Maximum Mean Discrepancy Gradient Flow

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

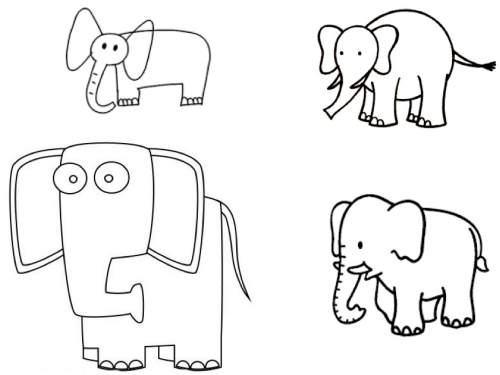
[1] Gatsby Computational Neuroscience Unit, UCL, London

[2] Visual Computing Center, KAUST, Saudi Arabia

September 27, 2019

General Problem

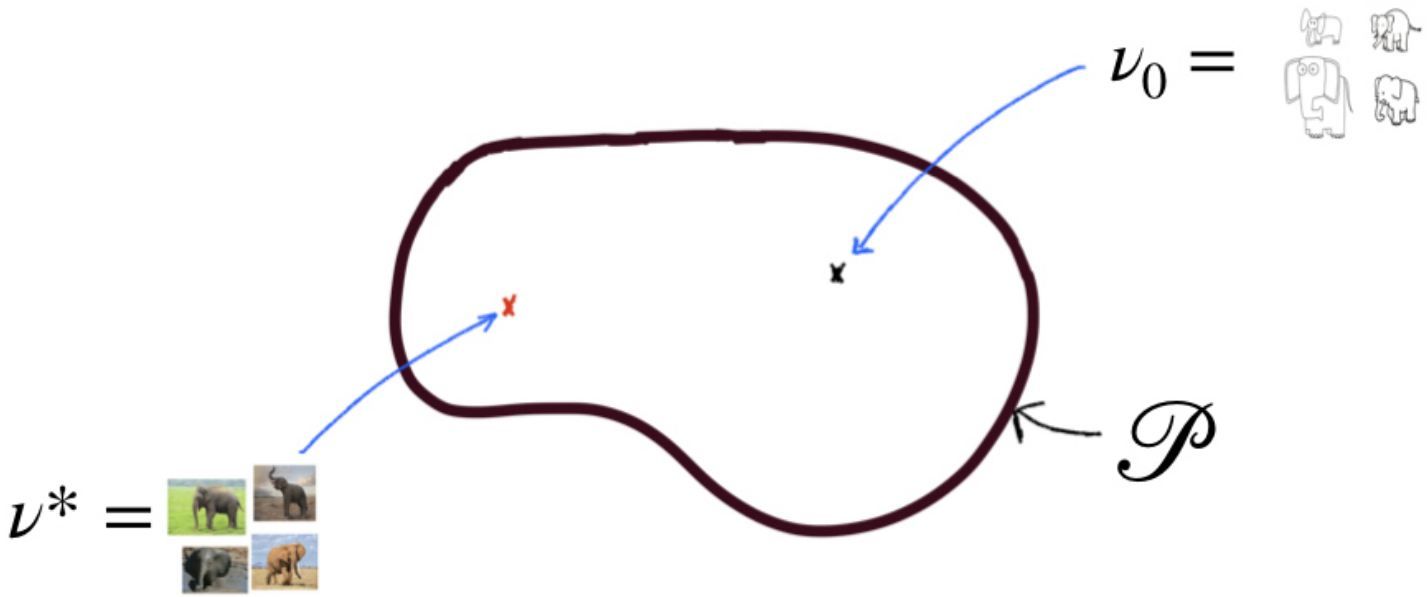
ν^*



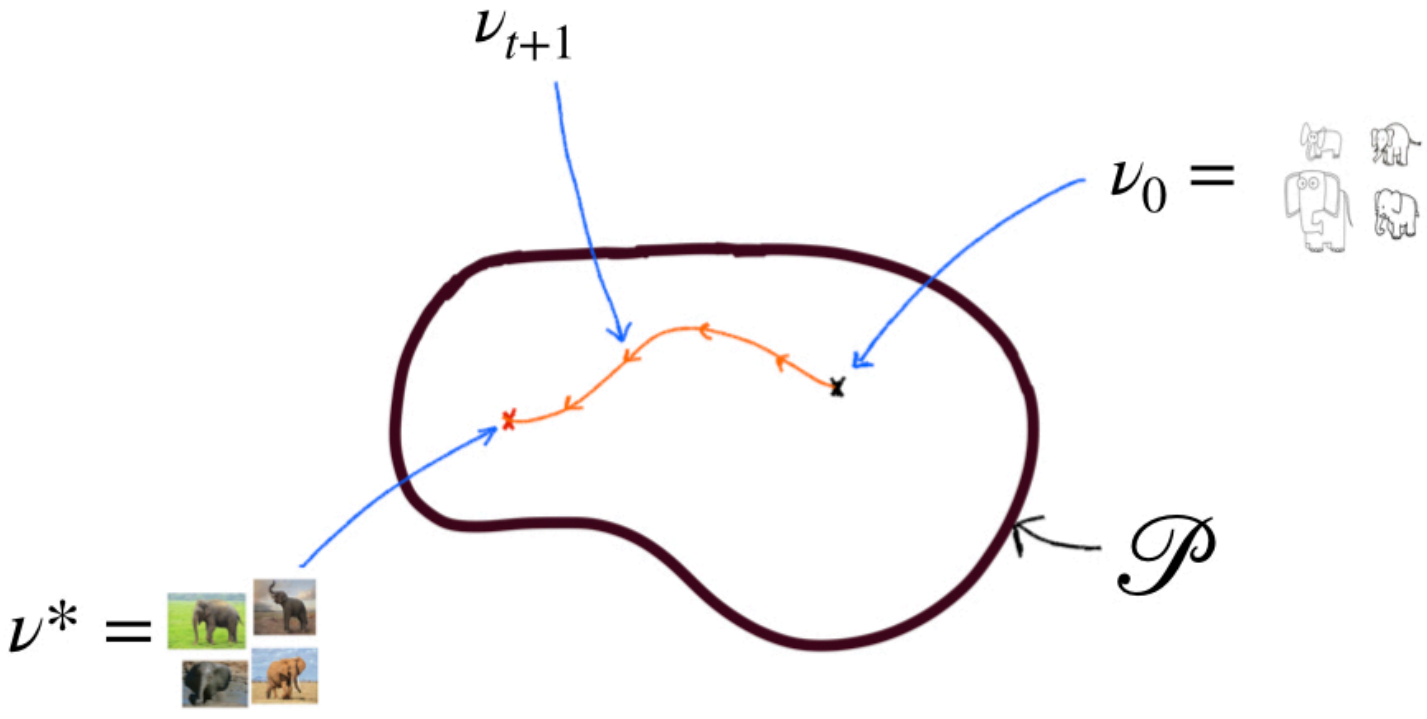
ν_0

[Goodfellow et al., 2014]

General Problem

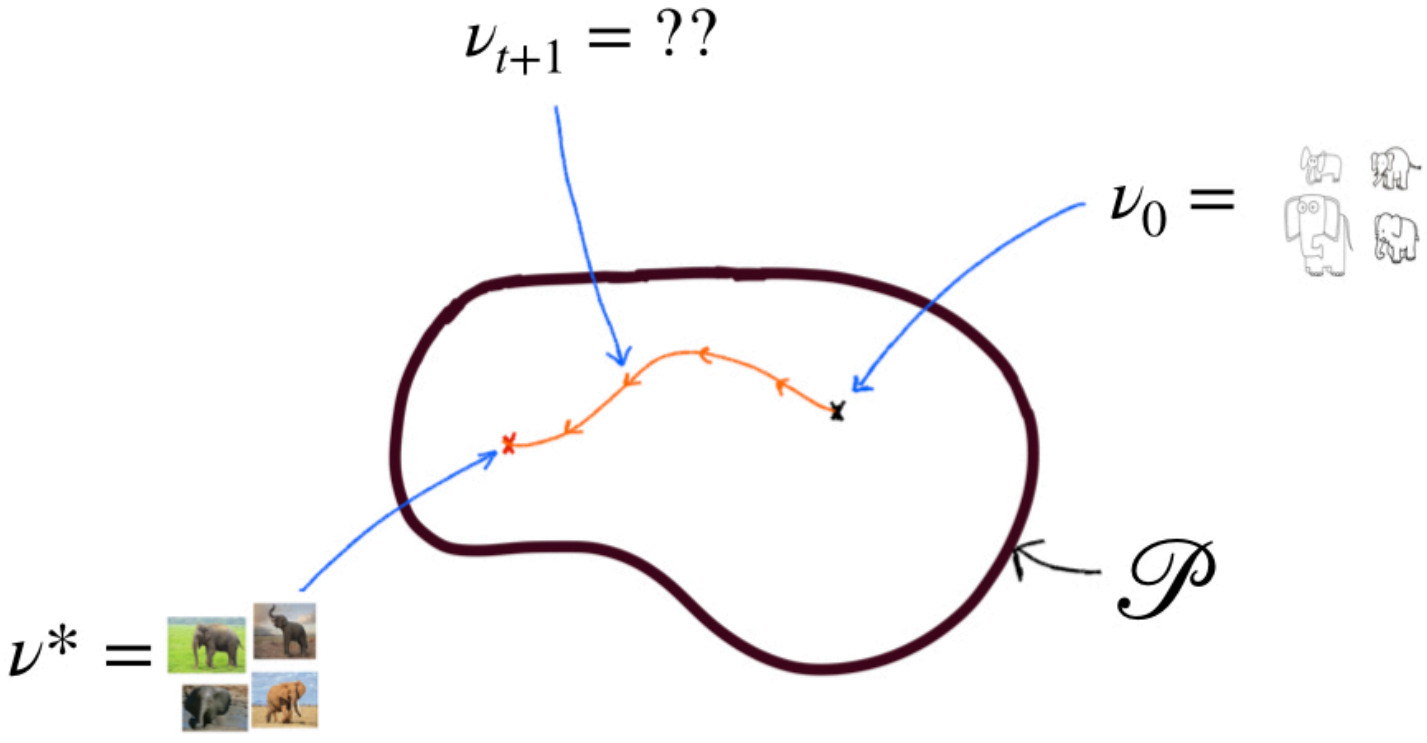


General Problem



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

General Problem

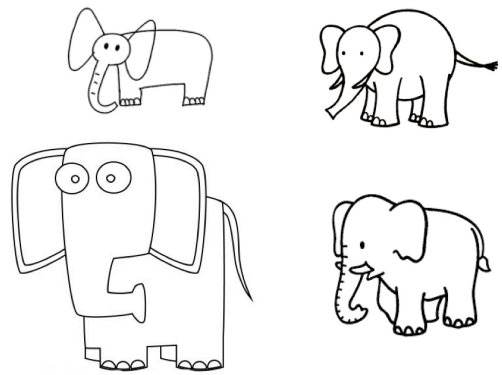


$$\min_{\nu \in \mathscr{P}} MMD(\nu^*, \nu)$$

Motivation: Implicit generative models [Goodfellow et al., 2014]

Given samples from a distribution ν^* over $\mathscr{X}$, want a model that can produce new samples from $\nu_\psi \approx \nu^*$

ν^*

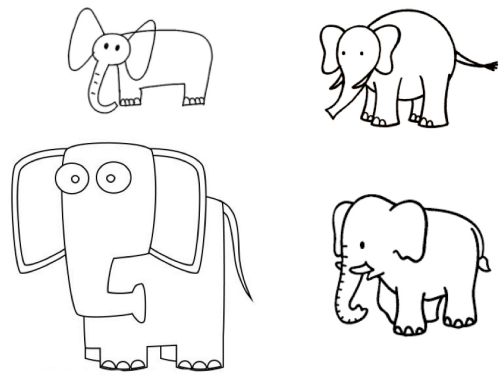


ν_ψ

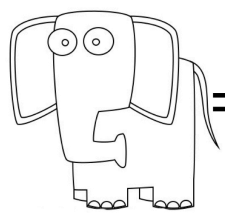
Motivation: Implicit generative models [Goodfellow et al., 2014]

Given samples from a distribution ν^* over $\mathscr{X}$, want a model that can produce new samples from $\nu_\psi \approx \nu^*$

ν^*



ν_ψ



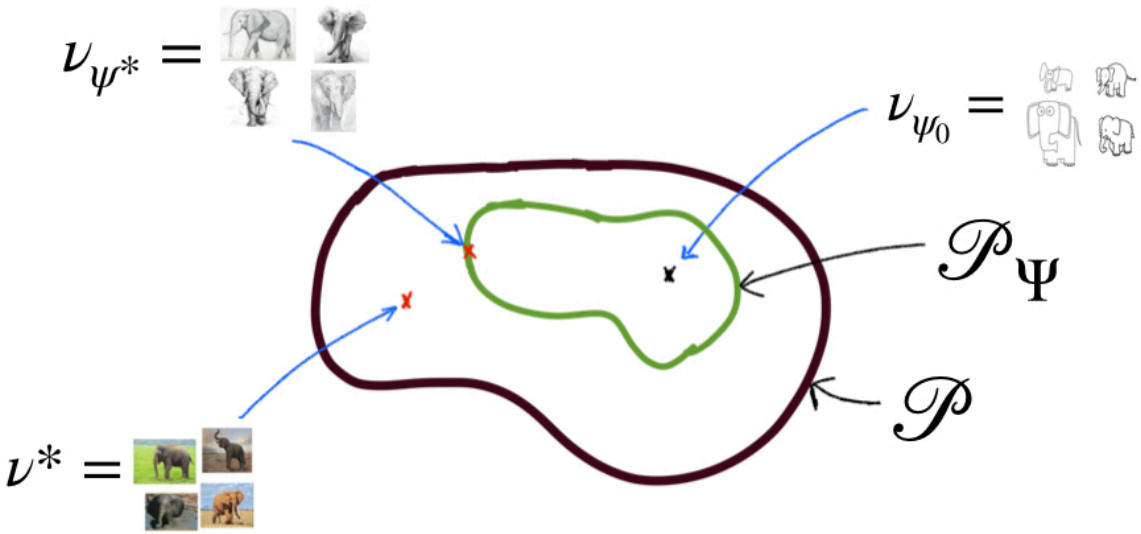
$= G_\psi(z)$

z sampled from a fixed known distribution.

