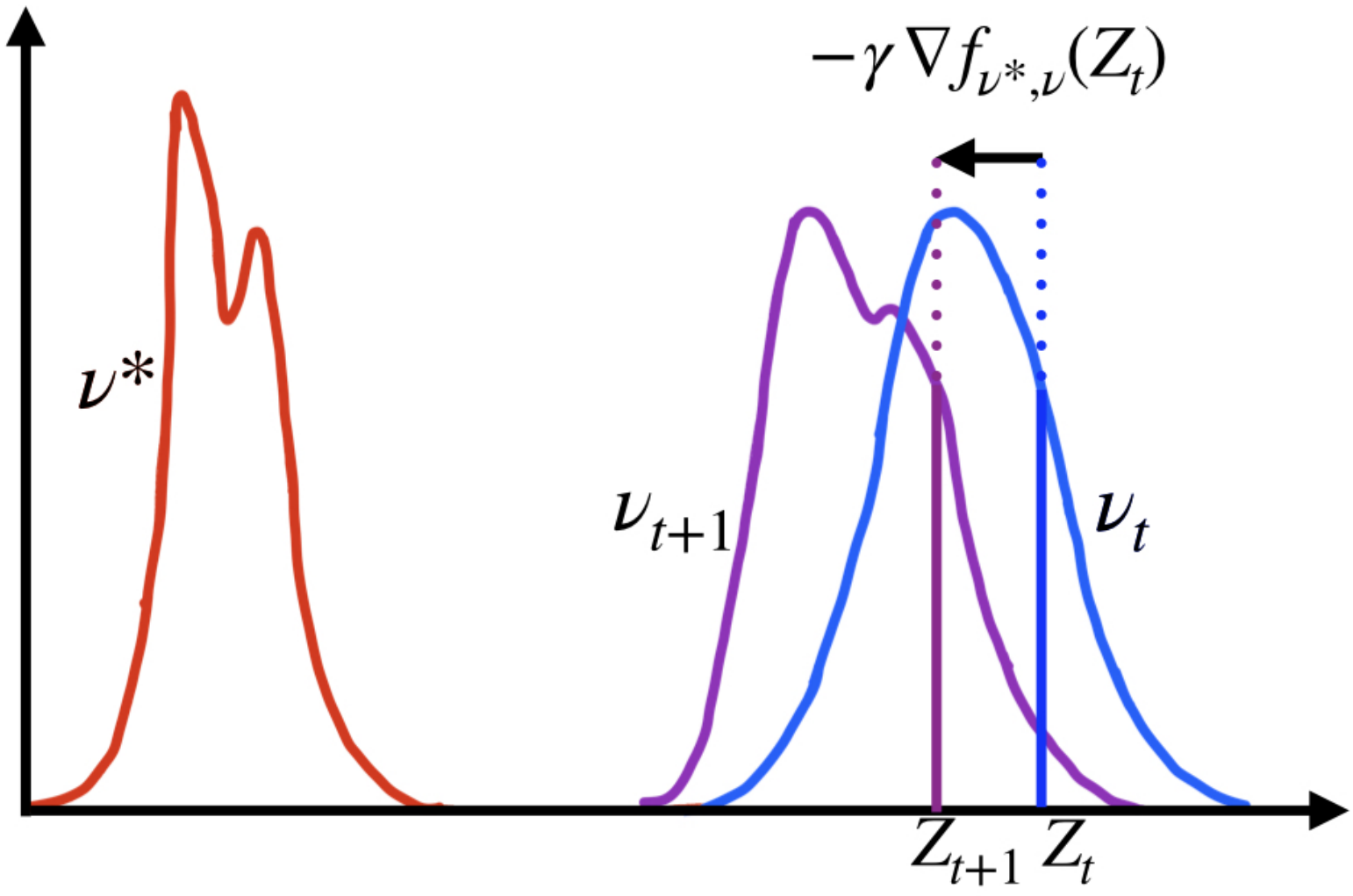
$$f_{\nu^*,\nu_t}(Z_t) = \mathbb{E}_{Z' \sim \nu_t}[k(Z_t, Z')] - \mathbb{E}_{U \sim \nu^*}[k(Z_t, U)]$$

Wasserstein gradient descent



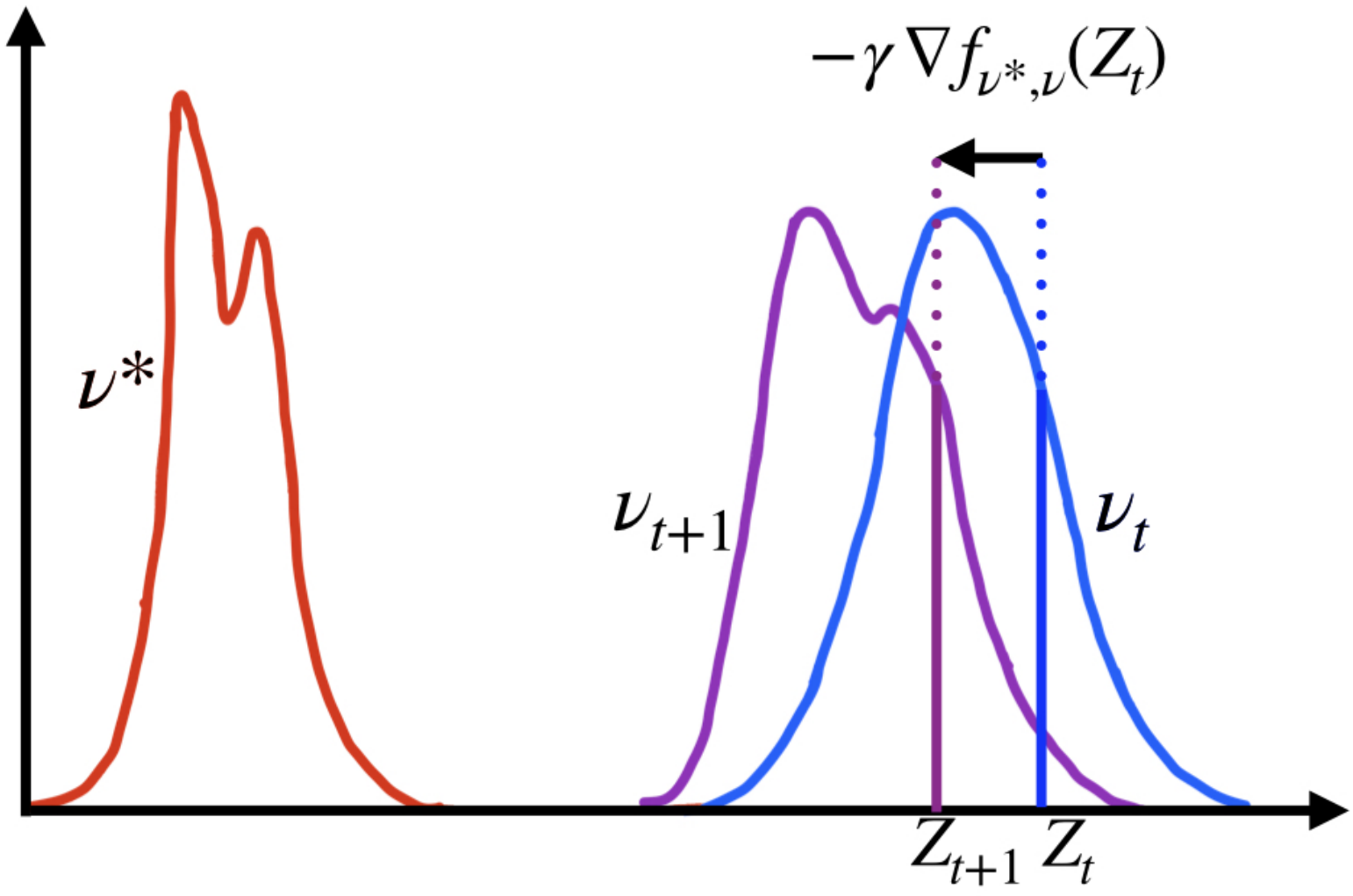
$$f_{\nu^*,\nu_t}(Z_t) = \mathbb{E}_{Z' \sim \nu_t}[k(Z_t, Z')] - \mathbb{E}_{U \sim \nu^*}[k(Z_t, U)]$$

Wasserstein gradient descent



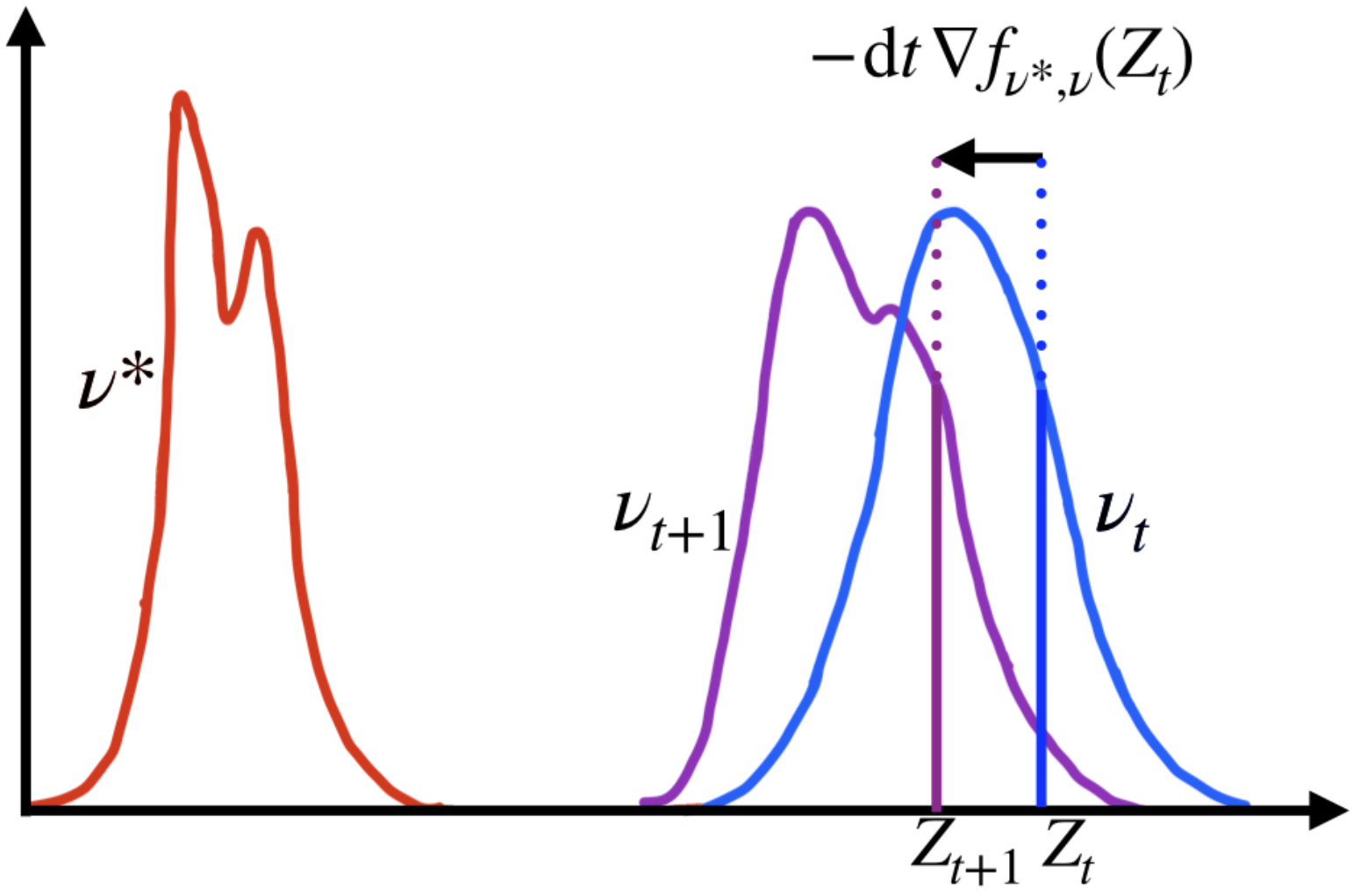
$$f_{\nu^*,\nu_t}(Z_t) = \mathbb{E}_{Z' \sim \nu_t}[k(Z_t, Z')] - \mathbb{E}_{U \sim \nu^*}[k(Z_t, U)]$$

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

Continuity Equation

- Continuous time equation: Mc-Kean Vlasov dynamics

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

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

Continuity Equation

- 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 ν_t:

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

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

Continuity Equation

- 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 ν_t:

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

- Interpretation as a gradient flow in probability space [3]:

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

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

A criterion for convergence

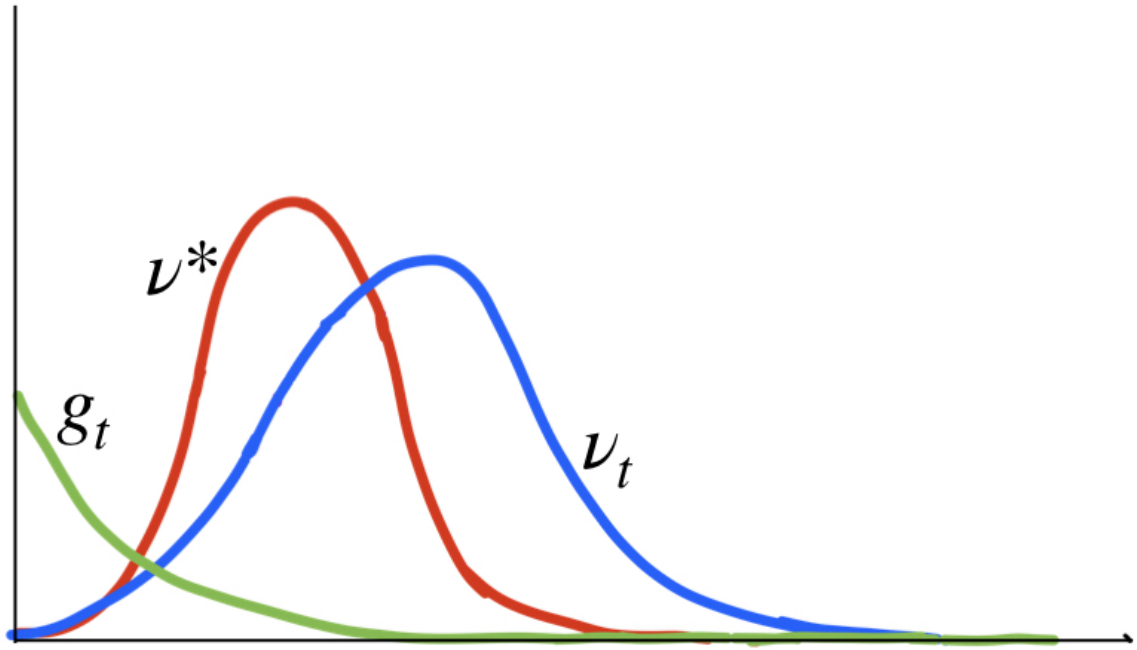
- Define the Negative Sobolev distance:

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

A criterion for convergence

- Define the Negative Sobolev distance:

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



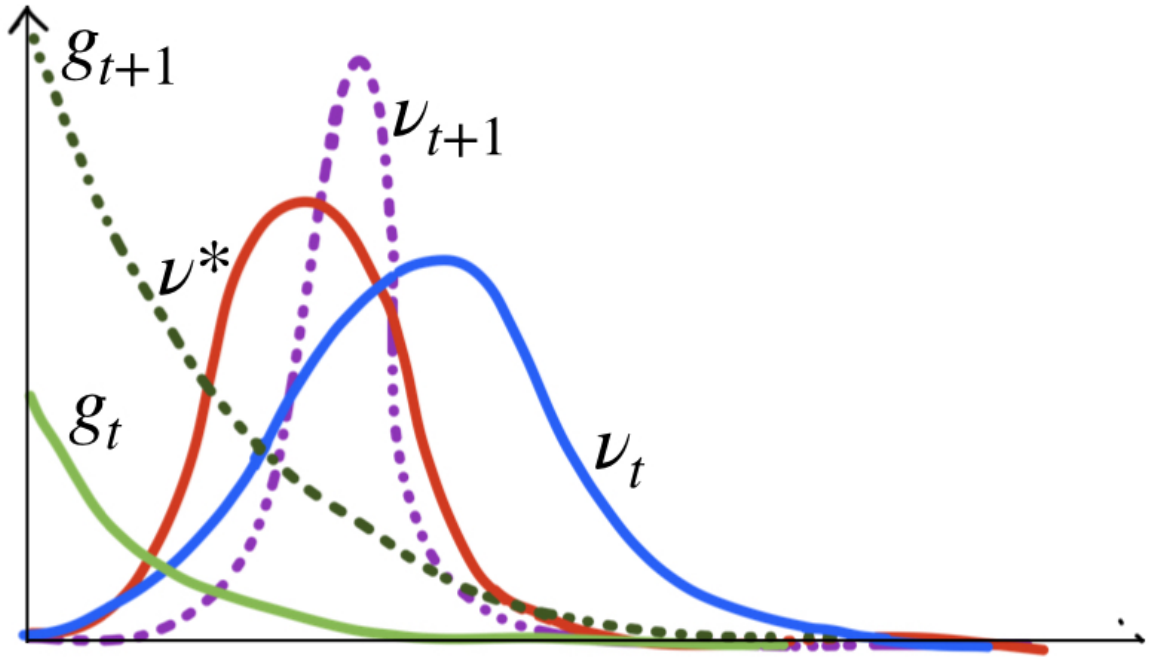
$$\int \|\nabla g_t(x)\|^2 \, d\nu_t(x) \leq 1$$

$$S(\nu^* \,|\, \nu_t) = |\, \mathbb{E}_{Z\sim\nu_t}[g_t(Z)] - \mathbb{E}_{U\sim\nu^*}[g_t(U)] \,|$$

A criterion for convergence

- Define the Negative Sobolev distance:

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

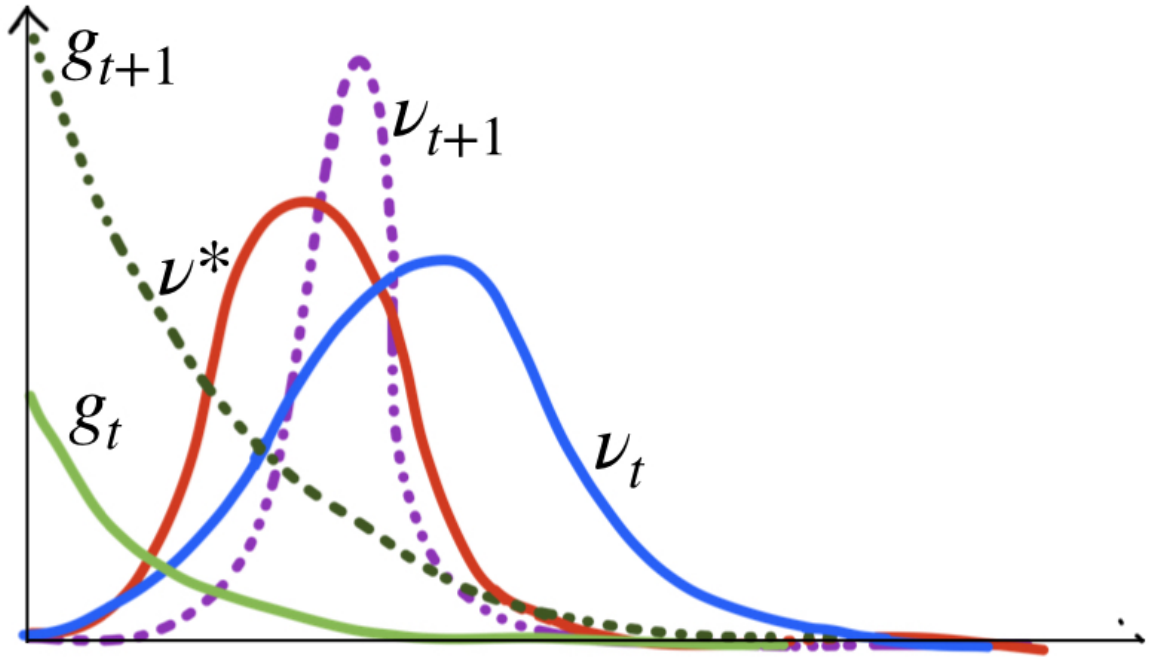


$$S(\nu^* \,|\, \nu_{t+1}) - S(\nu^* \,|\, \nu_t) > 0$$

A criterion for convergence

- Define the Negative Sobolev distance:

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



$$S(\nu^* \,|\, \nu_{t+1}) - S(\nu^* \,|\, \nu_t) > C$$

A criterion for convergence

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

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

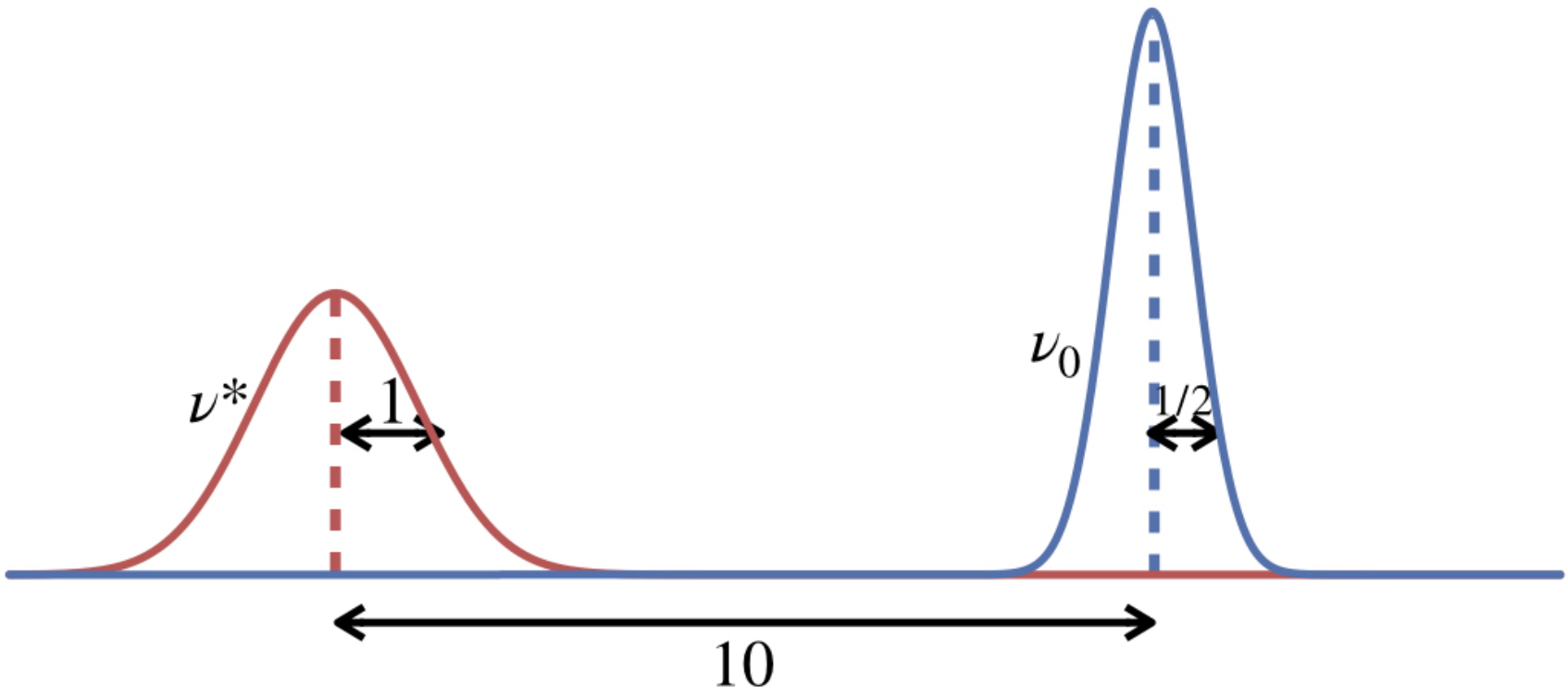
A criterion for convergence

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

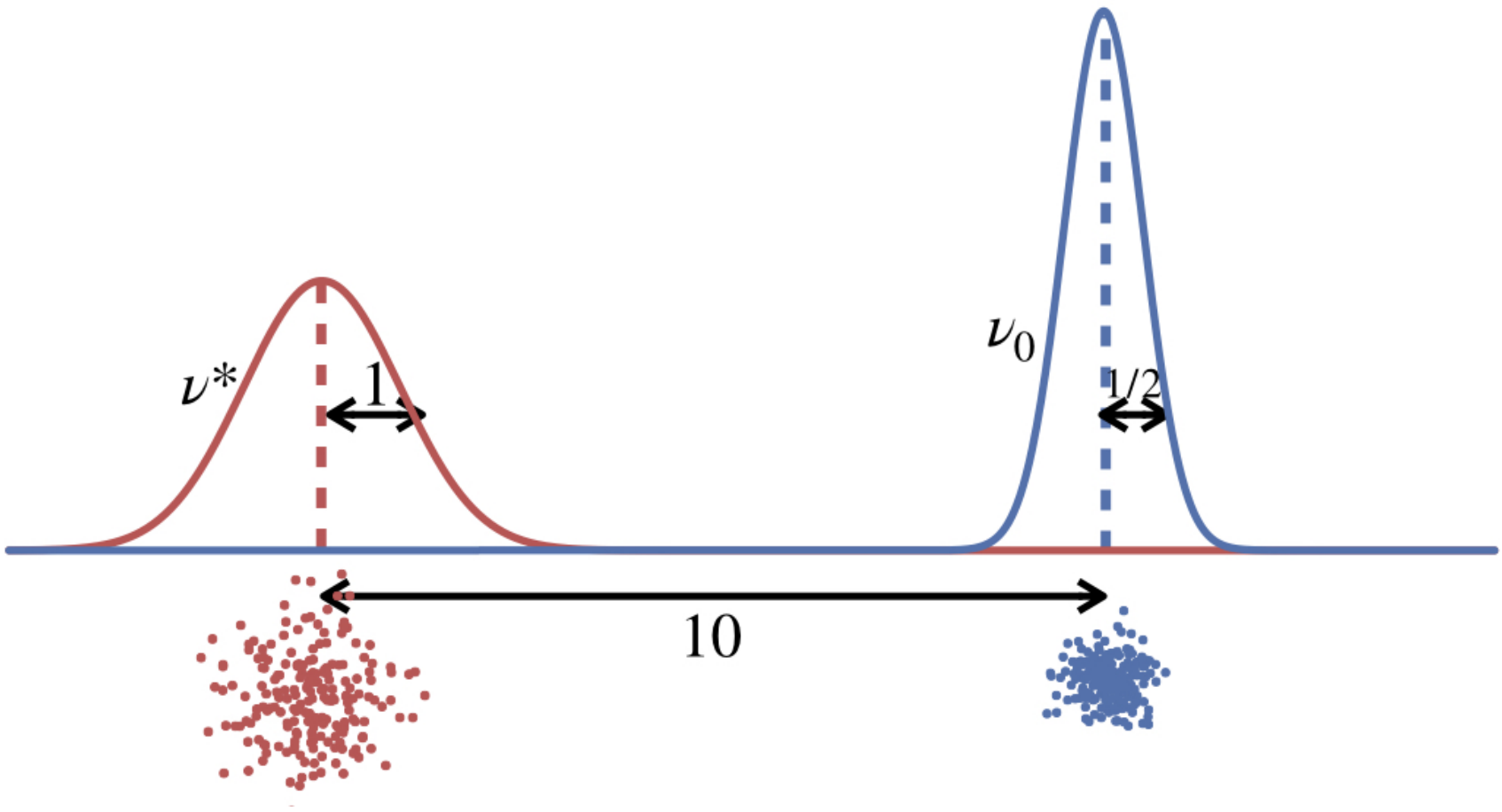
$$MMD^2(\nu^*, \nu_t) \leq \frac{1}{MMD^2(\nu^*, \nu_0) + 8\gamma C^{-1} t}$$

- Depends on the whole sequence ν_t: Hard to verify in general
- Can be checked for simple examples.

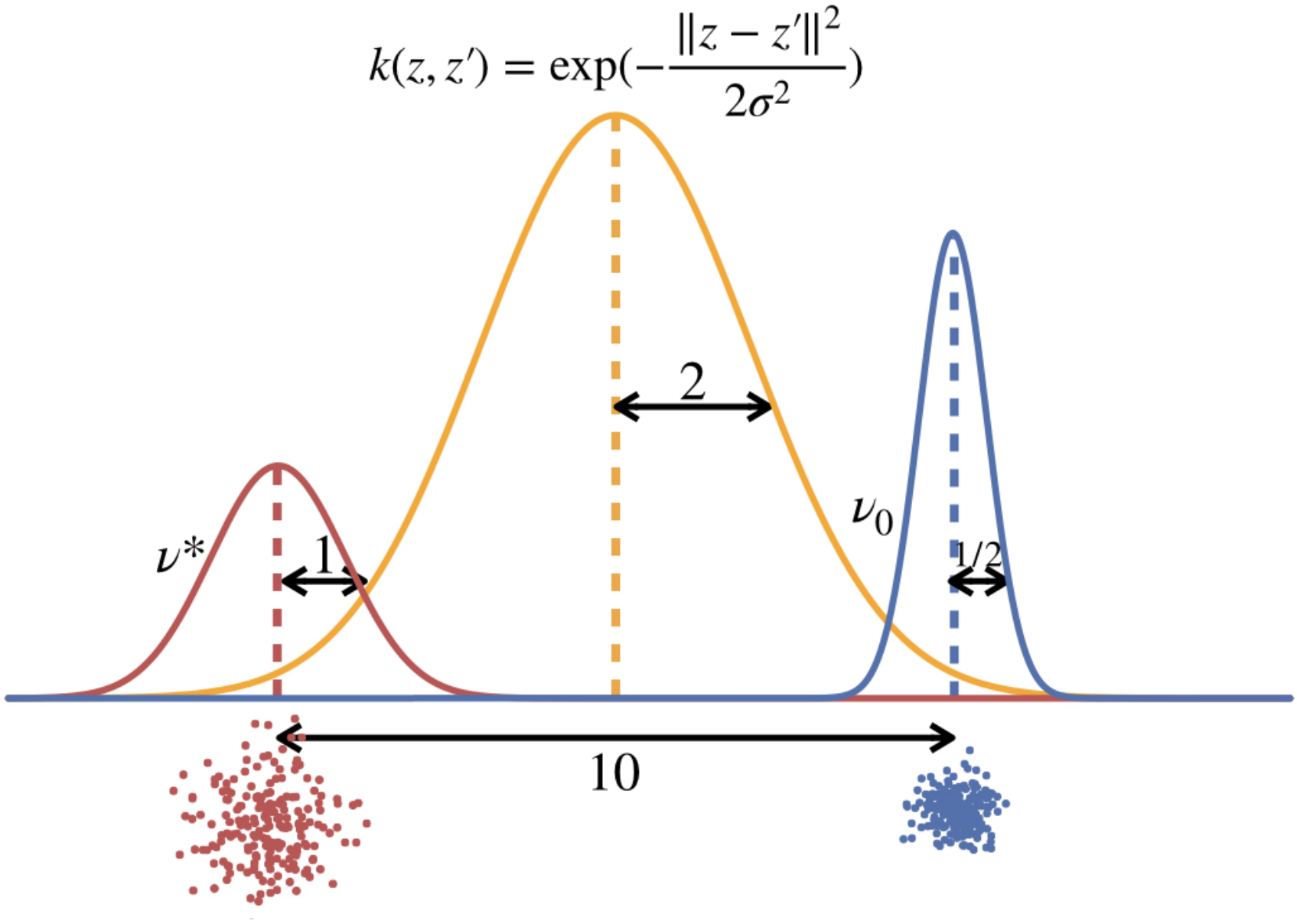
Convergence: Failure case



Convergence: Failure case



Convergence: Failure case



Convergence: Failure case

Noise Injection

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

[4] [Chaudhari et al., 2017, Hazan et al., 2015]

[5] [Mei et al., 2018]

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:

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

[4] [Chaudhari et al., 2017, Hazan et al., 2015]

[5] [Mei et al., 2018]

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* [4], but extended to interacting particles.

[4] [Chaudhari et al., 2017, Hazan et al., 2015]

[5] [Mei et al., 2018]

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:

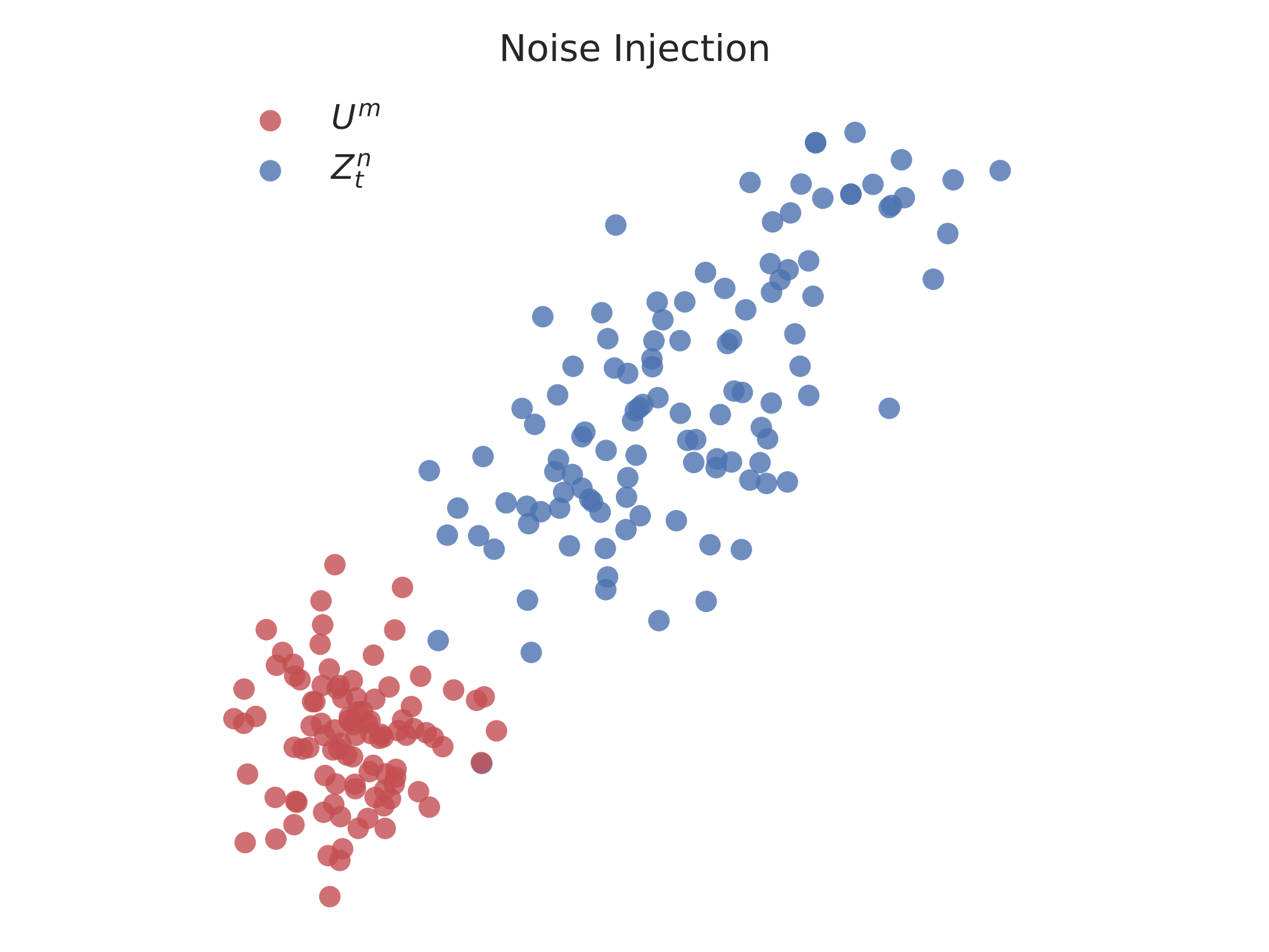
$$Z_{t+1} = Z_t - \gamma \nabla f_t(Z_t + \beta_t u_t); \qquad Z_t \sim \nu_t$$