[Goodfellow et al., 2014]

Implicit generative models

Learn best ν_{ψ^*} by minimizing a distance/divergence d between ν_ψ and ν^*: Constrained optimization



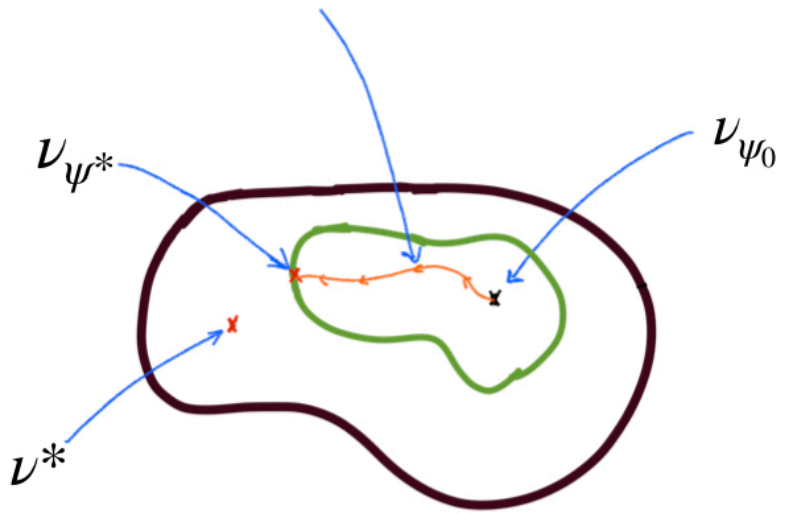
$$\min_{\nu_\psi \in \mathscr{P}_\Psi} d(\nu^*, \nu_\psi)$$

[Goodfellow et al., 2014, Nowozin et al., 2016, Arjovsky et al., 2017]

Implicit generative models

Learn best ν_{ψ^*} by minimizing a distance/divergence d between ν_ψ and ν^* : Constrained optimization

$$\nu_{\psi_{t+1}} : \psi_{t+1} = \psi_t - \gamma_t \nabla_\psi \widehat{d(\nu^*, \nu_{\psi_t})}$$



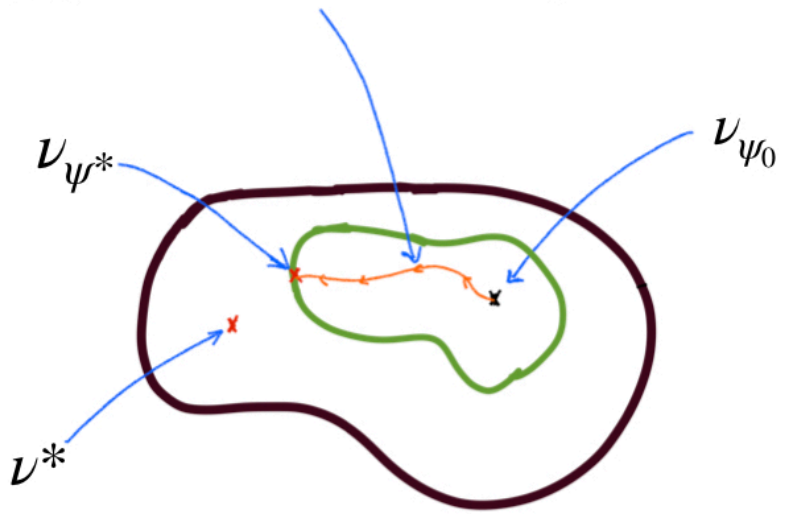
$$\min_{\nu_\psi \in \mathscr{P}_\Psi} d(\nu^*, \nu_\psi)$$

[Goodfellow et al., 2014, Li et al., 2015, Nowozin et al., 2016, Arjovsky et al., 2017]

Implicit generative models

Learn best ν_{ψ^*} by minimizing a distance/divergence d between ν_ψ and ν^* : Constrained optimization

$$\nu_{\psi_{t+1}} : \psi_{t+1} = \psi_t - \gamma_t \nabla_\psi \widehat{MMD}^2(\nu^*, \nu_{\psi_t})$$

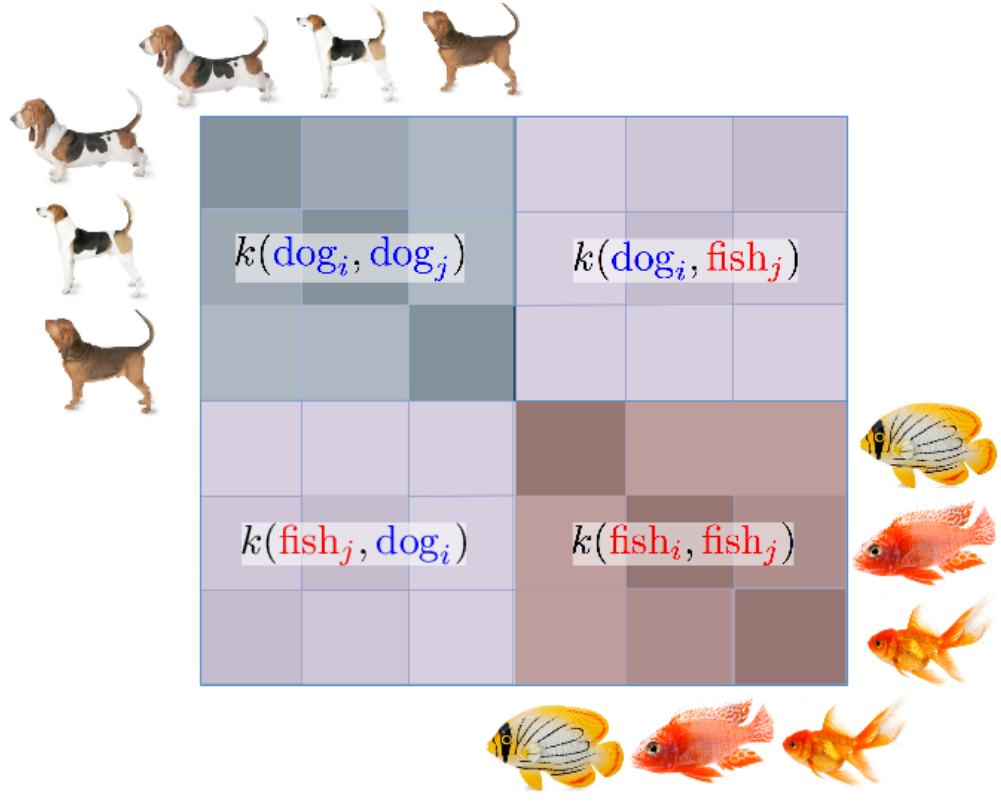


$$\min_{\nu_\psi \in \mathcal{P}_\Psi} MMD(\nu^*, \nu_\psi)$$

[Goodfellow et al., 2014, Nowozin et al., 2016, Arjovsky et al., 2017, Li et al., 2015, Bińkowski et al., 2018]

Choice of d: The Maximum Mean Discrepancy

First choose a kernel k, then compute a similarity matrix:



$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(dog_i, dog_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(fish_i, fish_j) - \frac{2}{n^2} \sum_{i,j} k(dog_i, fish_j)$$

[Gretton et al., 2012]

Choice of d: The Maximum Mean Discrepancy

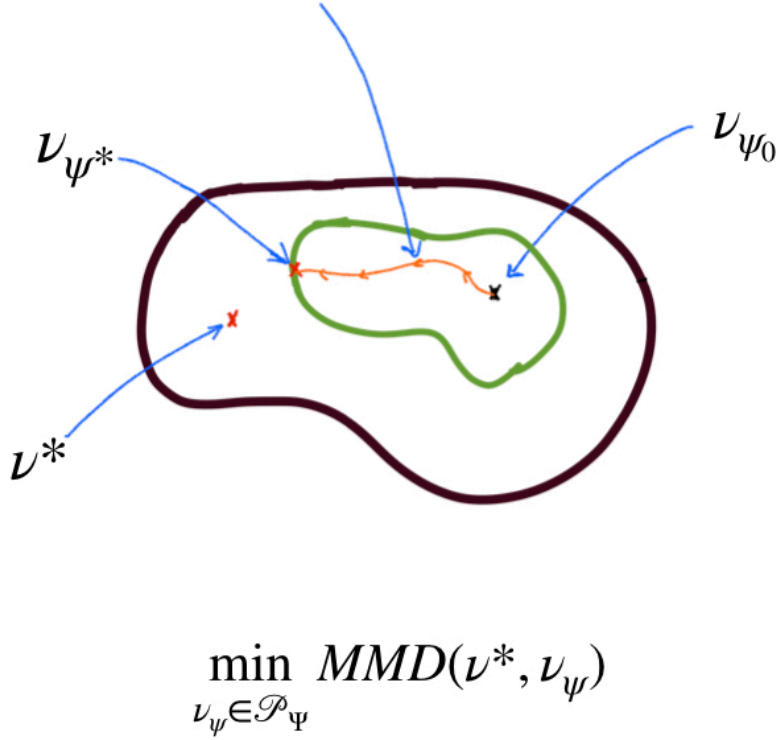
- Easy to estimate using samples
- Nice statistical guarantees [Gretton et al., 2012]
- Works well in practice[1] [Li et al., 2015, Bińkowski et al., 2018, Arbel et al., 2018]

[1] When the kernel is parameterized by a deep neural network

Implicit generative models

Guarantees that $\nu_{\psi_t} \to \nu_{\psi^*}$?

$$\nu_{\psi_{t+1}} : \psi_{t+1} = \psi_t - \gamma_t \nabla_\psi \widehat{MMD}^2(\nu^*, \nu_{\psi_t})$$



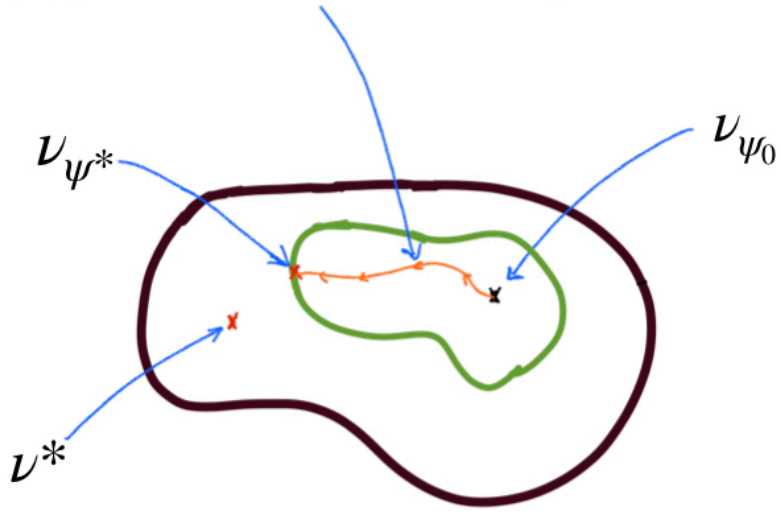
$$\min_{\nu_\psi \in \mathcal{P}_\Psi} MMD(\nu^*, \nu_\psi)$$

[Goodfellow et al., 2014, Nowozin et al., 2016, Arjovsky et al., 2017, Li et al., 2015, Bińkowski et al., 2018, Arbel et al., 2018]

Implicit generative models

Hard to say anything if $\mathscr{P}_\Psi$ is not convex.

$$\nu_{\psi_{t+1}} : \psi_{t+1} = \psi_t - \gamma_t \nabla_\psi \widehat{MMD}^2(\nu^*, \nu_{\psi_t})$$

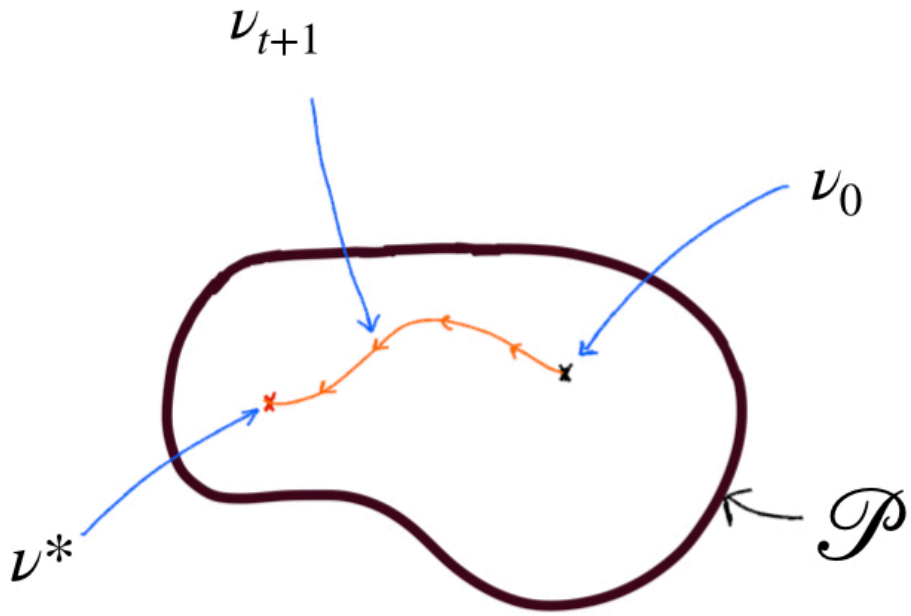


$$\min_{\nu_\psi \in \mathscr{P}_\Psi} MMD(\nu^*, \nu_\psi)$$

[Goodfellow et al., 2014, Nowozin et al., 2016, Arjovsky et al., 2017, Li et al., 2015, Bińkowski et al., 2018, Arbel et al., 2018]

Implicit generative model

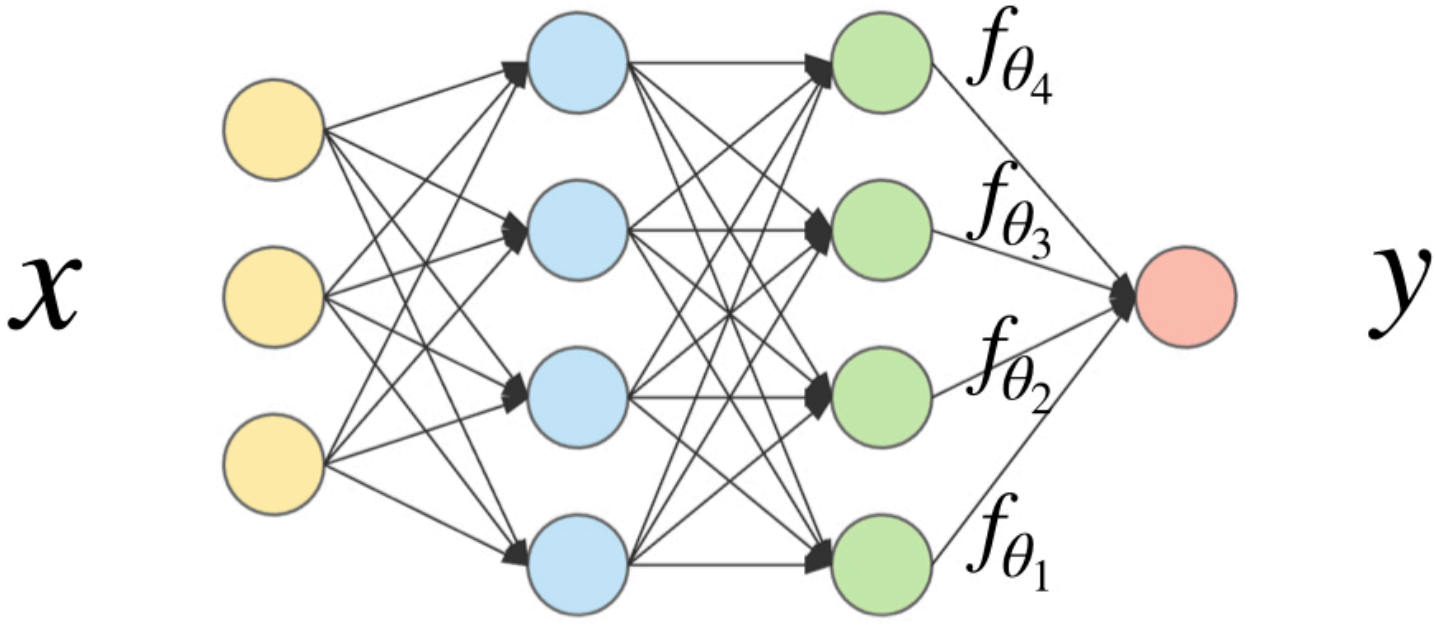
Unconstrained optimization: Replace $\mathscr{P}_\Psi$ by $\mathscr{P}$