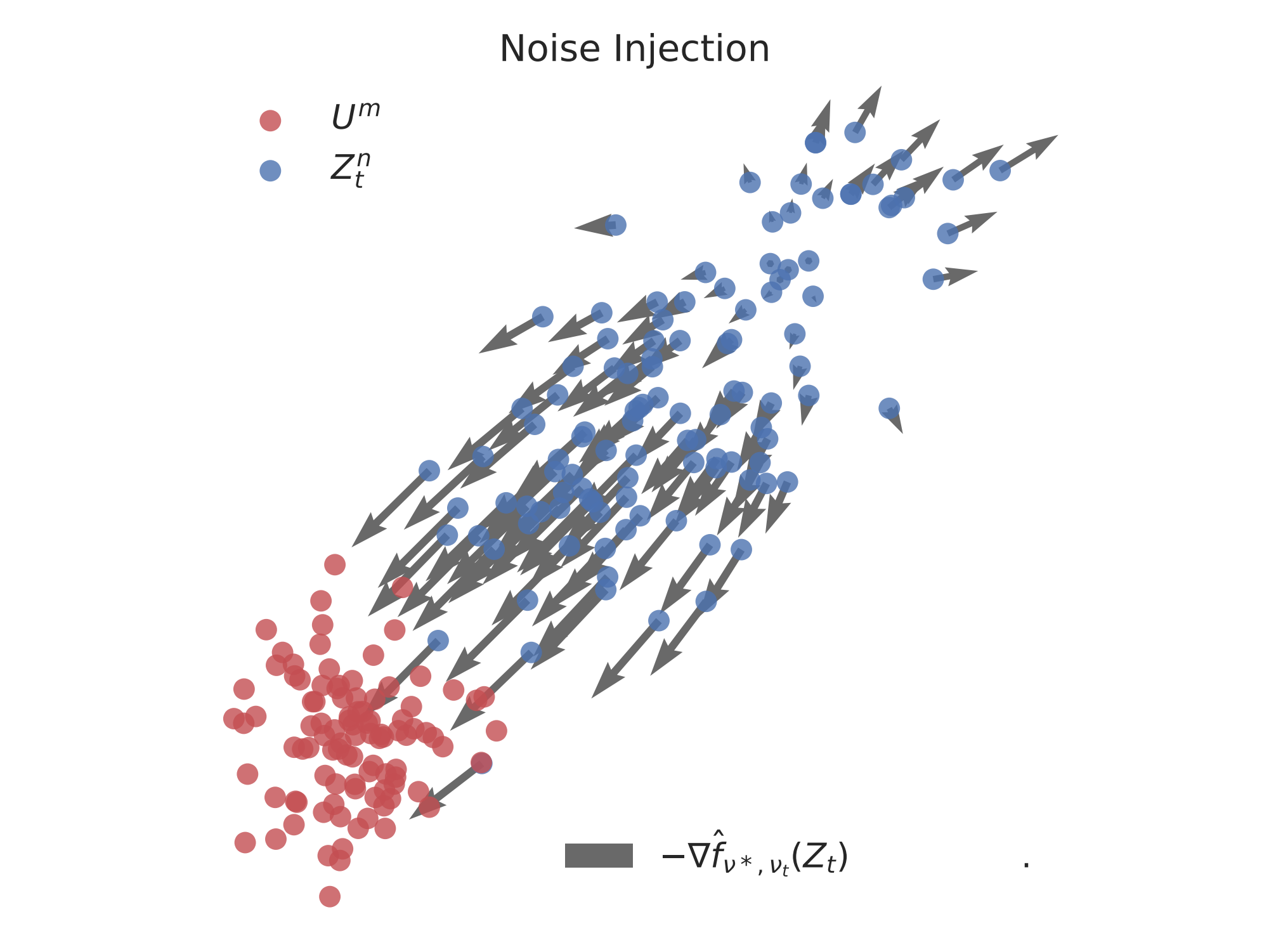
- Similar to *continuation methods* [4], but extended to interacting particles.
- Different from entropic regularization[5]

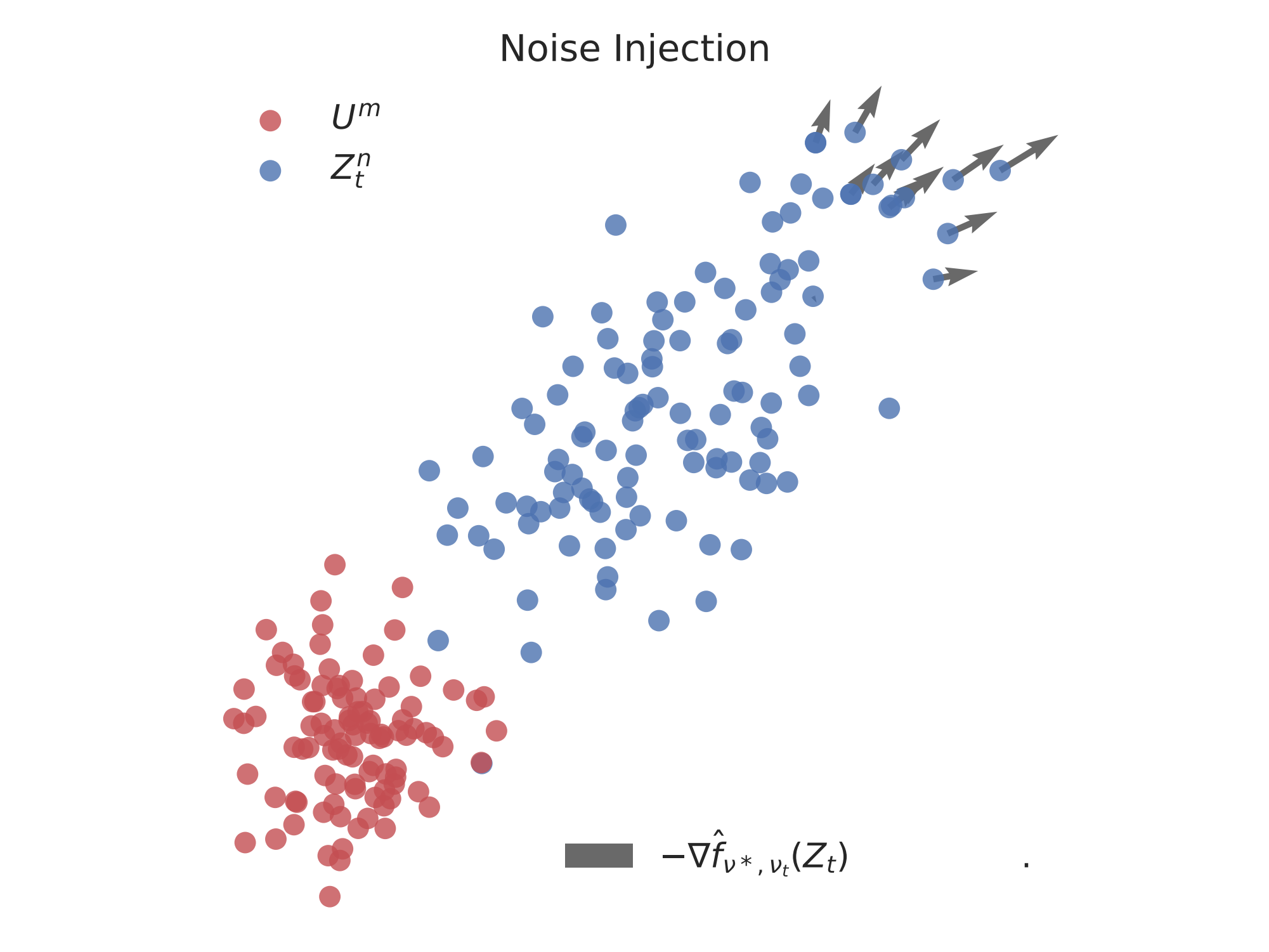
$$Z_{t+1} = Z_t - \gamma \nabla f_{\nu^*,\nu_t}(Z_t) + \beta_t u_t$$

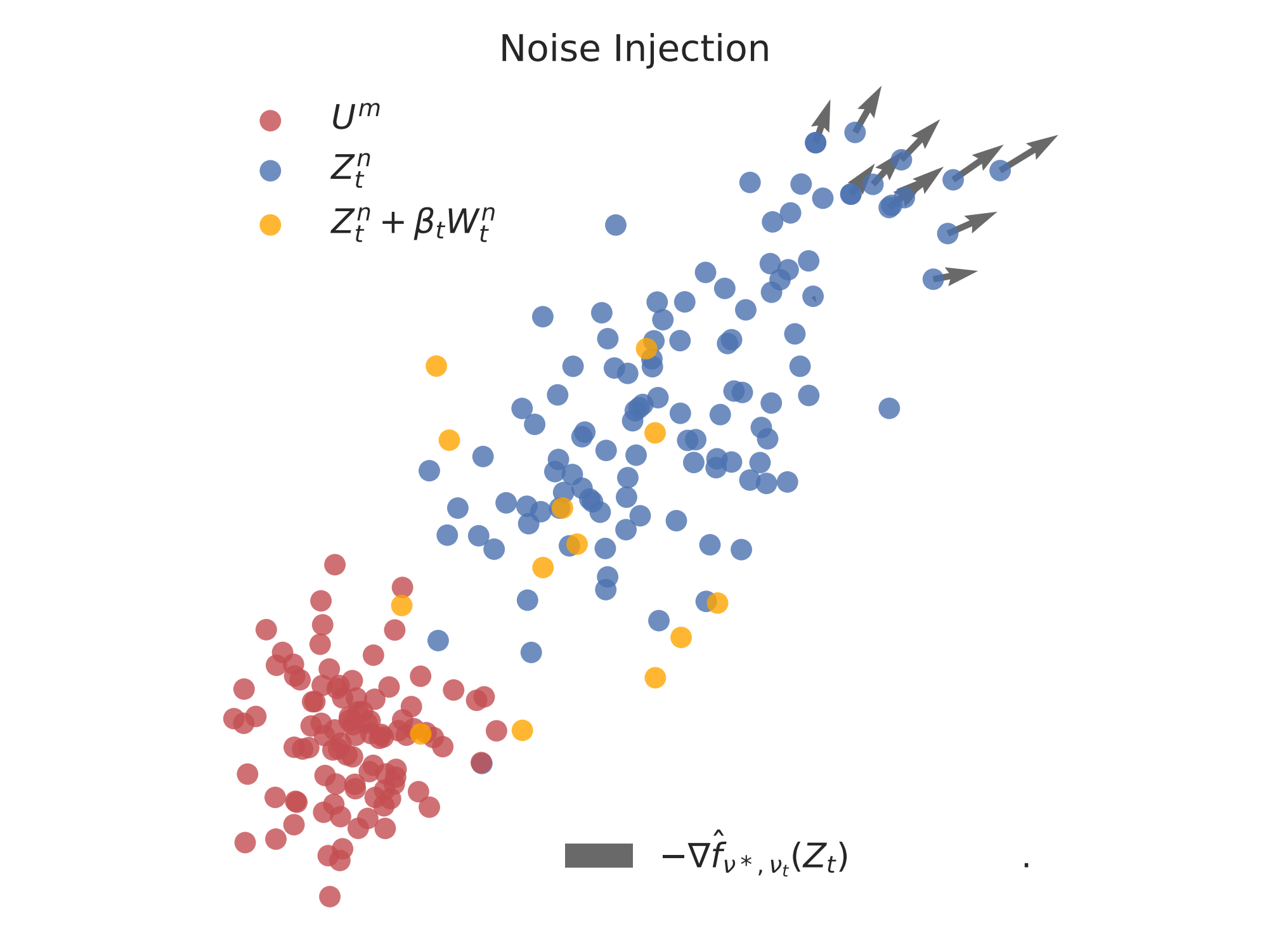
[4] [Chaudhari et al., 2017, Hazan et al., 2015]

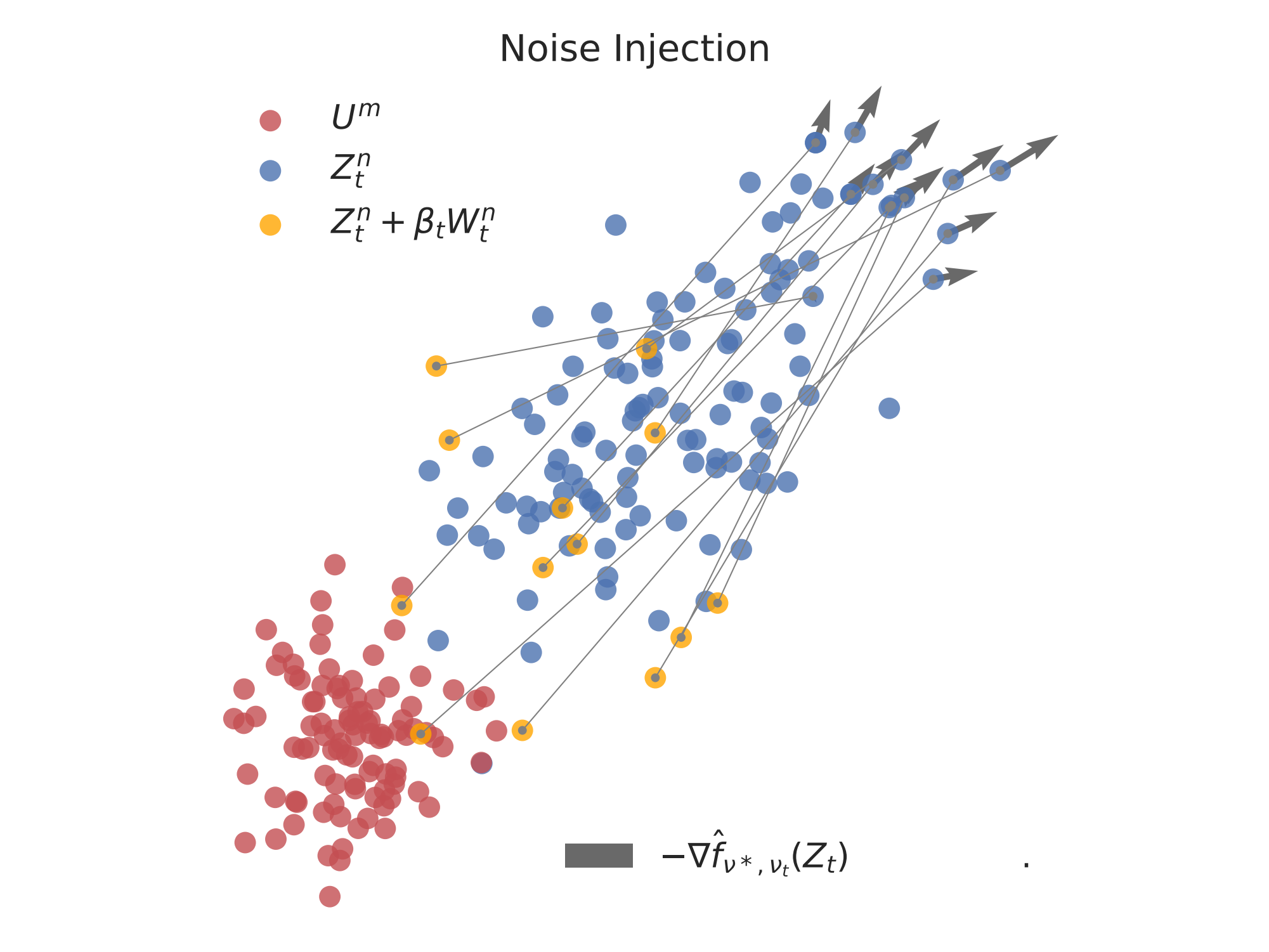
[5] [Mei et al., 2018]