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

Motivation: Neural Networks Optimization

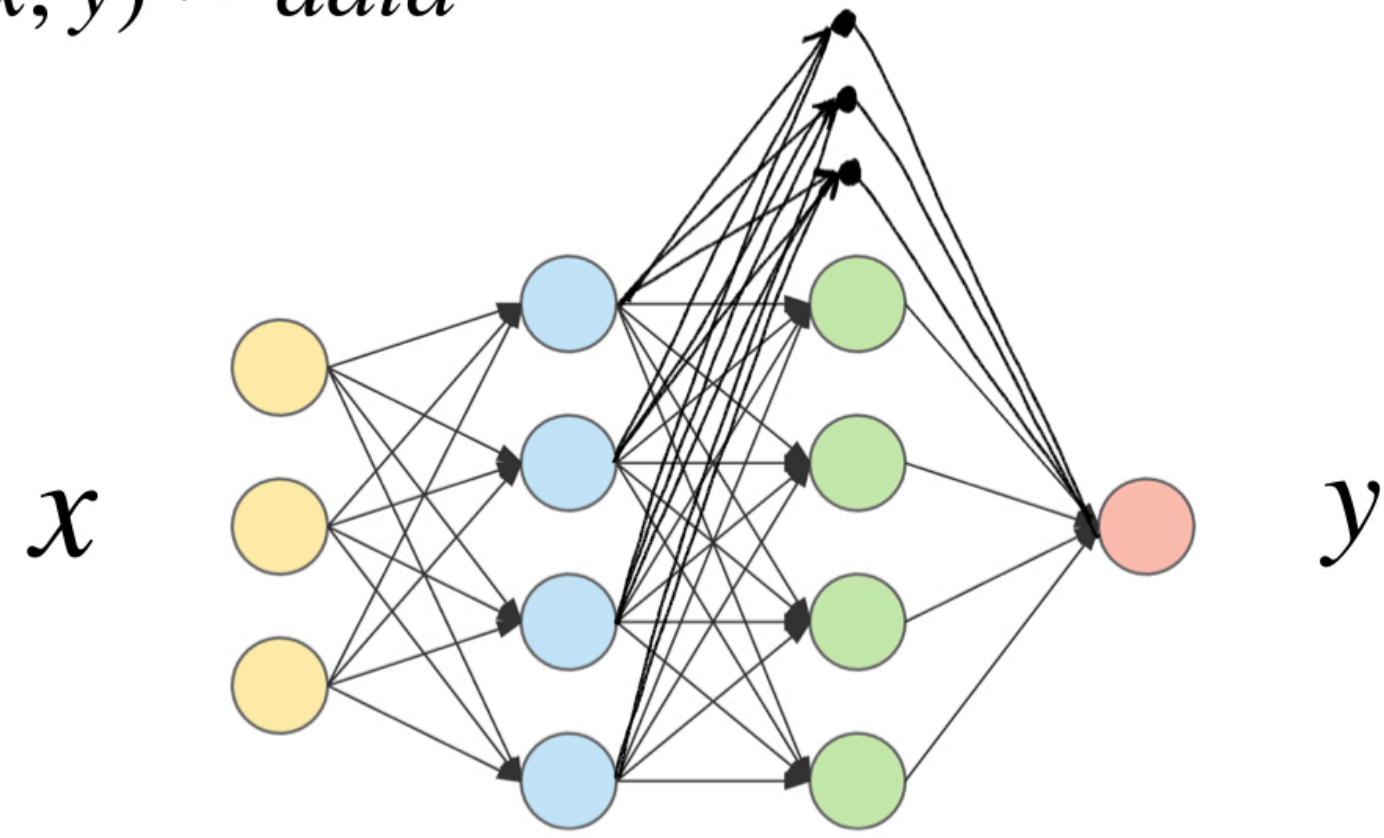
$$(x, y) \sim data$$



$$\min_{\theta_1,\ldots,\theta_N} \mathbb{E}_{data}[\|y - \frac{1}{N}\sum_{i=1}^{N} f_{\theta_i}(x)\|^2]$$

Motivation: Neural Networks Optimization

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



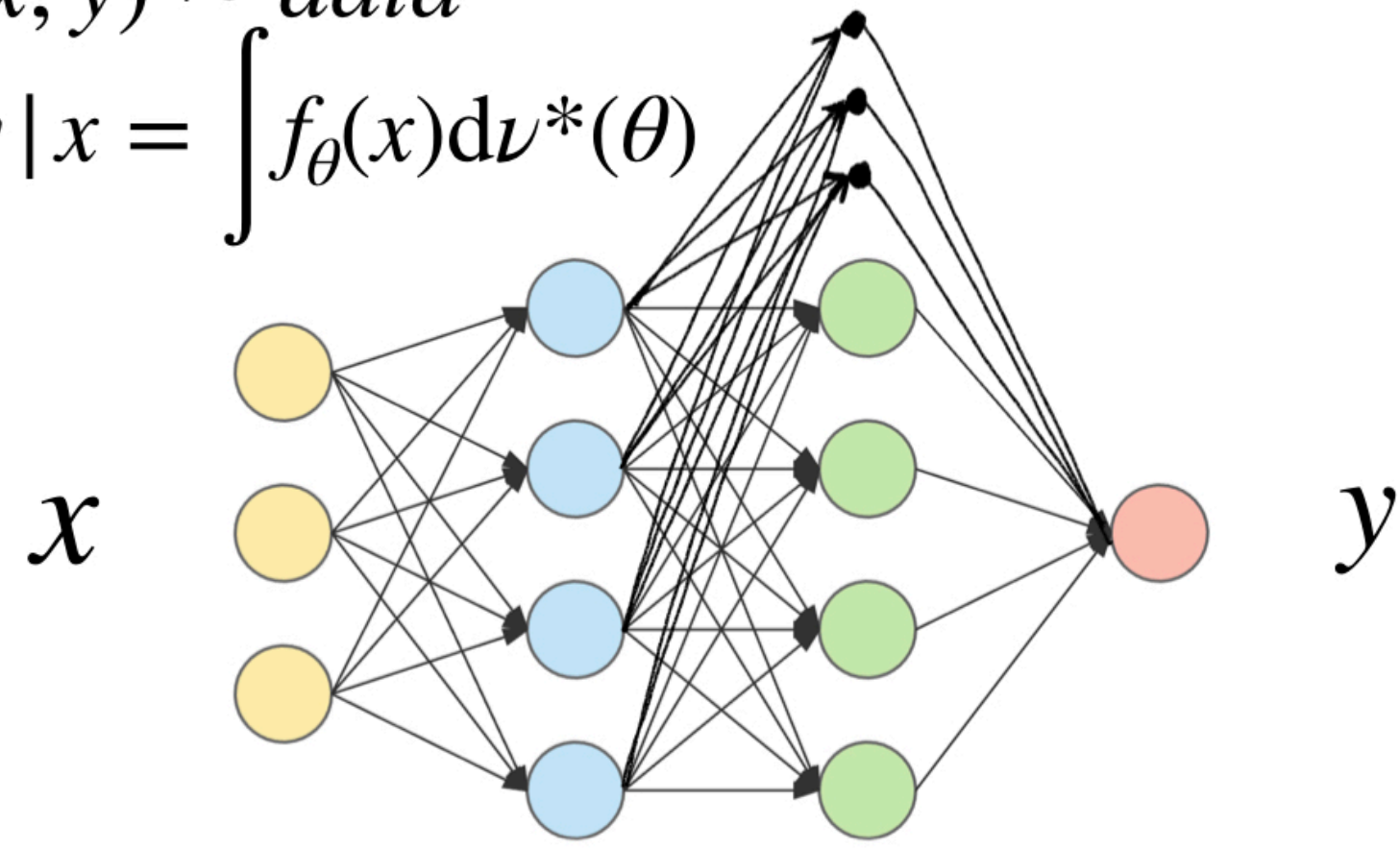
$$\min_{\nu \in \mathcal{P}} \mathbb{E}_{data}[\|y - \int f_\theta(x) \mathrm{d}\nu(\theta)\|^2]$$

[Rotskoff and Vanden-Eijnden, 2018, Chizat and Bach, 2018, Mei et al., 2018]

Motivation: Neural Networks Optimization

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

$$y \,|\, x = \int f_\theta(x) \mathrm{d}\nu^*(\theta)$$

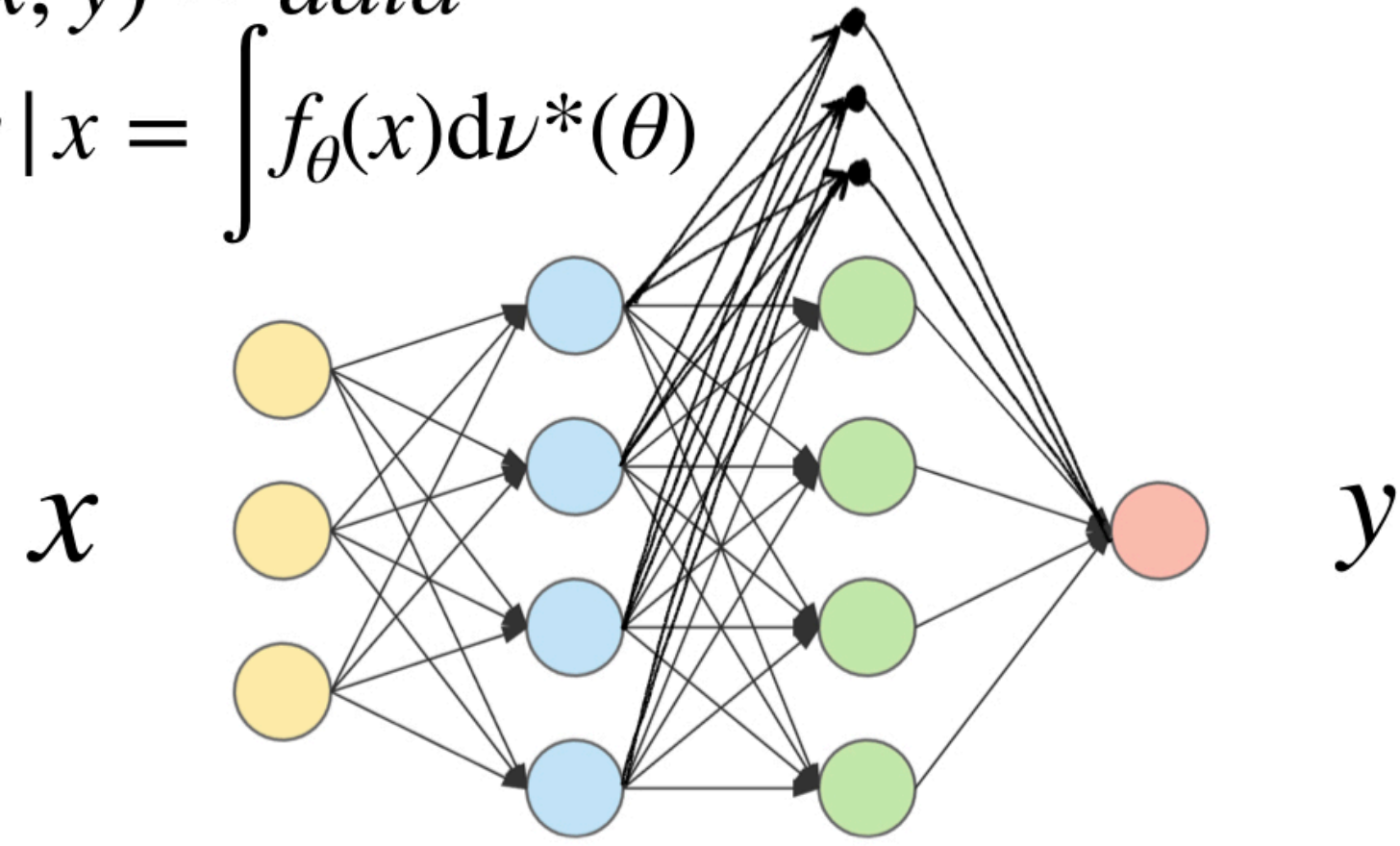


$$\min_{\nu \in \mathcal{P}} \mathbb{E}_{data}[\|y - \int f_\theta(x) \mathrm{d}\nu(\theta)\|^2]$$

Motivation: Neural Networks Optimization

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

$$y \,|\, x = \int f_\theta(x) \mathrm{d}\nu^*(\theta)$$

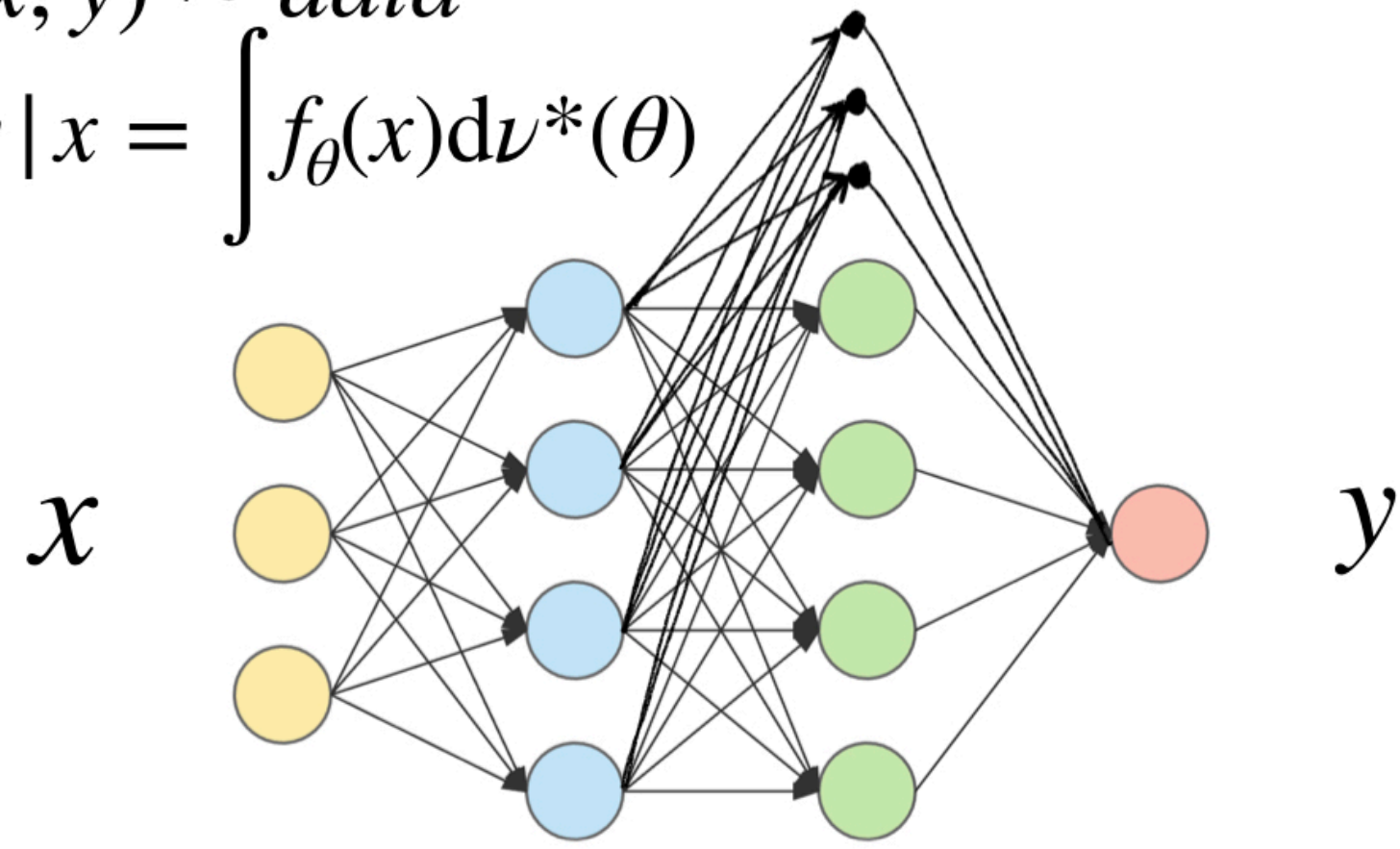


$$\min_{\nu \in \mathscr{P}} \mathbb{E}_{data}[\| \int f_\theta(x) \mathrm{d}\nu^*(\theta) - \int f_\theta(x) \mathrm{d}\nu(\theta) \|^2]$$