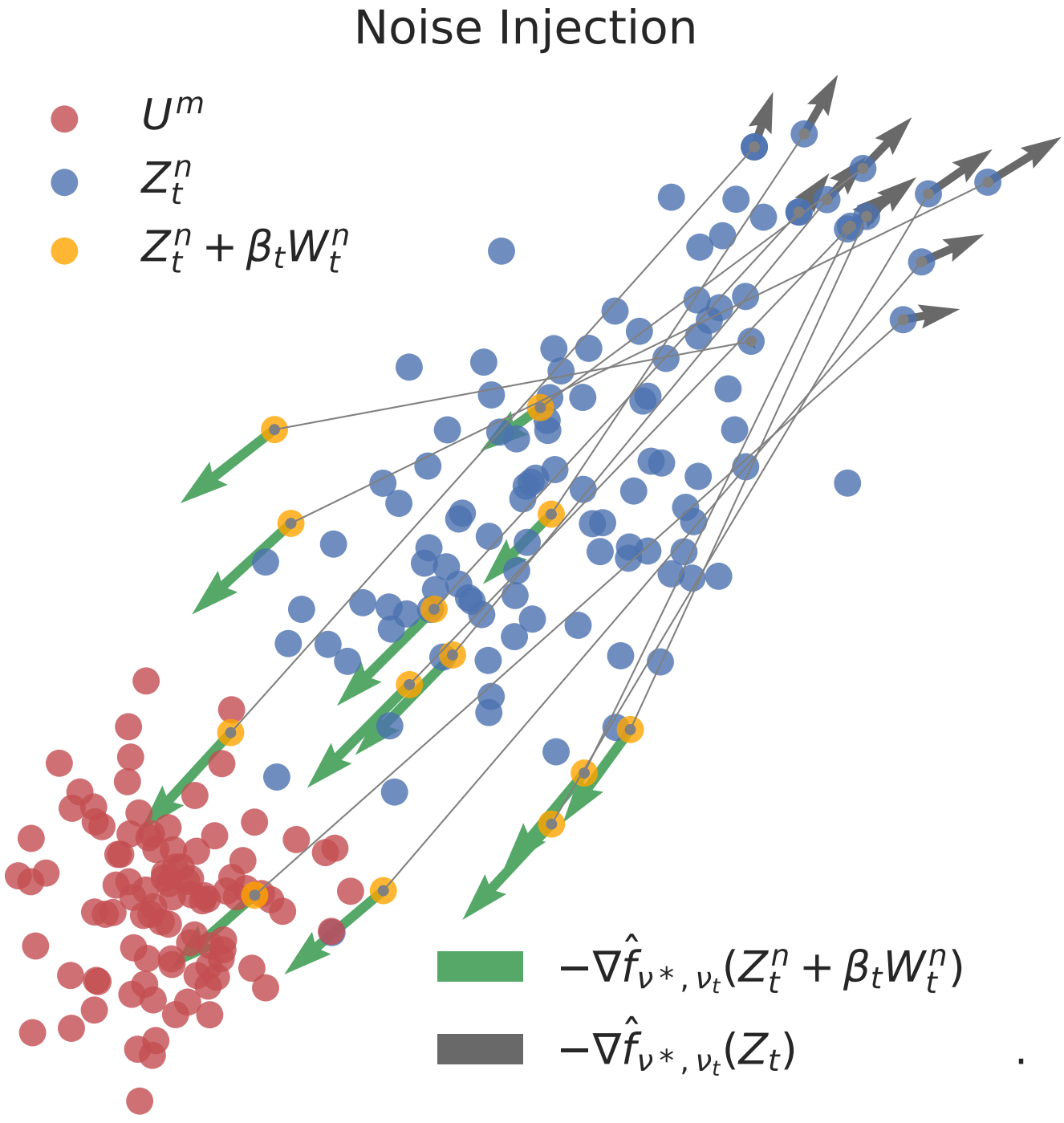

Noise Injection
U^m
Z_t^n

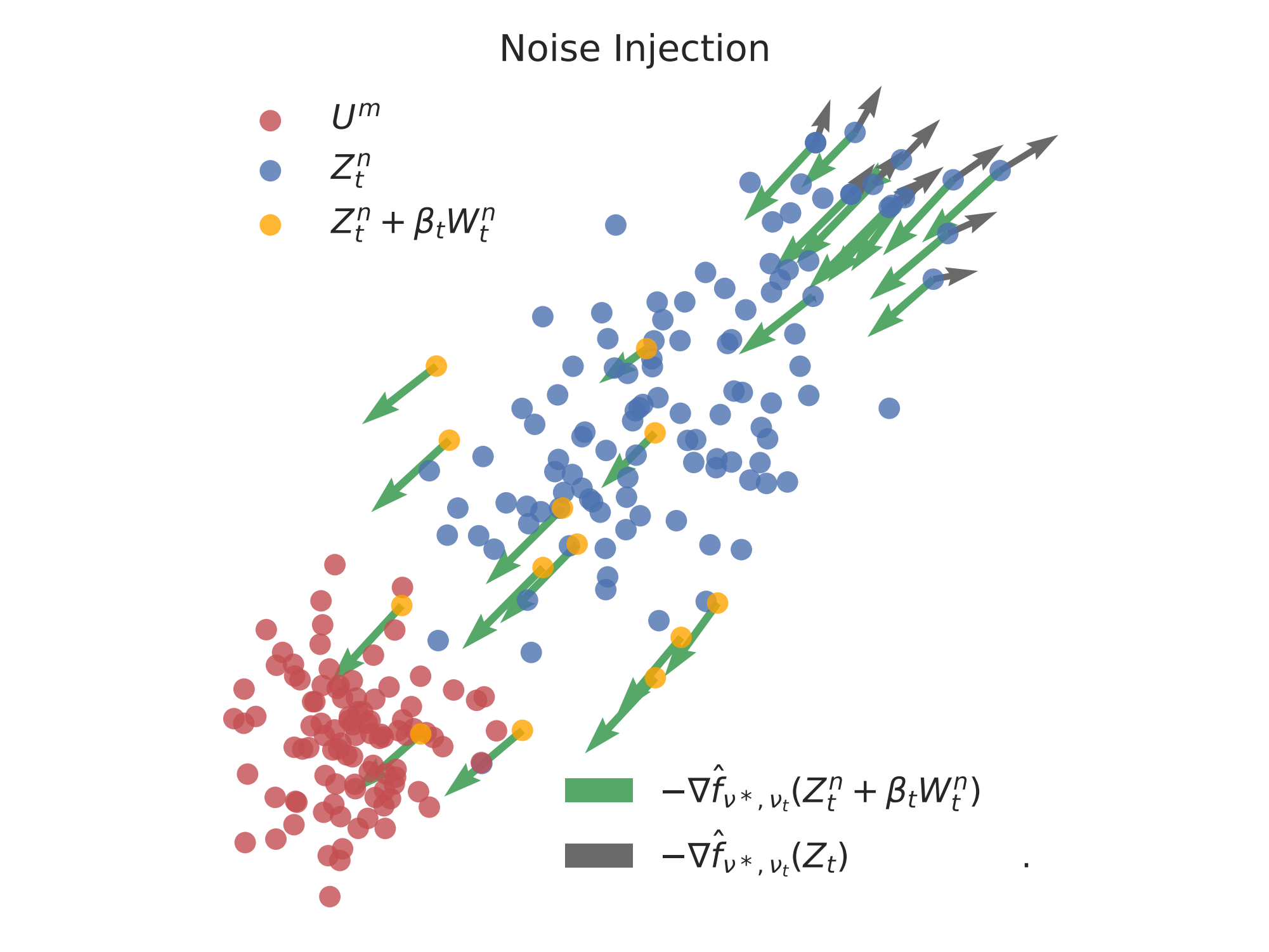

Noise Injection
U^m
Z_t^n
$-\nabla \hat{f}_{\nu^*, \nu_t}(Z_t)$

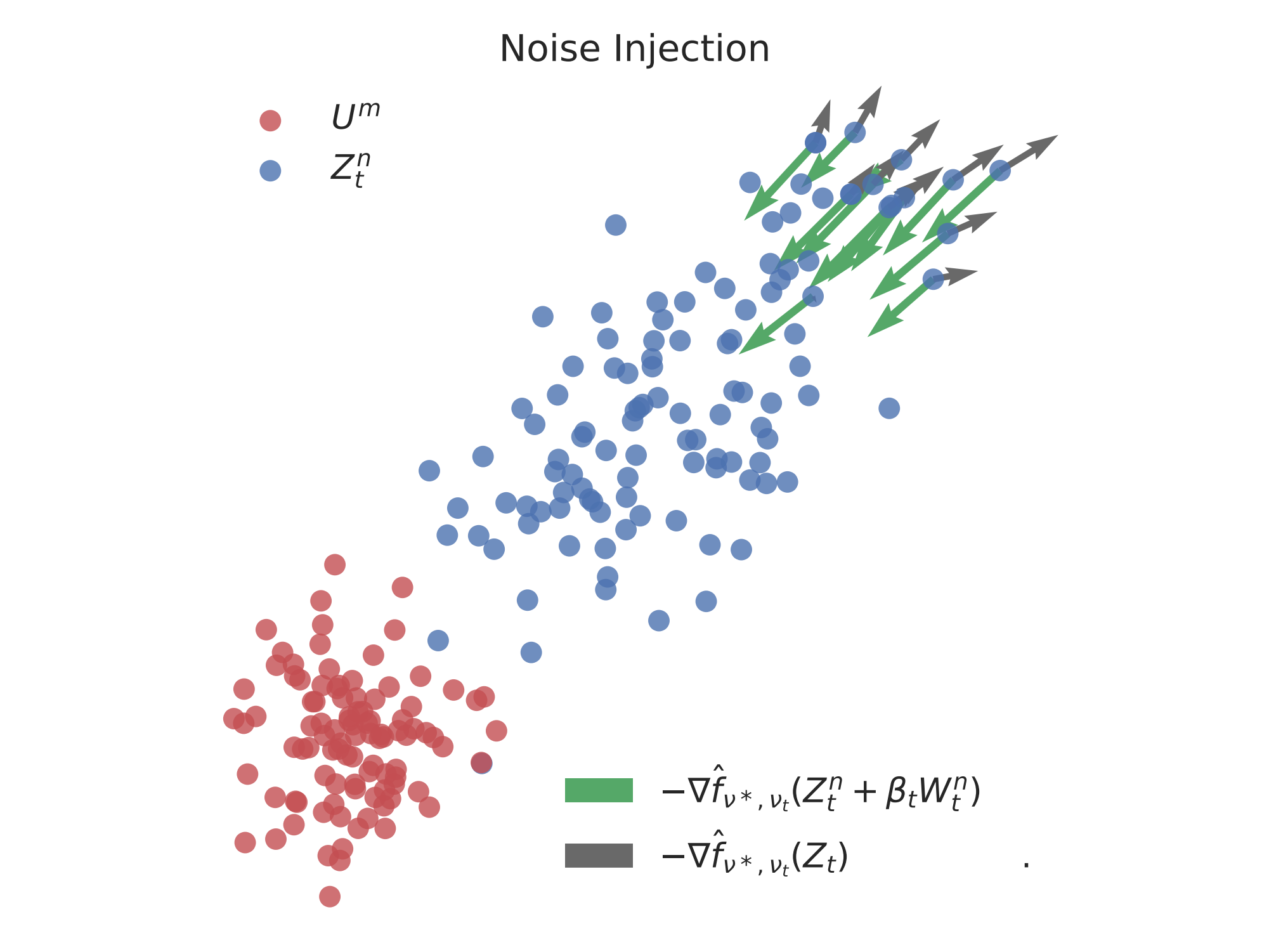

Noise Injection
U^m
Z_t^n
$-\nabla \hat{f}_{\nu^*, \nu_t}(Z_t)$

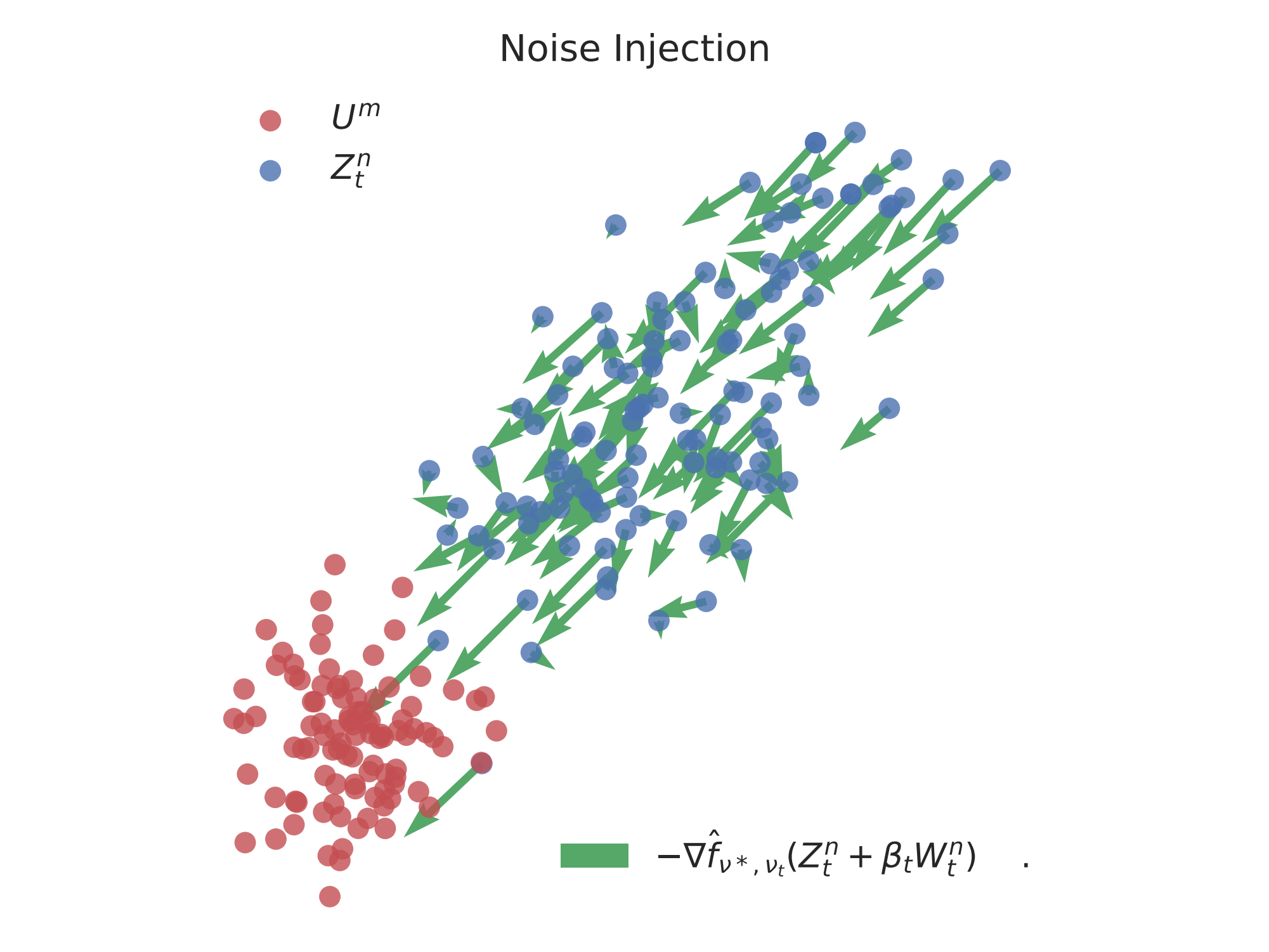

Noise Injection
U^m
Z_t^n
$Z_t^n + \beta_t W_t^n$
$-\nabla \hat{f}_{\nu^*, \nu_t}(Z_t)$


Noise Injection
U^m
Z_t^n
$Z_t^n + \beta_t W_t^n$
$-\nabla \hat{f}_{\nu^*, \nu_t}(Z_t)$


Noise Injection
U^m
Z_t^n
$Z_t^n + \beta_t W_t^n$
$-\nabla \hat{f}_{\nu^*, \nu_t}(Z_t^n + \beta_t W_t^n)$
$-\nabla \hat{f}_{\nu^*, \nu_t}(Z_t)$


Noise Injection
U^m
Z_t^n
$Z_t^n + \beta_t W_t^n$
$-\nabla \hat{f}_{\nu^*, \nu_t}(Z_t^n + \beta_t W_t^n)$
$-\nabla \hat{f}_{\nu^*, \nu_t}(Z_t)$


Noise Injection
U^m
Z_t^n
$-\nabla \hat{f}_{\nu^*,\nu_t}(Z_t^n + \beta_t W_t^n)$
$-\nabla \hat{f}_{\nu^*,\nu_t}(Z_t)$

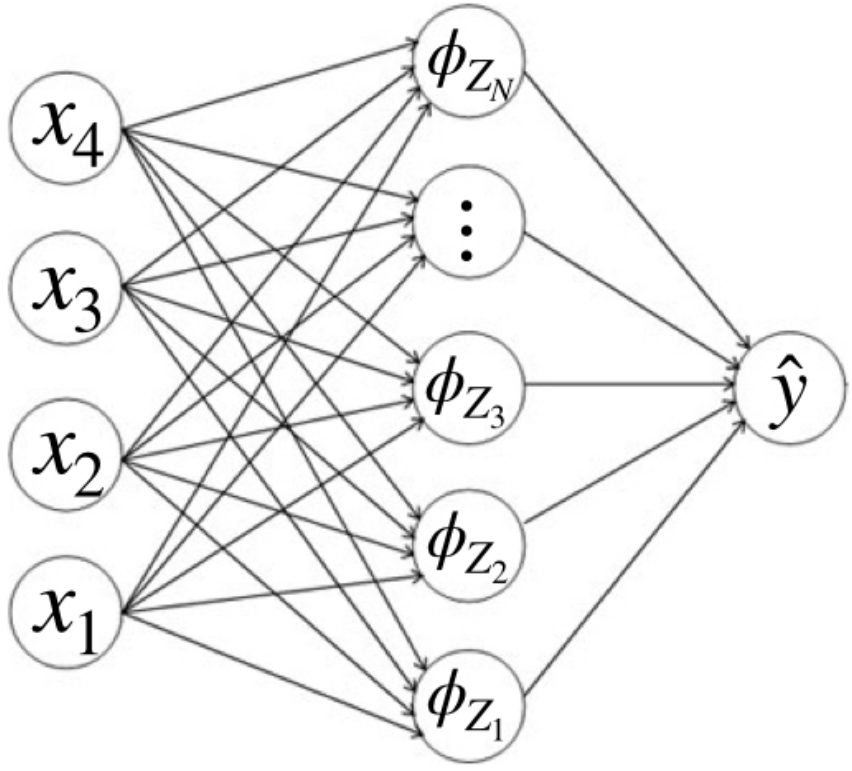

Noise Injection
U^m
Z_t^n
$-\nabla \hat{f}_{\nu^*,\nu_t}(Z_t^n + \beta_t W_t^n)$
.

Noise Injection: Experiments

Noise Injection: Experiments

Noise Injection: Experiments

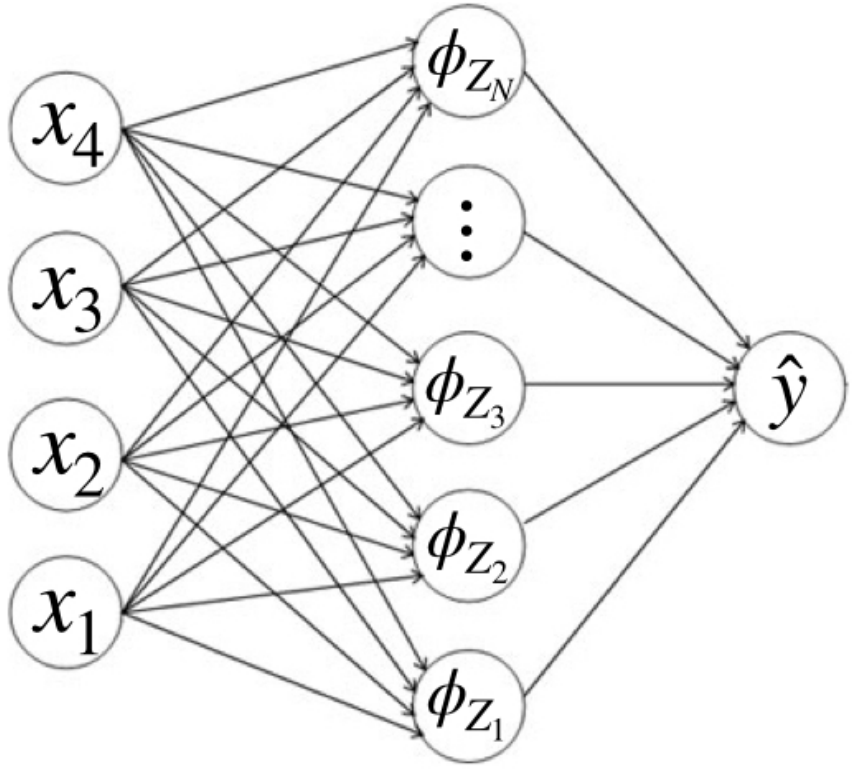
$$(x, y) \sim data$$



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

Noise Injection: Experiments

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

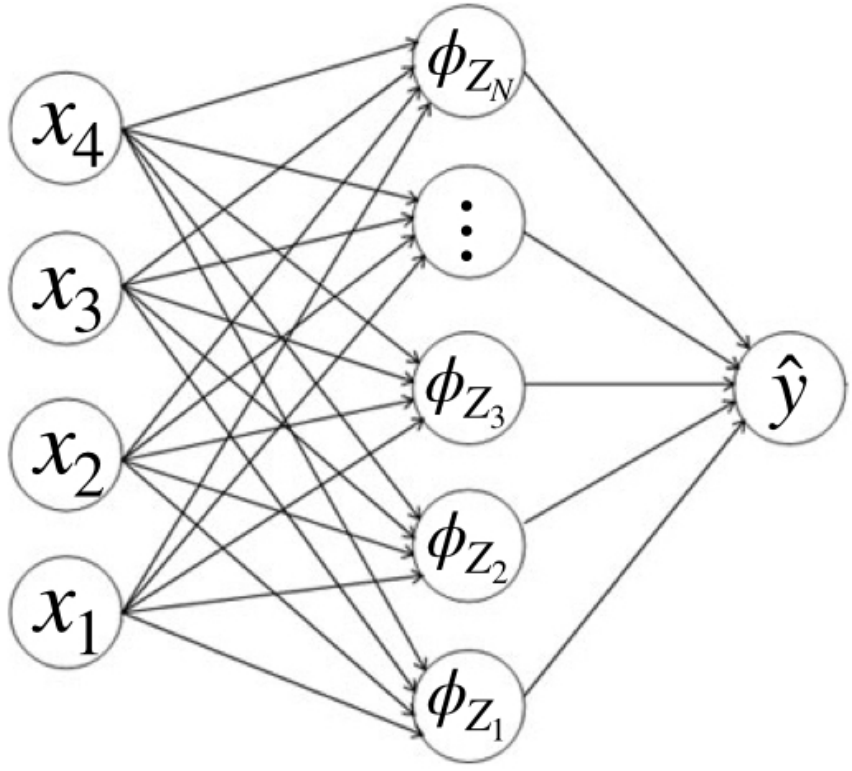


$$\min_{Z_1,\ldots,Z_N} MMD^2(\nu^*, \frac{1}{N}\sum_{n=1}^{N} \delta_{Z^n})$$

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

Noise Injection: Experiments

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



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

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

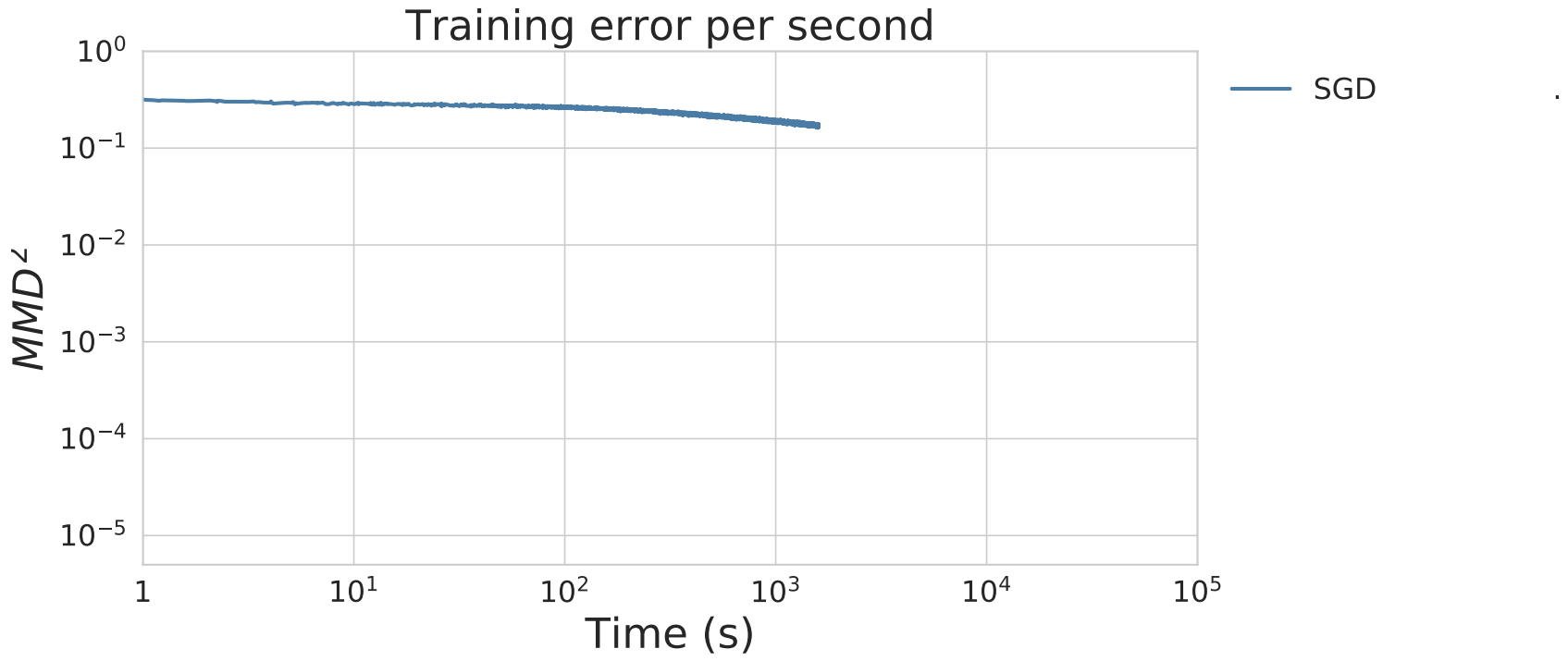
Noise Injection: Experiments

Methods:

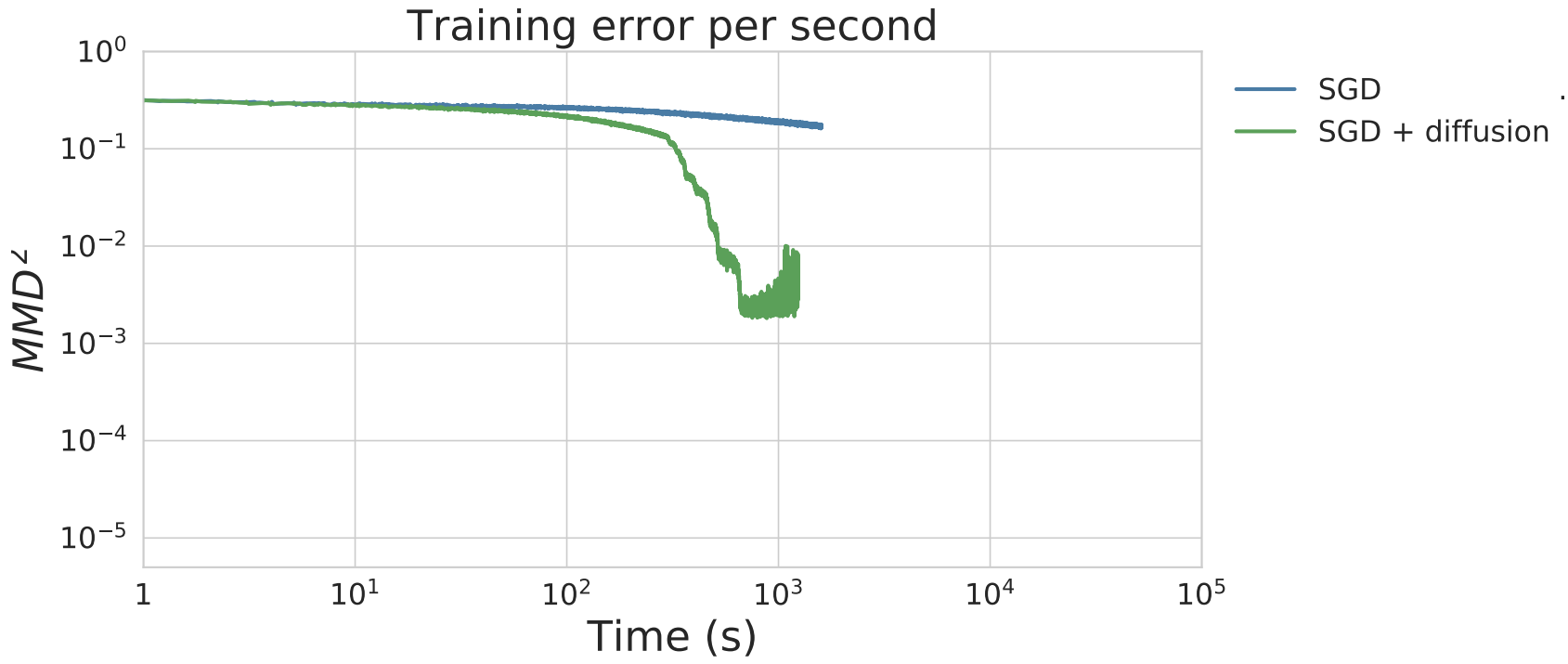
- SGD
- SGD + Noise injection
- SGD + diffusion
- KSD [6]: SGD using the Negative Sobolev distance $\nu \mapsto S(\nu^*|\nu)$ as a loss function: also minimizes the MMD.

[6] [Mroueh et al., 2019]