Motivation: Neural Networks Optimization

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

$$y \,|\, x = \int f_\theta(x) \mathrm{d}\nu^*(\theta)$$

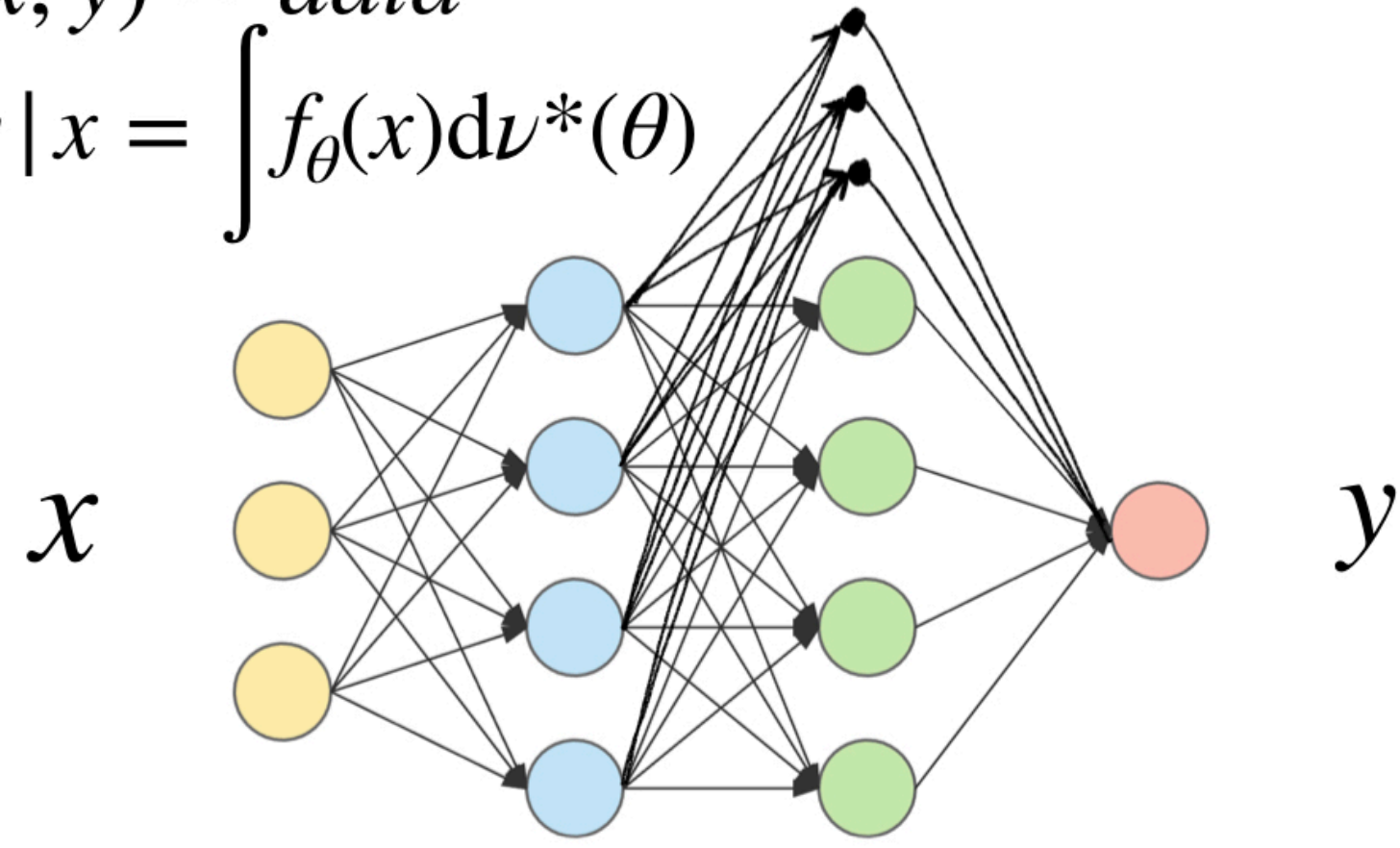


$$\min_{\nu \in \mathscr{P}} MMD^2(\nu^*, \nu)$$

Motivation: Neural Networks Optimization

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

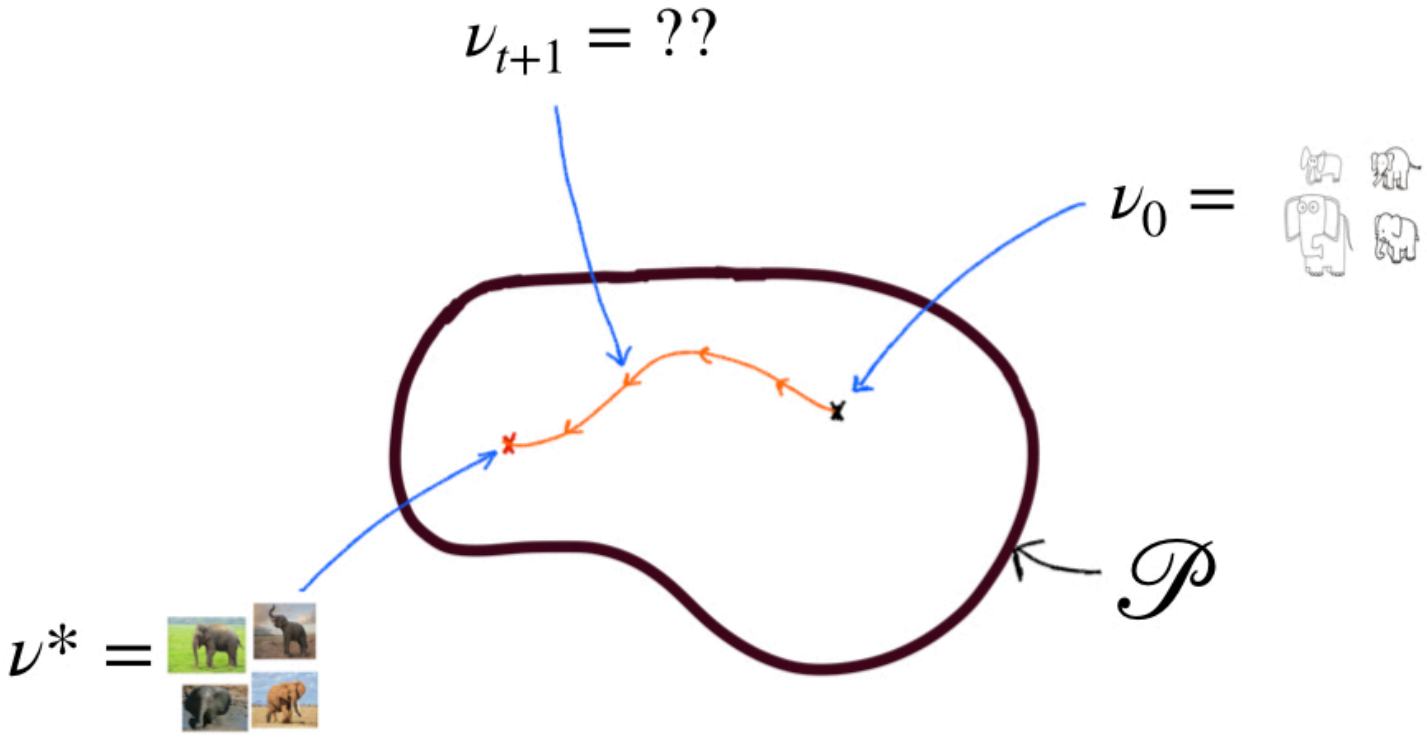
$$y \mid x = \int f_\theta(x) \mathrm{d}\nu^*(\theta)$$



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

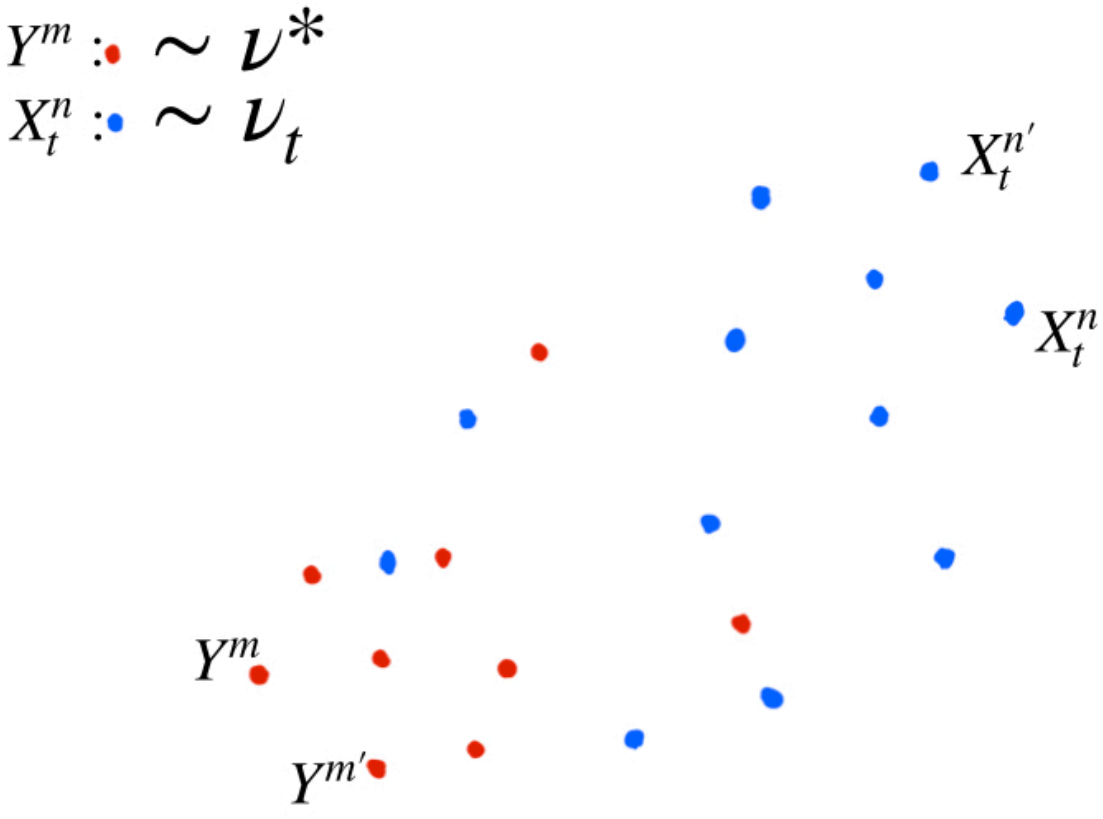
$$k(\theta, \theta') = \mathbb{E}_{data}[f_\theta(x) f_{\theta'}(x)]$$

General problem



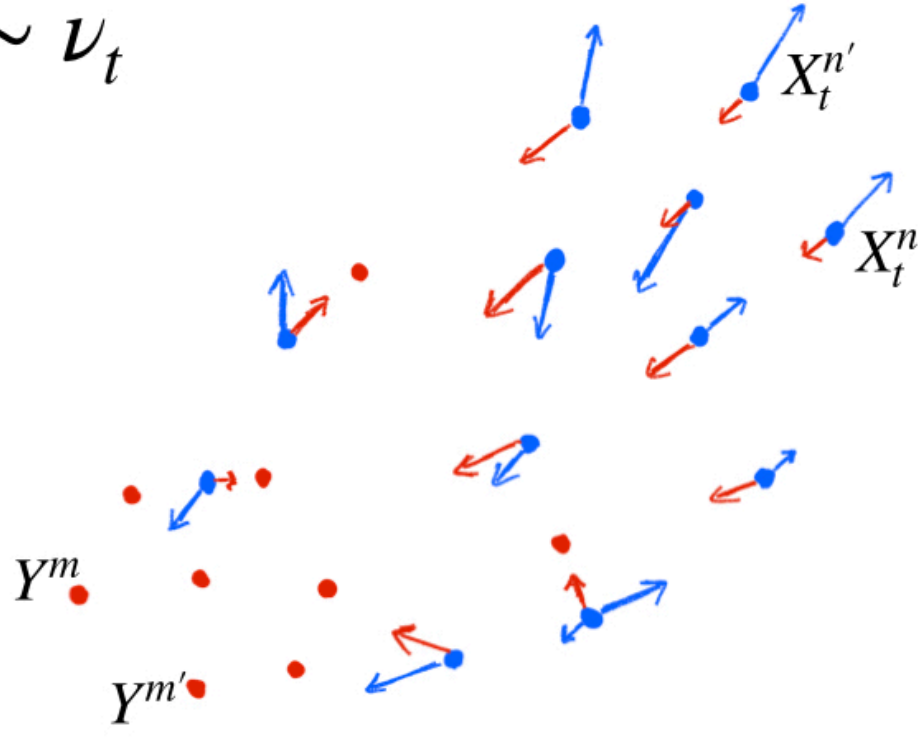
$$\min_{\nu \in \mathscr{P}} MMD(\nu^*, \nu)$$

A physical interpretation of the MMD



A physical interpretation of the MMD

$$Y^m : \sim \nu^*$$
$$X_t^n : \sim \nu_t$$

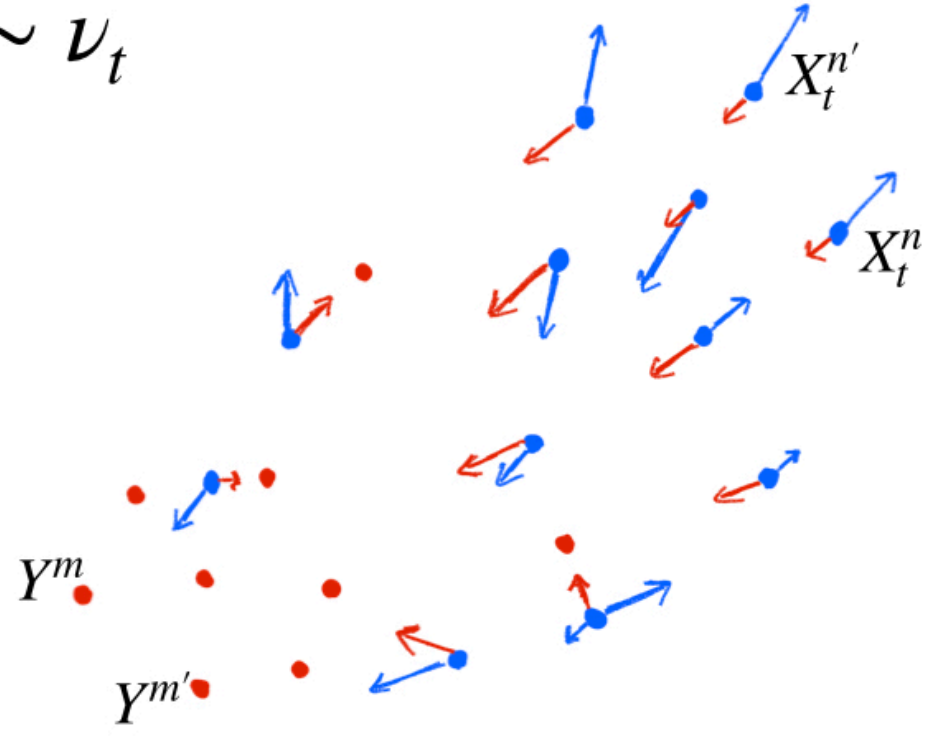


$$M\hat{M}D^2(\nu^*, \nu_t) = \underbrace{-\frac{2}{MN}\sum_{m,n} k(Y^m, X_t^n)}_{\text{confinement potential}} + \underbrace{\frac{1}{N(N-1)}\sum_{n,n'} k(X_t^n, X_t^{n'})}_{\text{interaction potential}} + \underbrace{\frac{1}{M(M-1)}\sum_{m,m'} k(Y^m, Y}_{\text{constant term}}$$