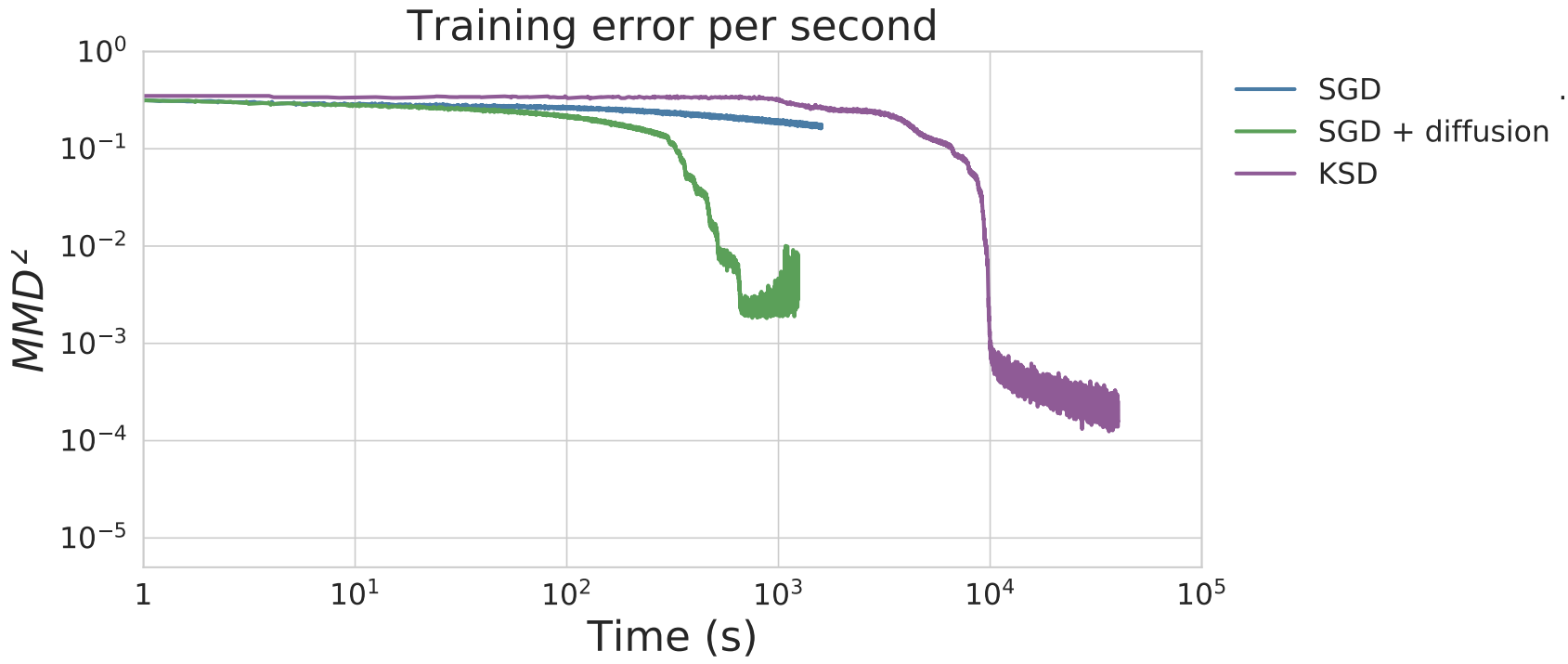
Noise Injection: Experiments



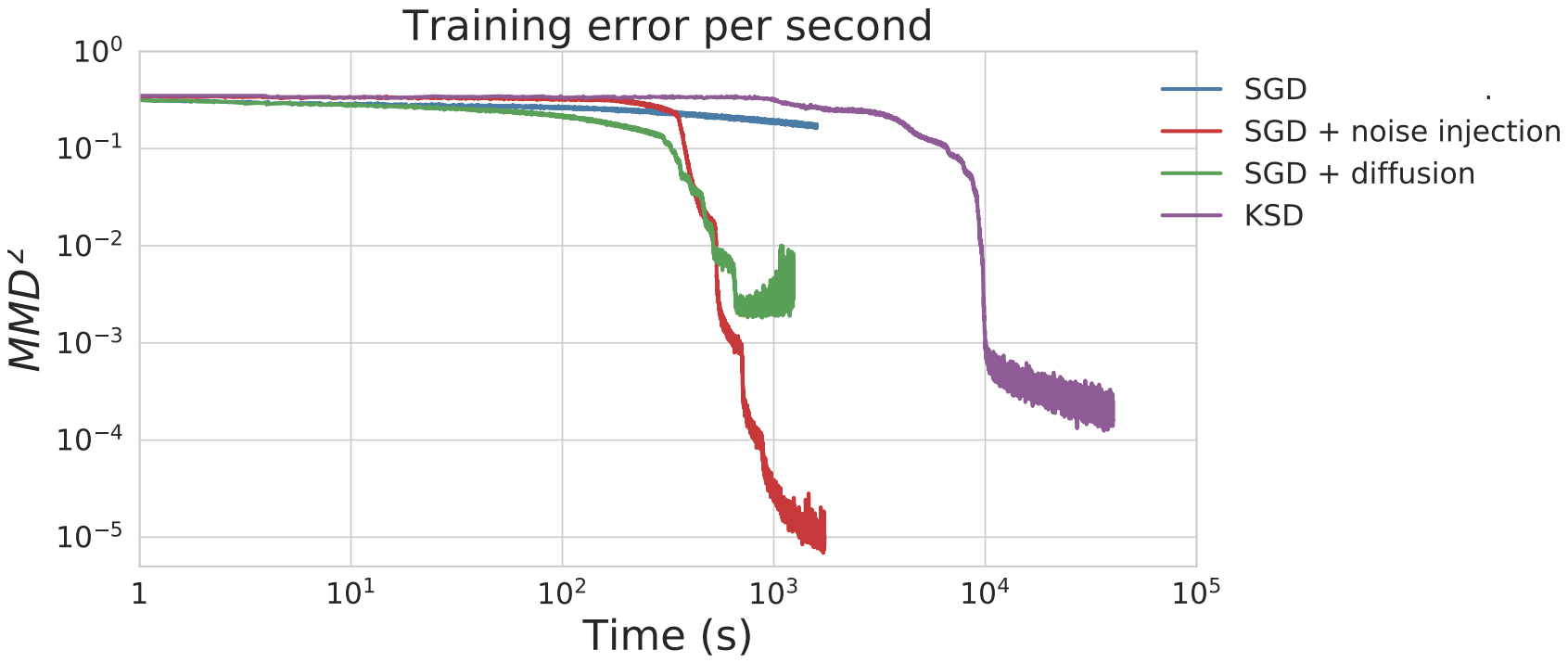
Noise Injection: Experiments



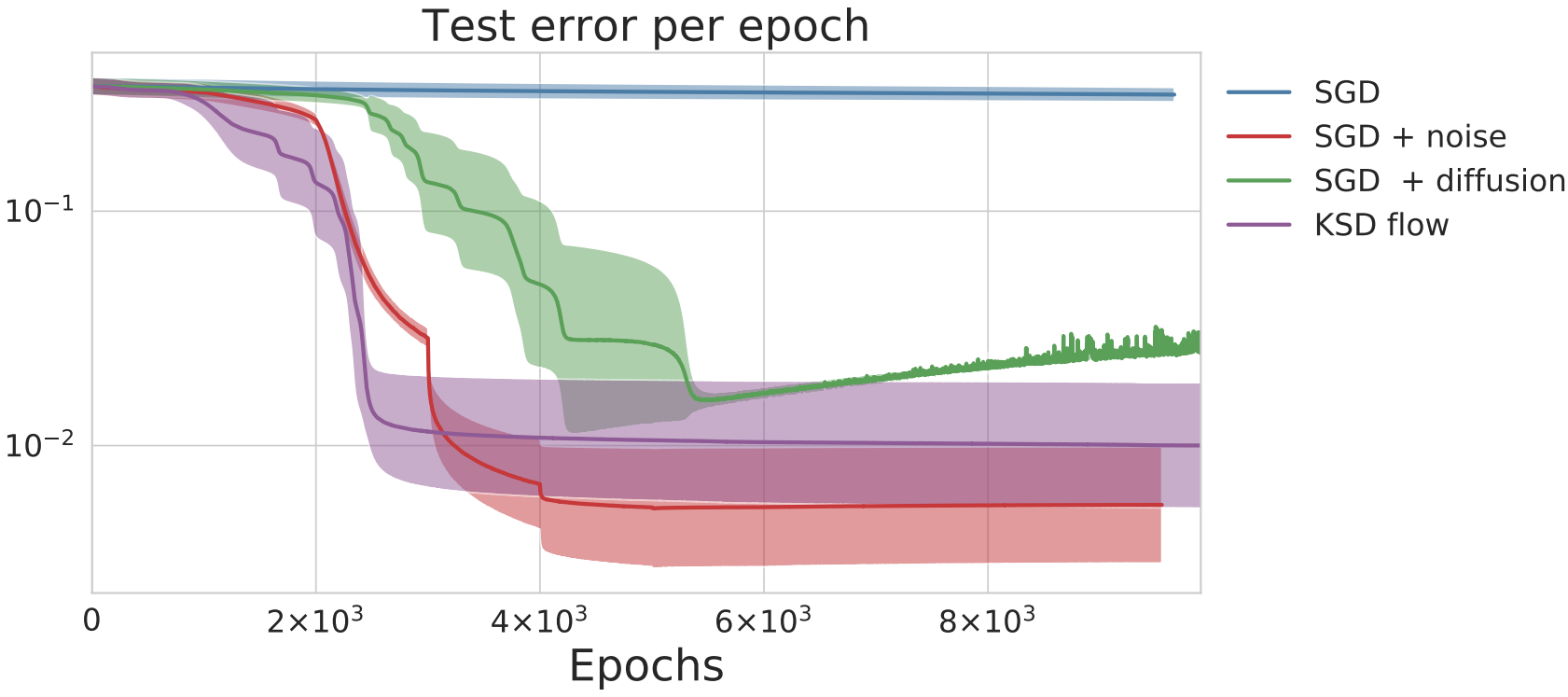
Noise Injection: Experiments



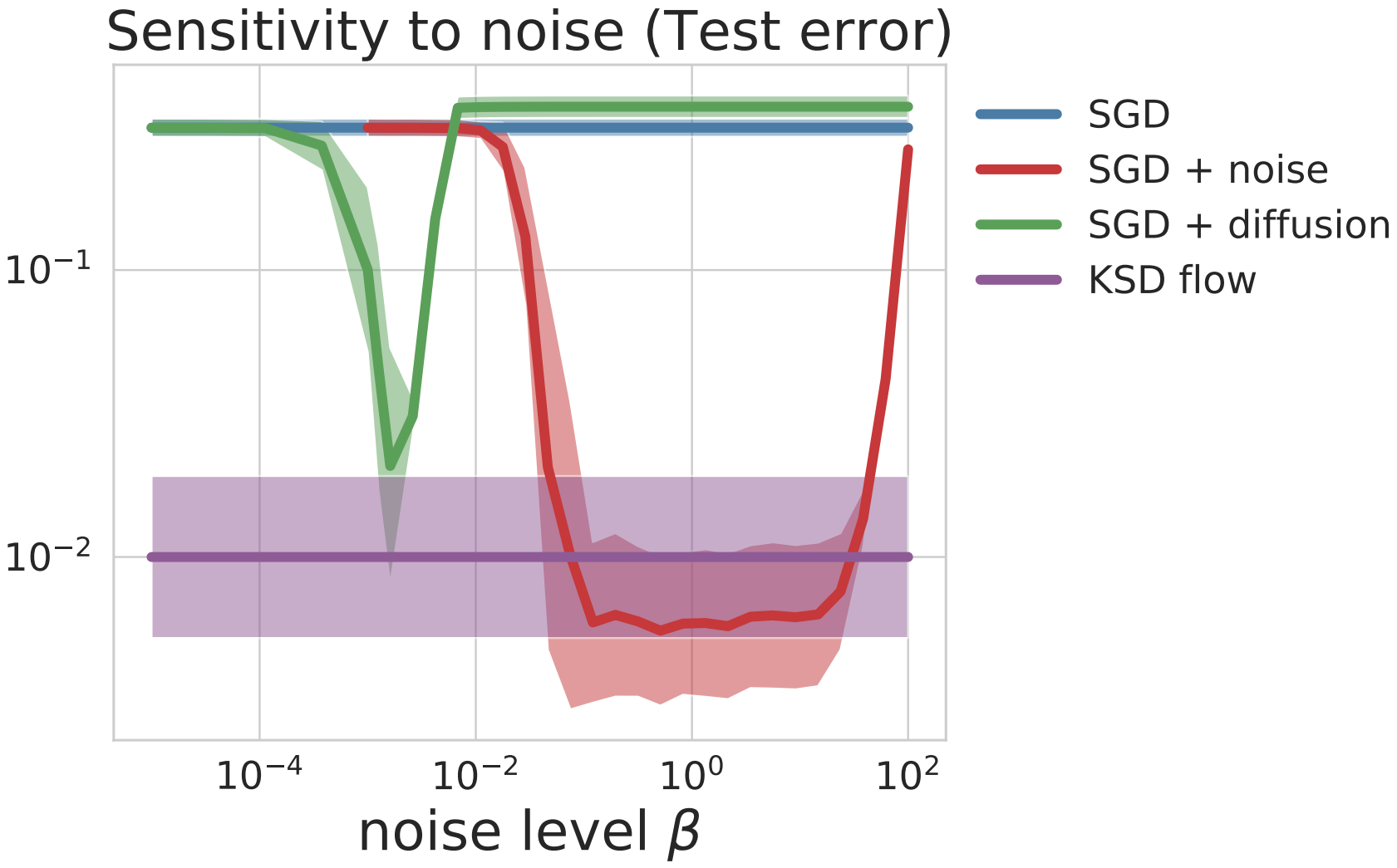
Noise Injection: Experiments



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.

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].
arXiv: 1704.04932.

Chizat, L. and Bach, F. (2018).
On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012).

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

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