A physical interpretation of the MMD

$Y^m : \sim \nu^*$

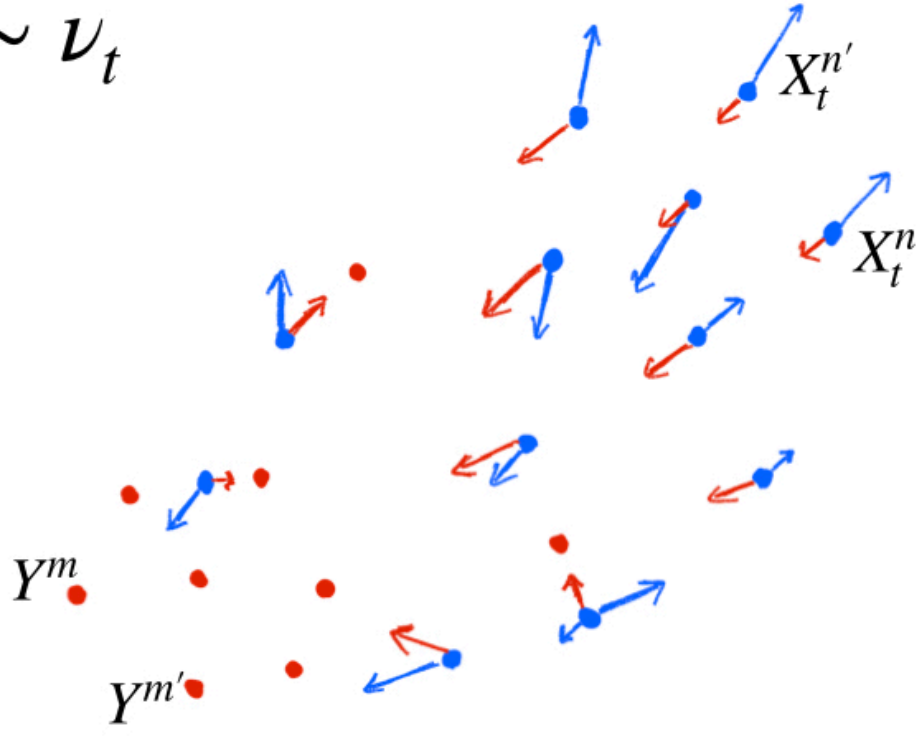
$X_t^n : \sim \nu_t$



$$\nearrow : \underbrace{\frac{1}{M}\sum_m \nabla_1 k(X_t^n, Y^m)}_{\text{confinement force}} \qquad \nearrow : \underbrace{-\frac{1}{N}\sum_{n'} \nabla_1 k(X_t^n, X_t^{n'})}_{\text{interaction force}}$$

A physical interpretation of the MMD

$$Y^m : \sim \nu^*$$
$$X_t^n : \sim \nu_t$$

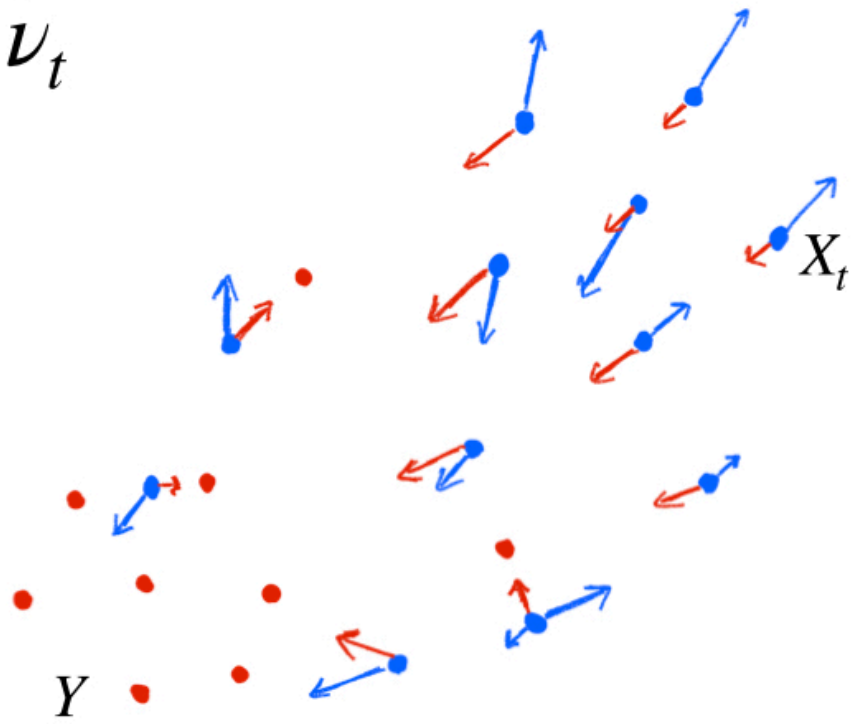


$$X_{t+1}^n = X_t^n + \gamma_t\left(\frac{1}{M}\sum_m \nabla_1 k(X_t^n, Y^m) - \frac{1}{N}\sum_{n'} \nabla_1 k(X_t^n, X_t^{n'})\right)$$

A physical interpretation of the MMD

$Y : \bullet \sim \nu^*$

$X_t : \bullet \sim \nu_t$



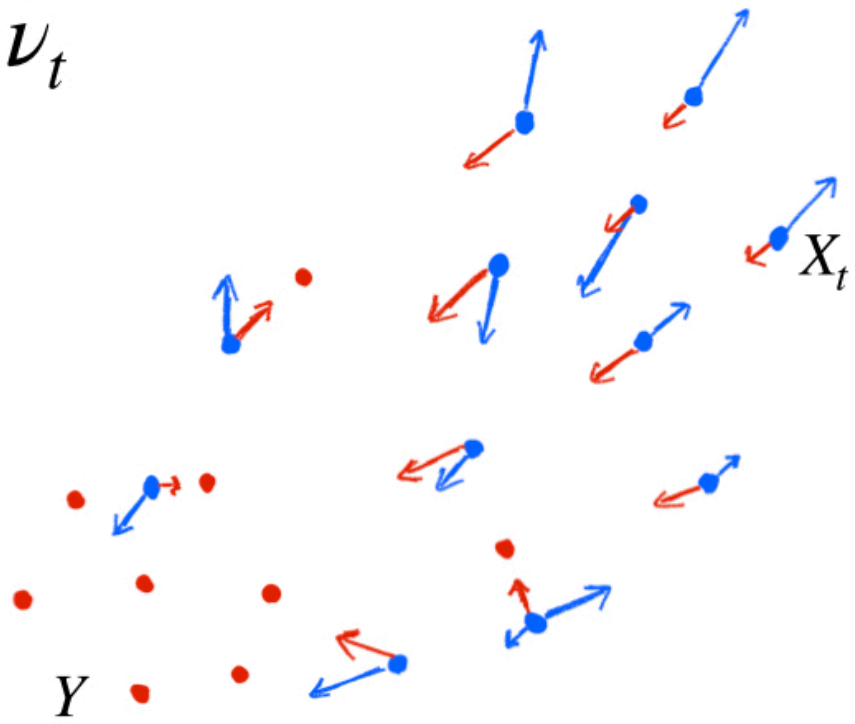
$$X_{t+1} = X_t + \gamma_t \underbrace{\left(\int \nabla_1 k(X_t, y) \mathrm{d}\nu^*(y) - \int \nabla_1 k(X_t, x_t') \mathrm{d}\nu_t(x_t') \right)}_{-\nabla f_t(X_t),}$$

f_t: **potential function**

A physical interpretation of the MMD

$Y : \bullet \sim \nu^*$

$X_t : \bullet \sim \nu_t$



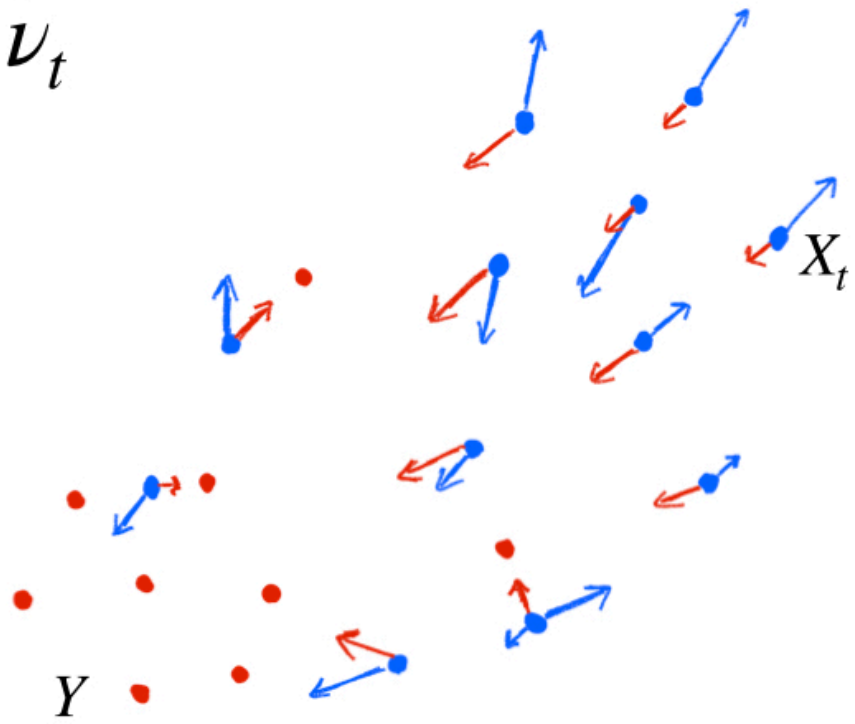
$$X_{t+1} = X_t + \gamma_t \underbrace{\left(\int \nabla_1 k(X_t, y) \mathrm{d}\nu^*(y) - \int \nabla_1 k(X_t, x_t') \mathrm{d}\nu_t(x_t') \right)}_{-\nabla f_t(X_t),}$$

f_t: **potential function**

A physical interpretation of the MMD

$Y : \bullet \sim \nu^*$

$X_t : \bullet \sim \nu_t$



$$\nu_{t+1} : \quad X_{t+1} = X_t - \gamma_t \nabla f_t(X_t), \quad X_t \sim \nu_t$$

Descent direction

- For a small step-size γ:

$$MMD^2(\nu^*, \nu_{t+1}) - MMD^2(\nu^*, \nu_t) \leq -C\gamma \int \|\nabla f_t(x)\|^2 \, \mathrm{d}\nu_t(x)$$

Descent direction

- For a small step-size γ:

$$MMD^2(\nu^*, \nu_{t+1}) - MMD^2(\nu^*, \nu_t) \leq -C\gamma \int \|\nabla f_t(x)\|^2 \, \mathrm{d}\nu_t(x)$$

- ν_{t+1} : Generalized gradient descent algorithm [Villani, 2004].

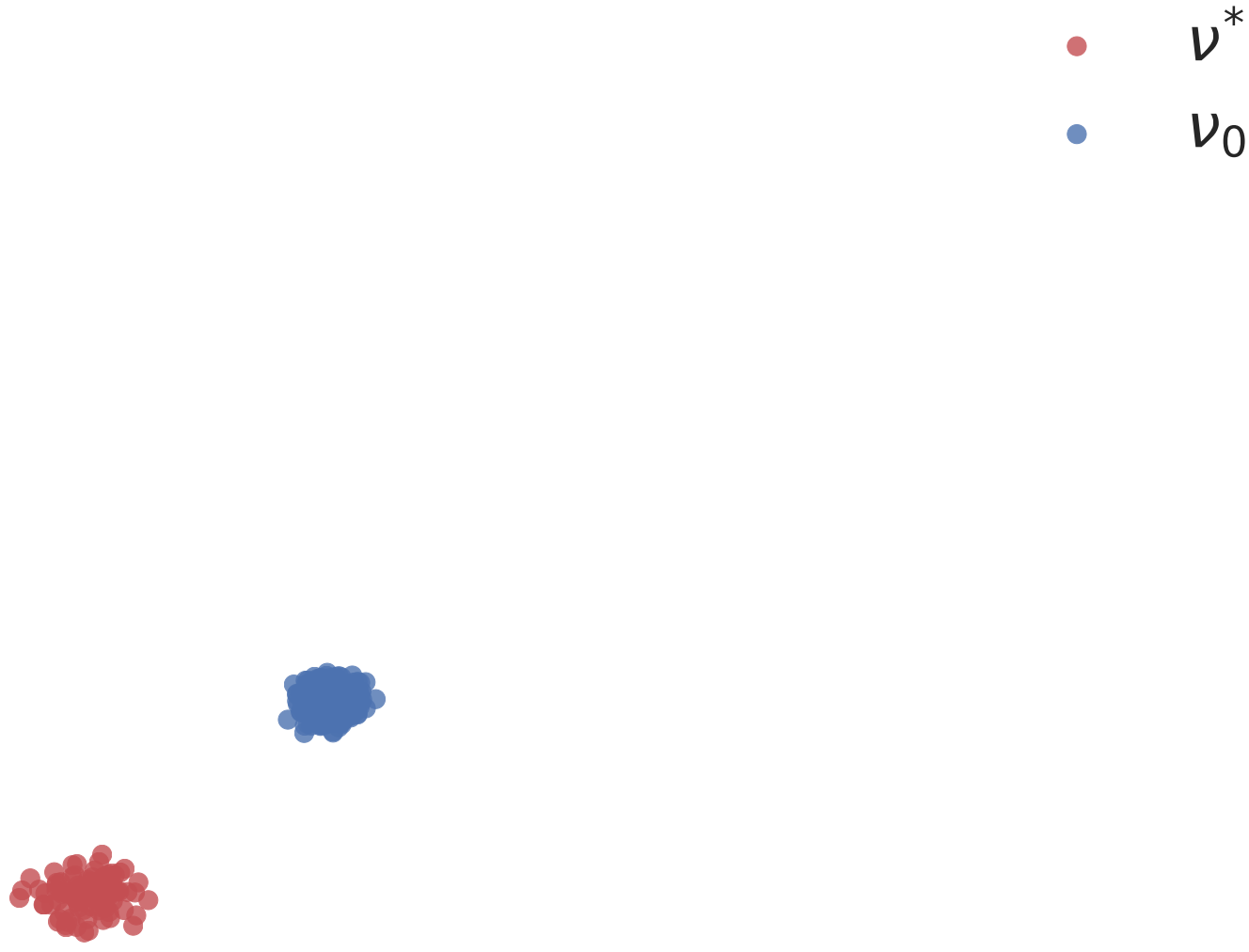
Descent direction

- For a small step-size γ:

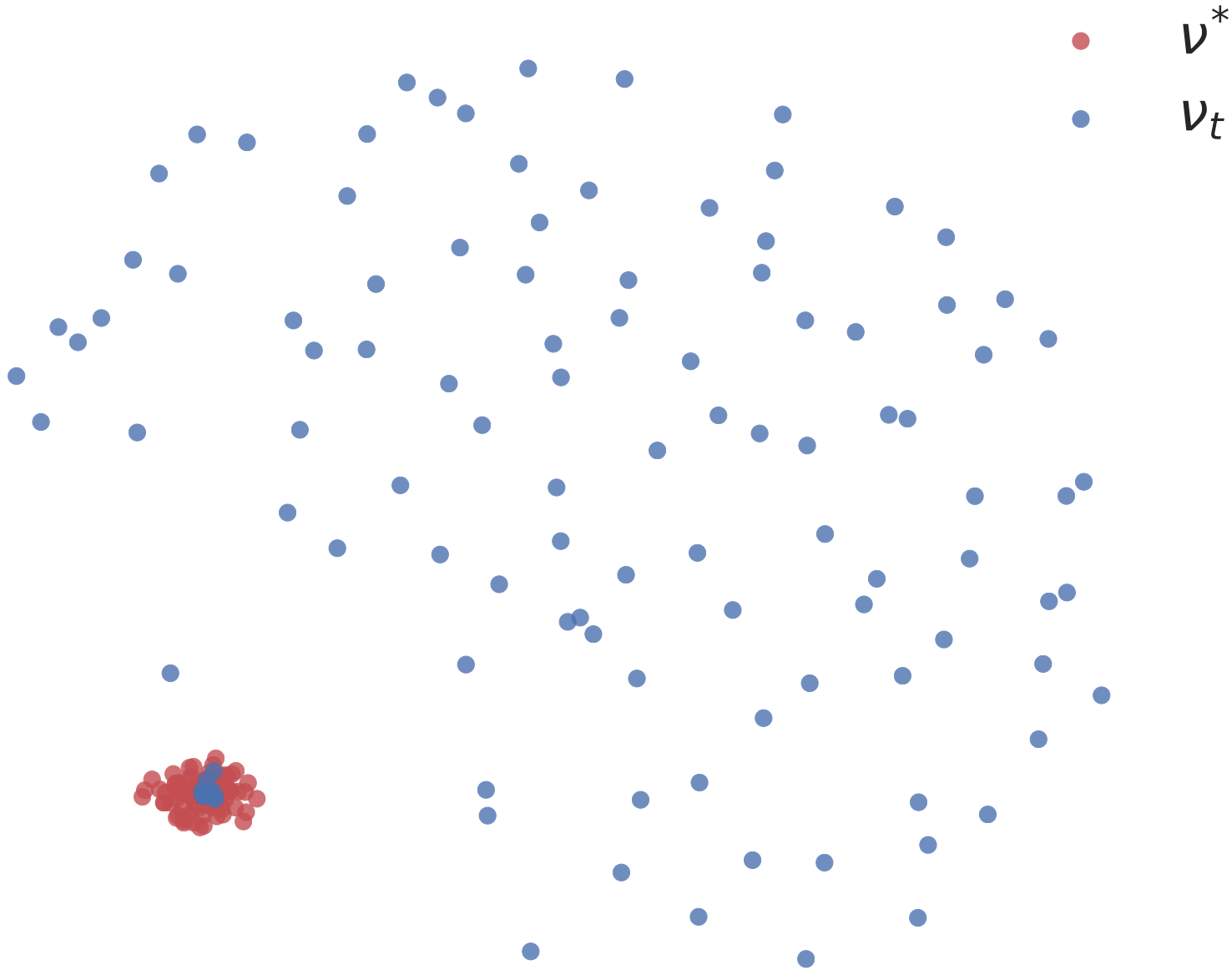
$$MMD^2(\nu^*, \nu_{t+1}) - MMD^2(\nu^*, \nu_t) \leq -C\gamma \int \|\nabla f_t(x)\|^2 \, \mathrm{d}\nu_t(x)$$

- ν_{t+1} : Generalized gradient descent algorithm [Villani, 2004].
- Convergence: $\nu_t \to \nu^*$?

Convergence: Failure case

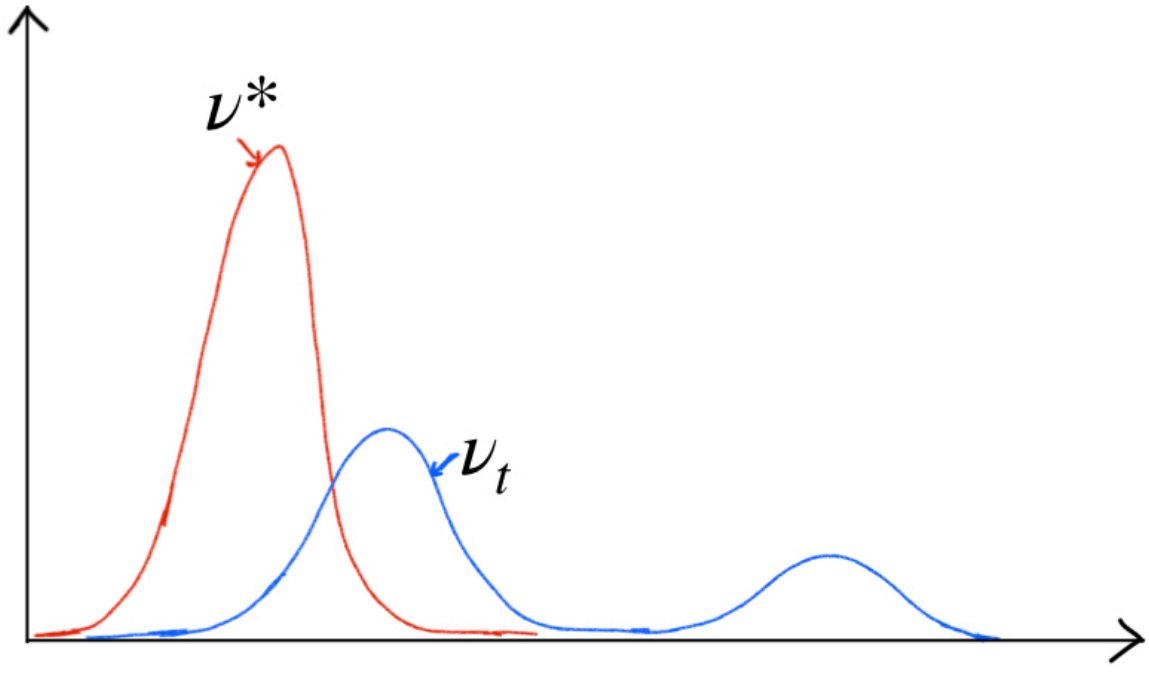


Convergence: Failure case



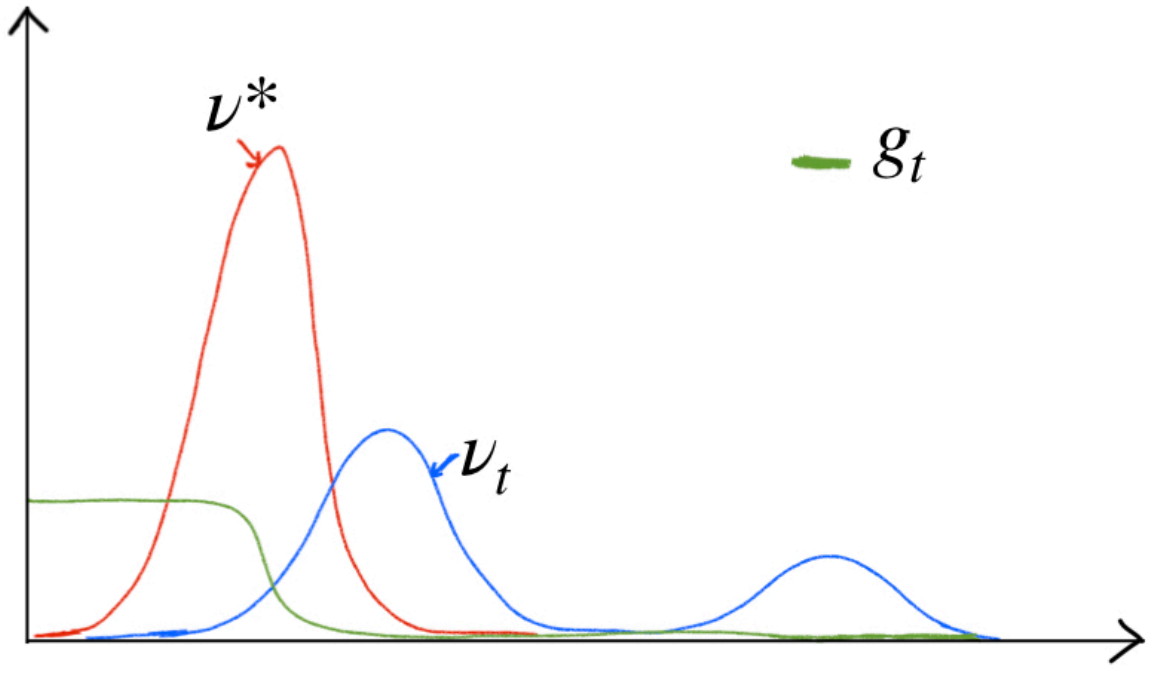
A criterion for convergence

Similar to a 'mode collapse'



A criterion for convergence

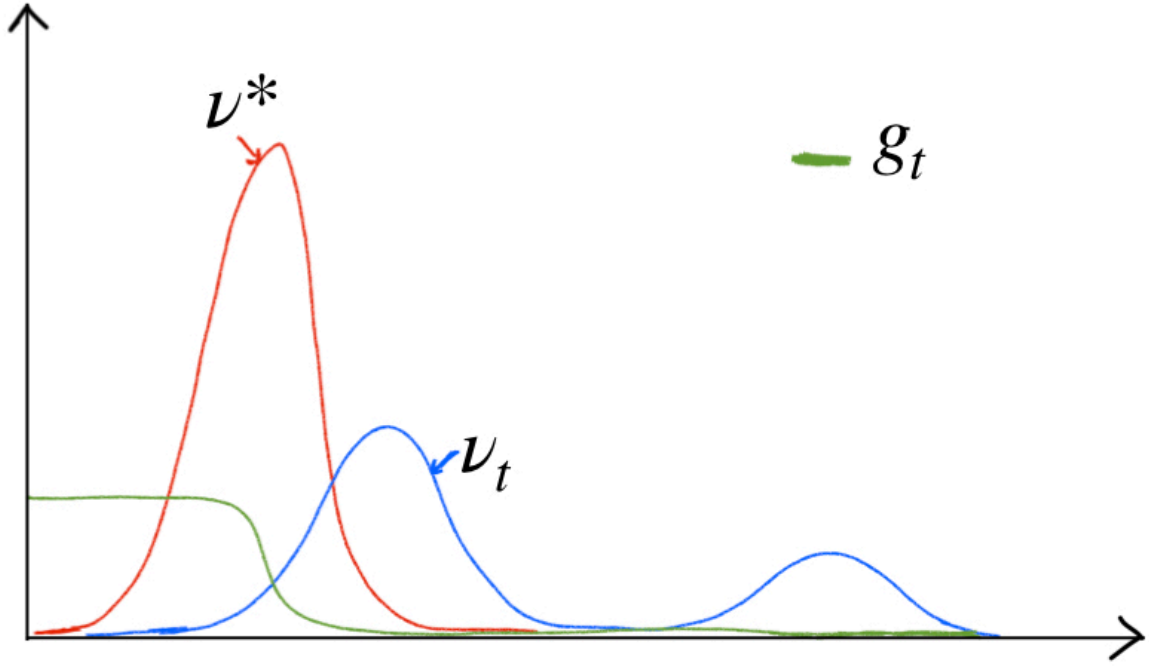
Similar to a 'mode collapse': Can construct g_t to be 'flat' on the support of ν_t



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

A criterion for convergence

Similar to a 'mode collapse': Can construct g_t to be 'flat' on the support of ν_t, and still witnesses the difference between ν^* and ν_t

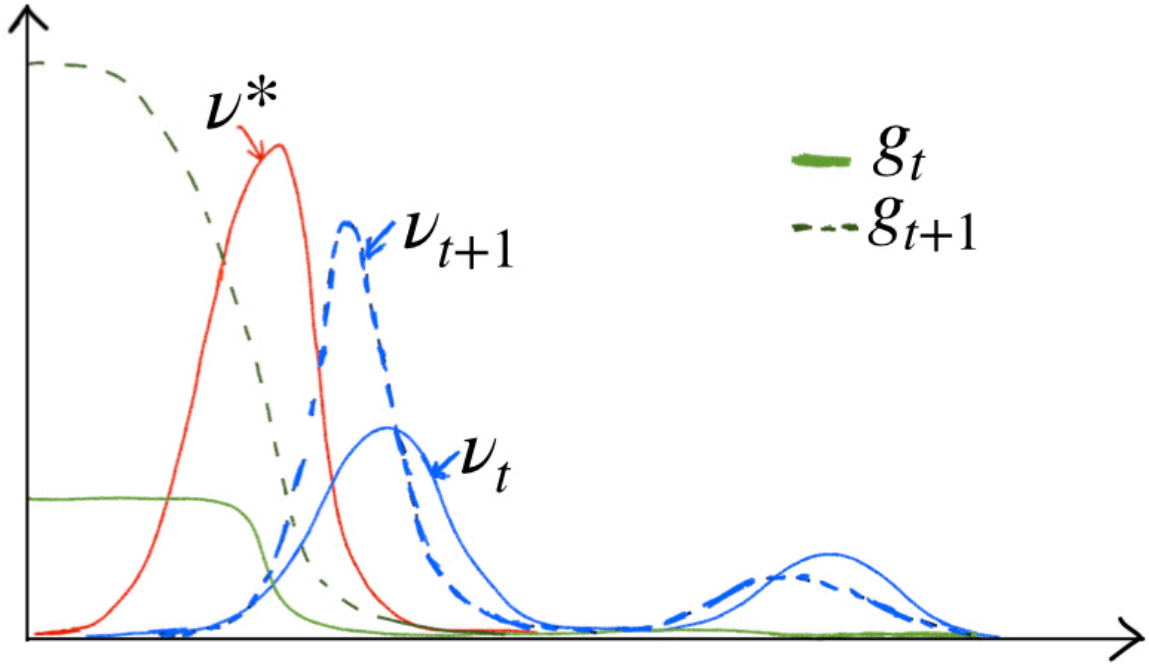


$$\int \|\nabla g_t(x)\|^2 \, d\nu_t(x) \leq 1$$

$$\left| \int g_t(x) d\nu^*(x) - \int g_t(x) d\nu_t(x) \right| > Ct^2$$

A criterion for convergence

Similar to a 'mode collapse': Can construct g_t to be 'flat' on the support of ν_t, and still witnesses the difference between ν^* and ν_t



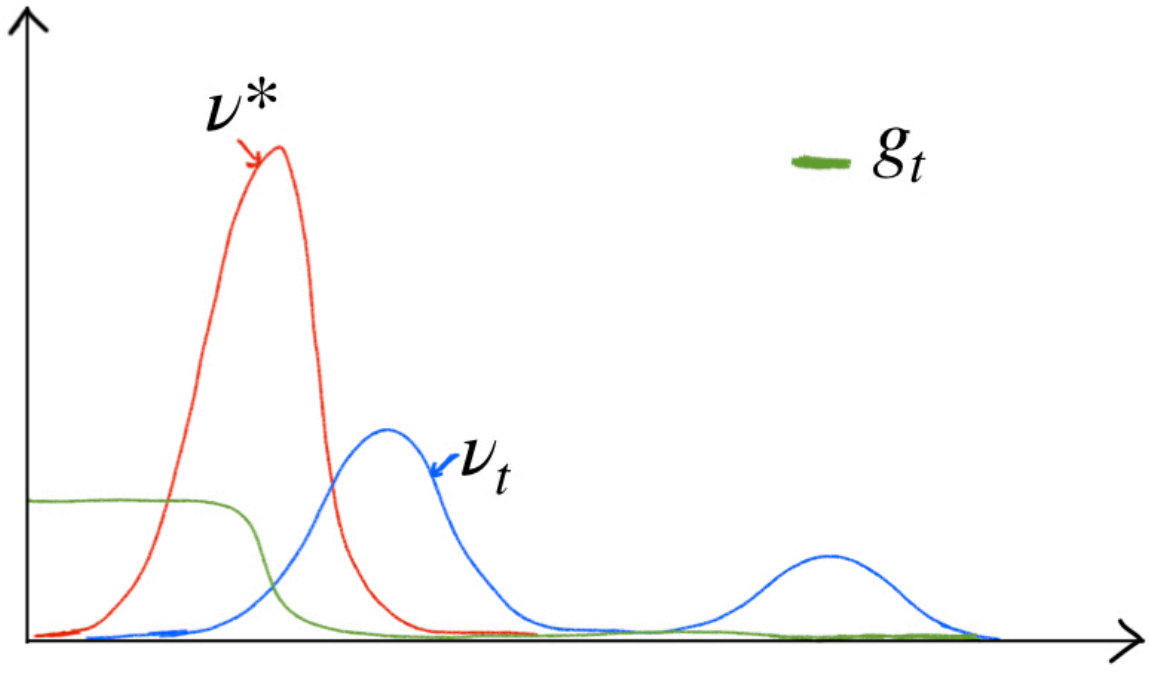
$$\int \|\nabla g_{t+1}(x)\|^2 \, d\nu_{t+1}(x) \leq 1$$

$$\left| \int g_{t+1}(x) d\nu^*(x) - \int g_{t+1}(x) d\nu_{t+1}(x) \right| > C(t+1)^2$$

A criterion for convergence: Negative Sobolev distance

Define:

$S(\nu^*|\nu_t) = \sup_{g,\int \|\nabla g(x)\|^2 \mathrm{d}\nu_t(x)\leq 1} |\int g(x)\,\mathrm{d}\nu^*(x) - \int g(x)\,\mathrm{d}\nu_t(x)|$



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

$$S(\nu^* \,|\, \nu_t) \geq |\int g_t(x) d\,\nu^*(x) - \int g_t(x) d\,\nu_t(x)\,| > Ct^2$$

A criterion for convergence: Negative Sobolev distance

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

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

A criterion for convergence: Negative Sobolev distance

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

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

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

Noise Injection

- Failure case when $\nabla f_t(x) \simeq 0$ on the support of ν_t but $\|\nabla f_t(x)\| \gg 1$ outside of it:

$$X_{t+1} = X_t - \gamma \nabla f_t(X_t) \simeq X_t$$

Noise Injection

- Failure case when $\nabla f_t(x) \simeq 0$ on the support of ν_t but $\|\nabla f_t(x)\| \gg 1$ outside of it:

$$X_{t+1} = X_t - \gamma \nabla f_t(X_t) \simeq X_t$$

- Idea: New sequence of updates μ_t, Evaluate ∇f_t outside of the support of μ_t

Noise Injection

- Failure case when $\nabla f_t(x) \simeq 0$ on the support of ν_t but $\|\nabla f_t(x)\| \gg 1$ outside of it:

$$X_{t+1} = X_t - \gamma \nabla f_t(X_t) \simeq X_t$$

- Idea: New sequence of updates μ_t, Evaluate ∇f_t outside of the support of μ_t
- Sample $u_t \sim \mathcal{N}(0,1)$ and β_t is the noise level:

$$X_{t+1} = X_t - \gamma \nabla f_t(X_t + \beta_t u_t); \qquad X_t \sim \mu_t$$

Noise Injection

- Failure case when $\nabla f_t(x) \simeq 0$ on the support of ν_t but $\|\nabla f_t(x)\| \gg 1$ outside of it:

$$X_{t+1} = X_t - \gamma \nabla f_t(X_t) \simeq X_t$$

- Idea: New sequence of updates μ_t, Evaluate ∇f_t outside of the support of μ_t
- Sample $u_t \sim \mathcal{N}(0,1)$ and β_t is the noise level:

$$X_{t+1} = X_t - \gamma \nabla f_t(X_t + \beta_t u_t); \qquad X_t \sim \mu_t$$

- Similar to *continuation methods* [Chaudhari et al., 2017, Hazan et al., 2015]
- But extended to interacting particles.

Noise Injection

- Failure case when $\nabla f_t(x) \simeq 0$ on the support of ν_t but $\|\nabla f_t(x)\| \gg 1$ outside of it:

$$X_{t+1} = X_t - \gamma \nabla f_t(X_t) \simeq X_t$$

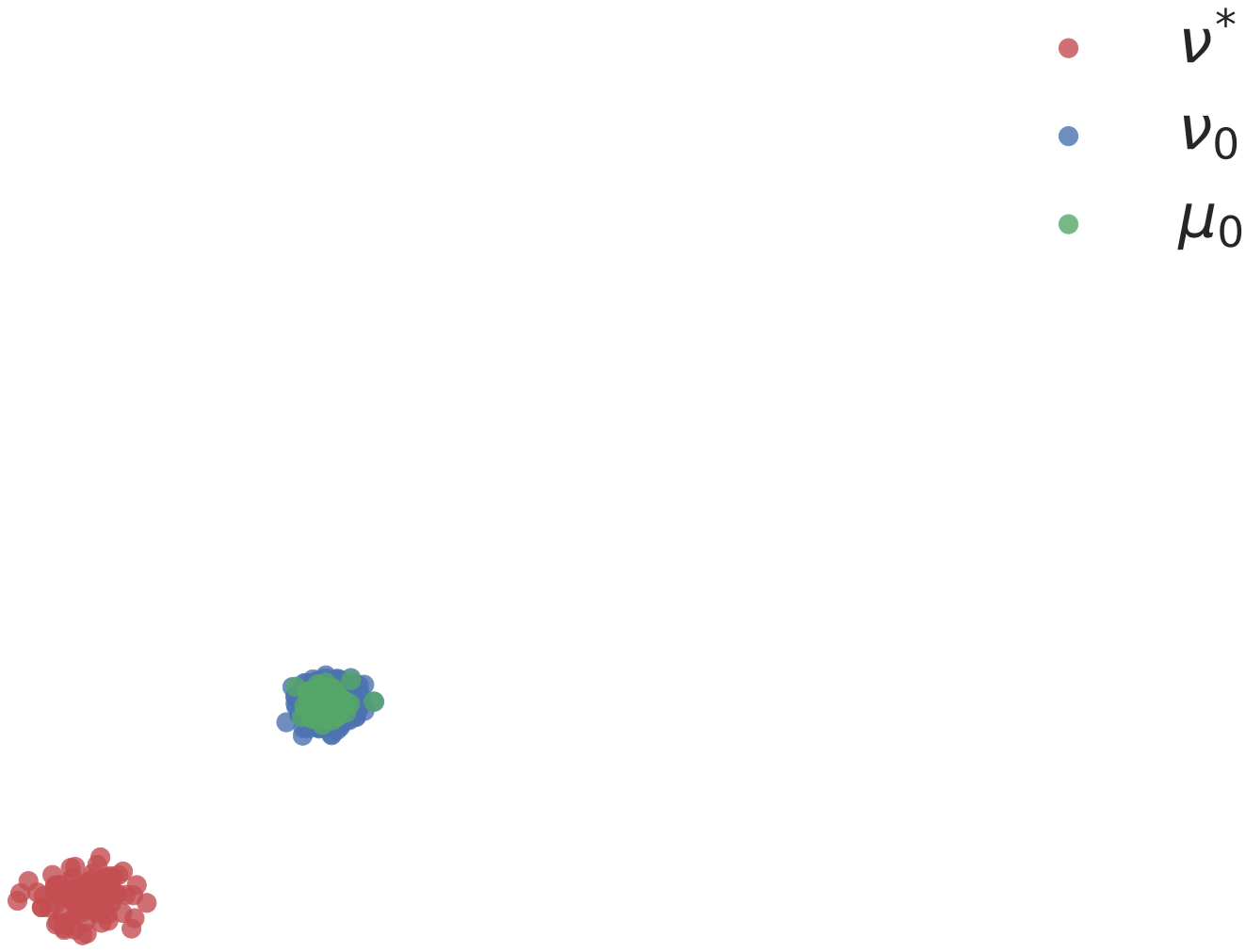
- Idea: New sequence of updates μ_t, Evaluate ∇f_t outside of the support of μ_t
- Sample $u_t \sim \mathcal{N}(0,1)$ and β_t is the noise level:

$$X_{t+1} = X_t - \gamma \nabla f_t(X_t + \beta_t u_t); \qquad X_t \sim \mu_t \tag{1}$$

- Different from entropic regularization [Mei et al., 2018]

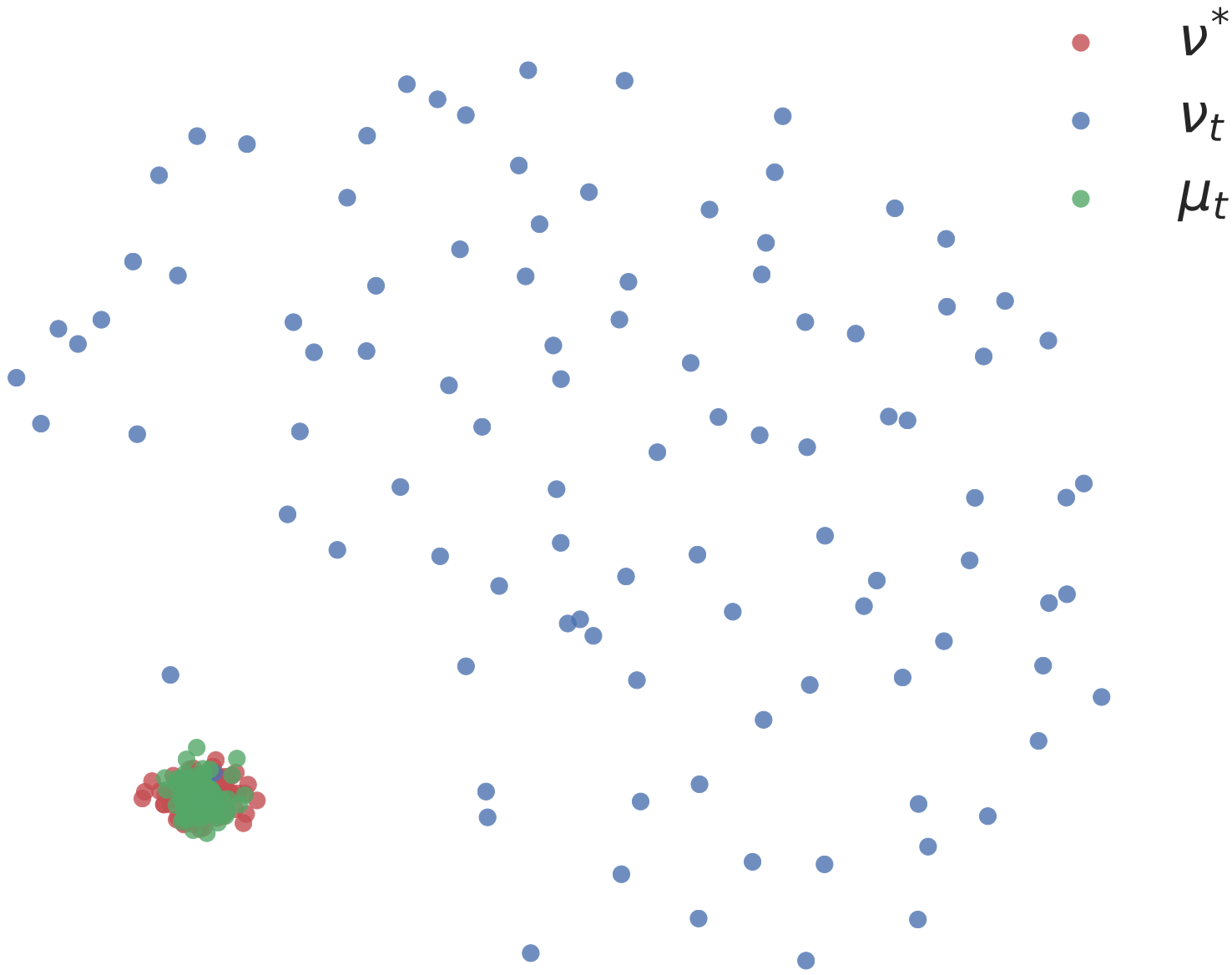
$$X_{t+1} = X_t - \gamma \nabla f_t(X_t) + \beta_t u_t \tag{2}$$

Noise Injection: Experiments



Noise Injection: Experiments

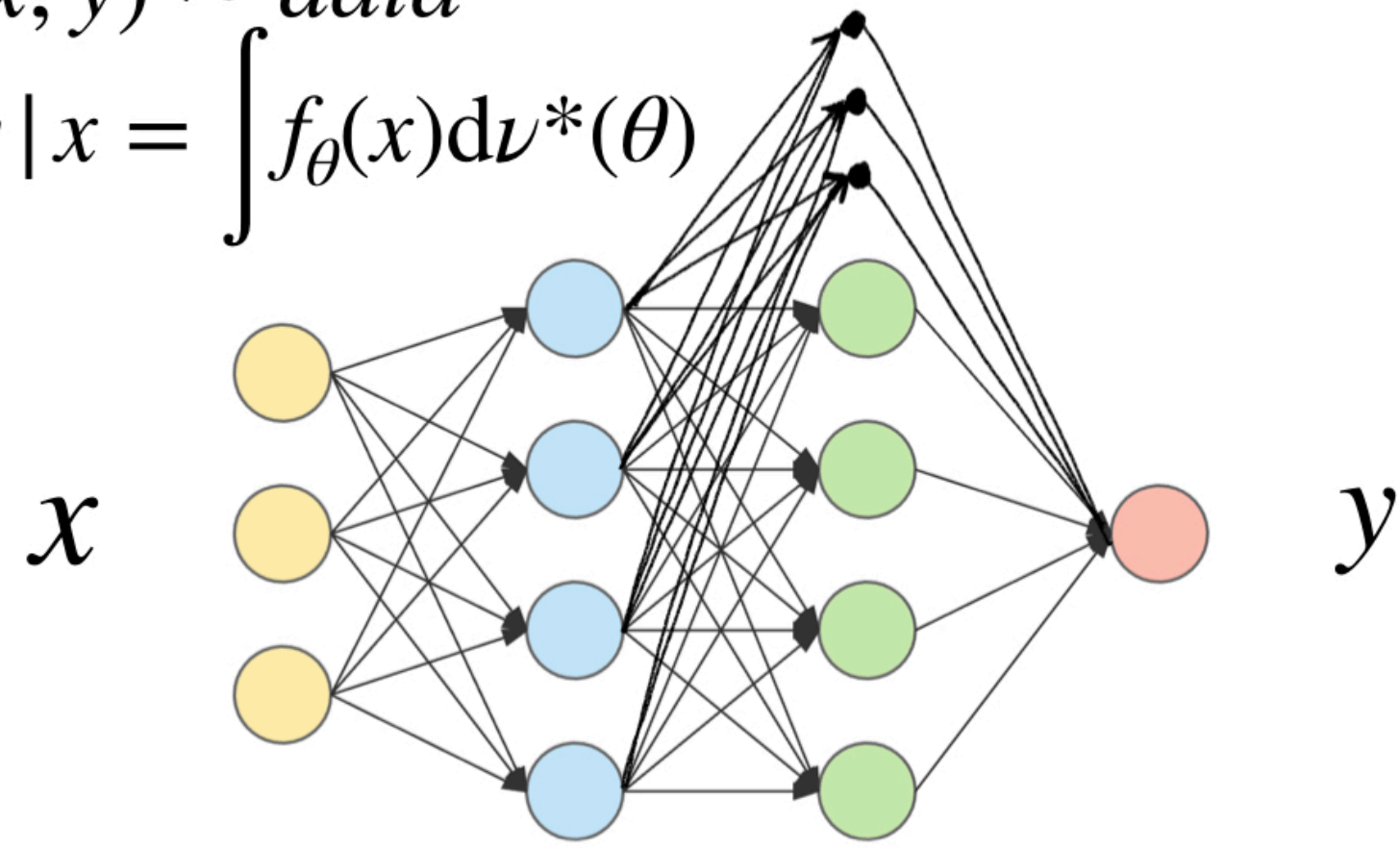
It works in practice:



Noise Injection: Experiments

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

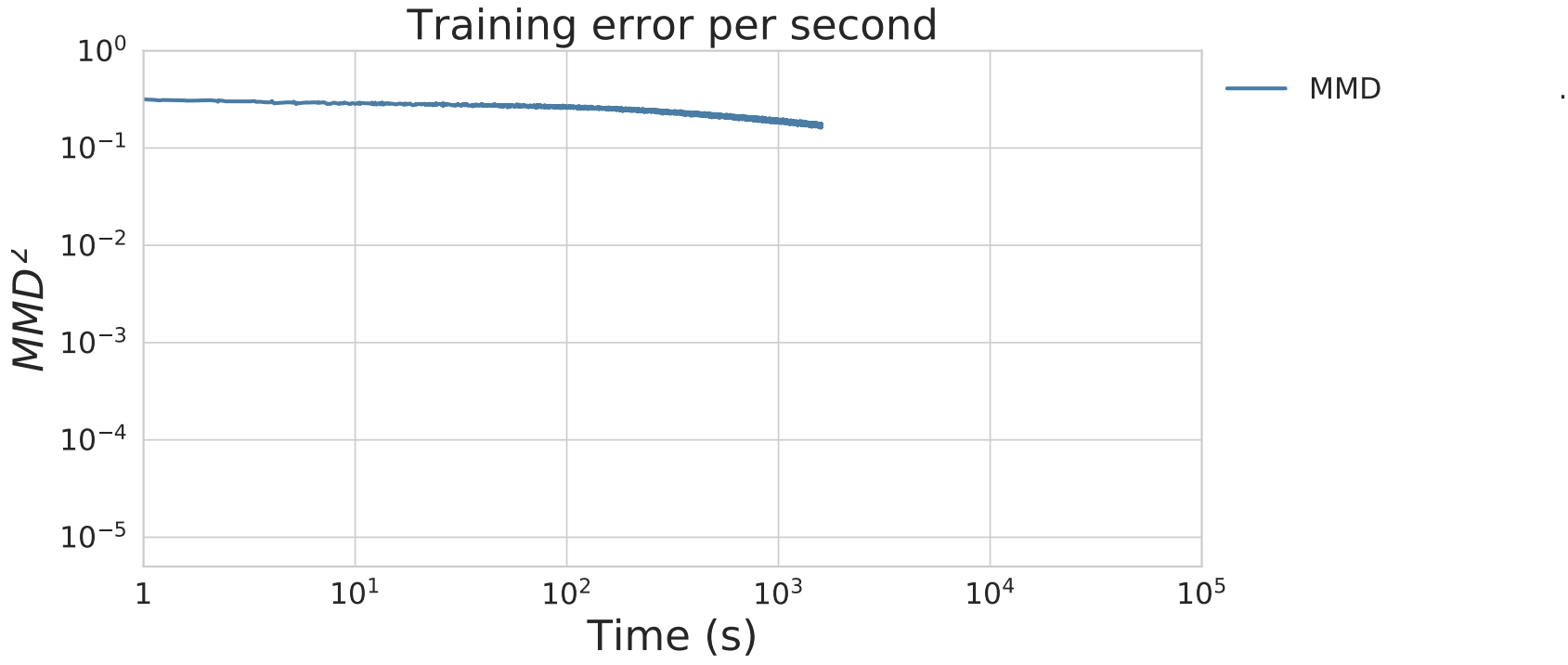
$$y \mid x = \int f_\theta(x) \mathrm{d}\nu^*(\theta)$$



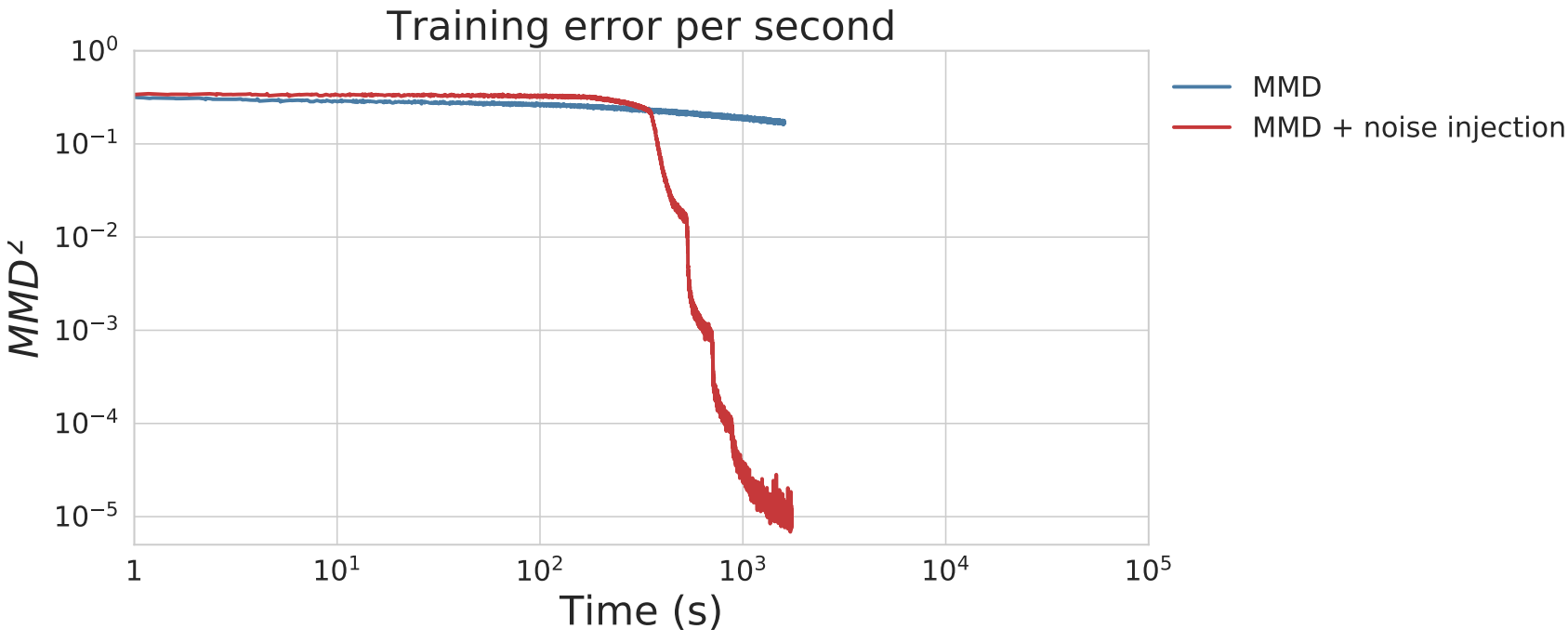
$$\min_{\nu \in \mathcal{P}} MMD^2(\nu^*, \nu)$$

$$k(\theta, \theta') = \mathbb{E}_{data}[f_\theta(x) f_{\theta'}(x)]$$

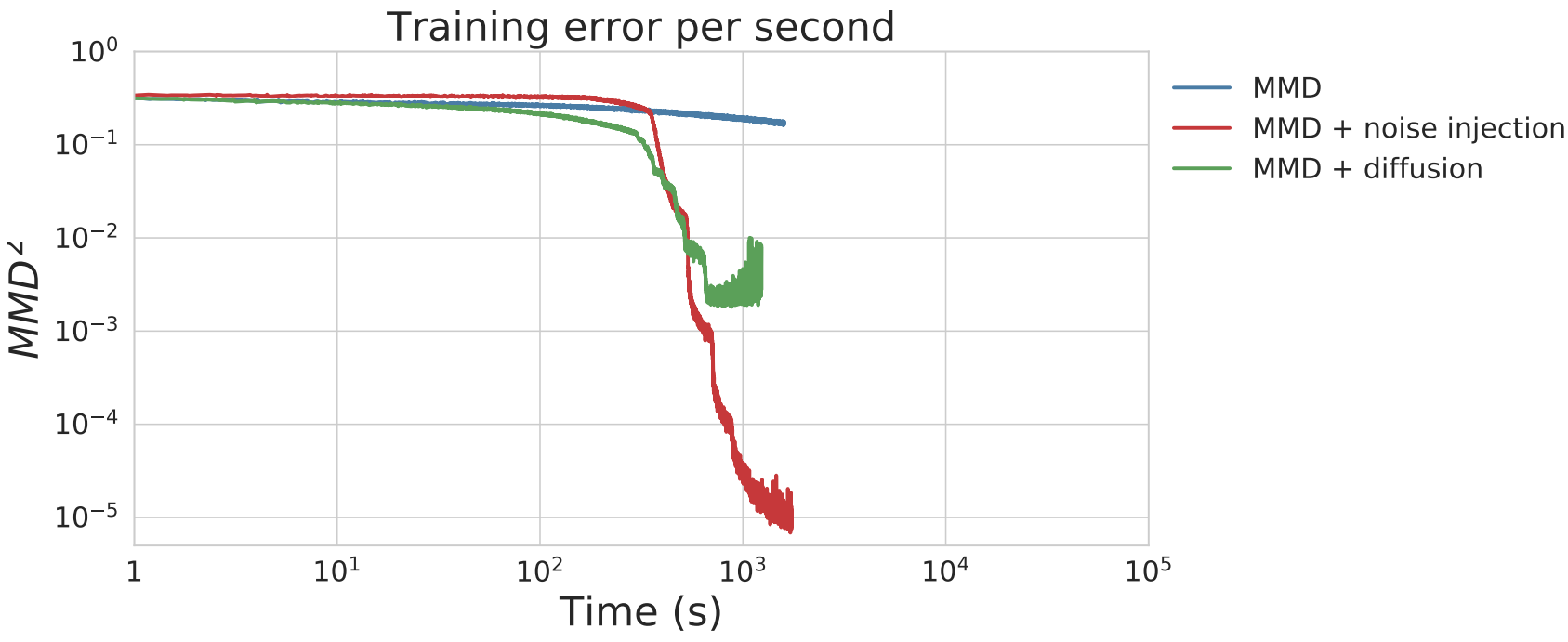
Noise Injection: Experiments



Noise Injection: Experiments

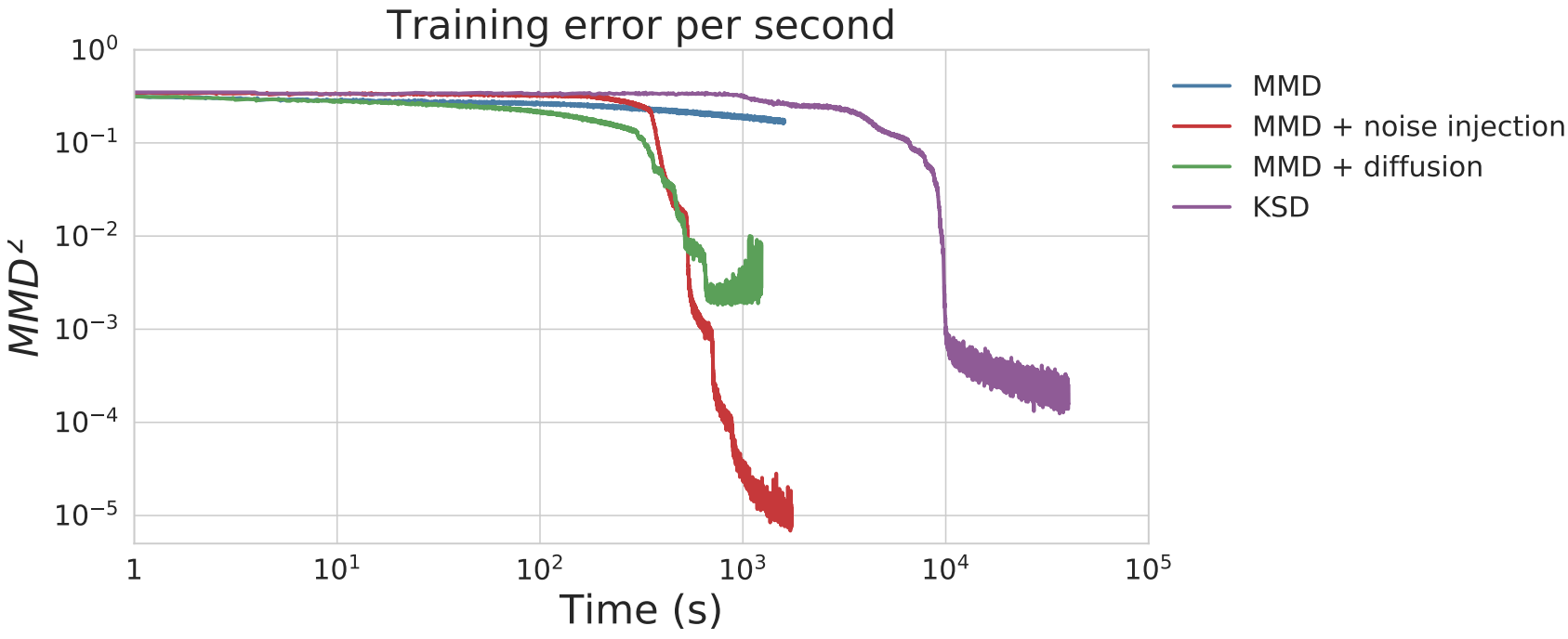


Noise Injection: Experiments



MMD + diffusion: [Mei et al., 2018]

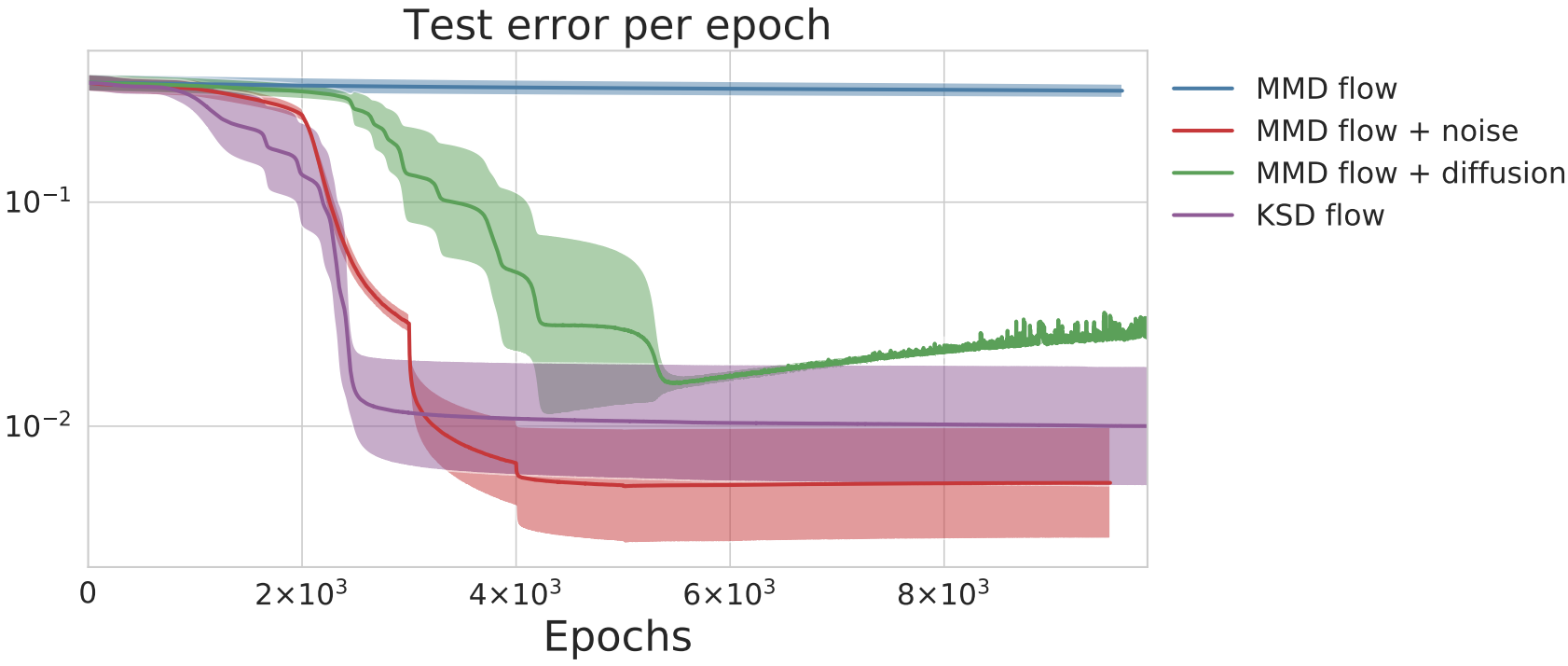
Noise Injection: Experiments



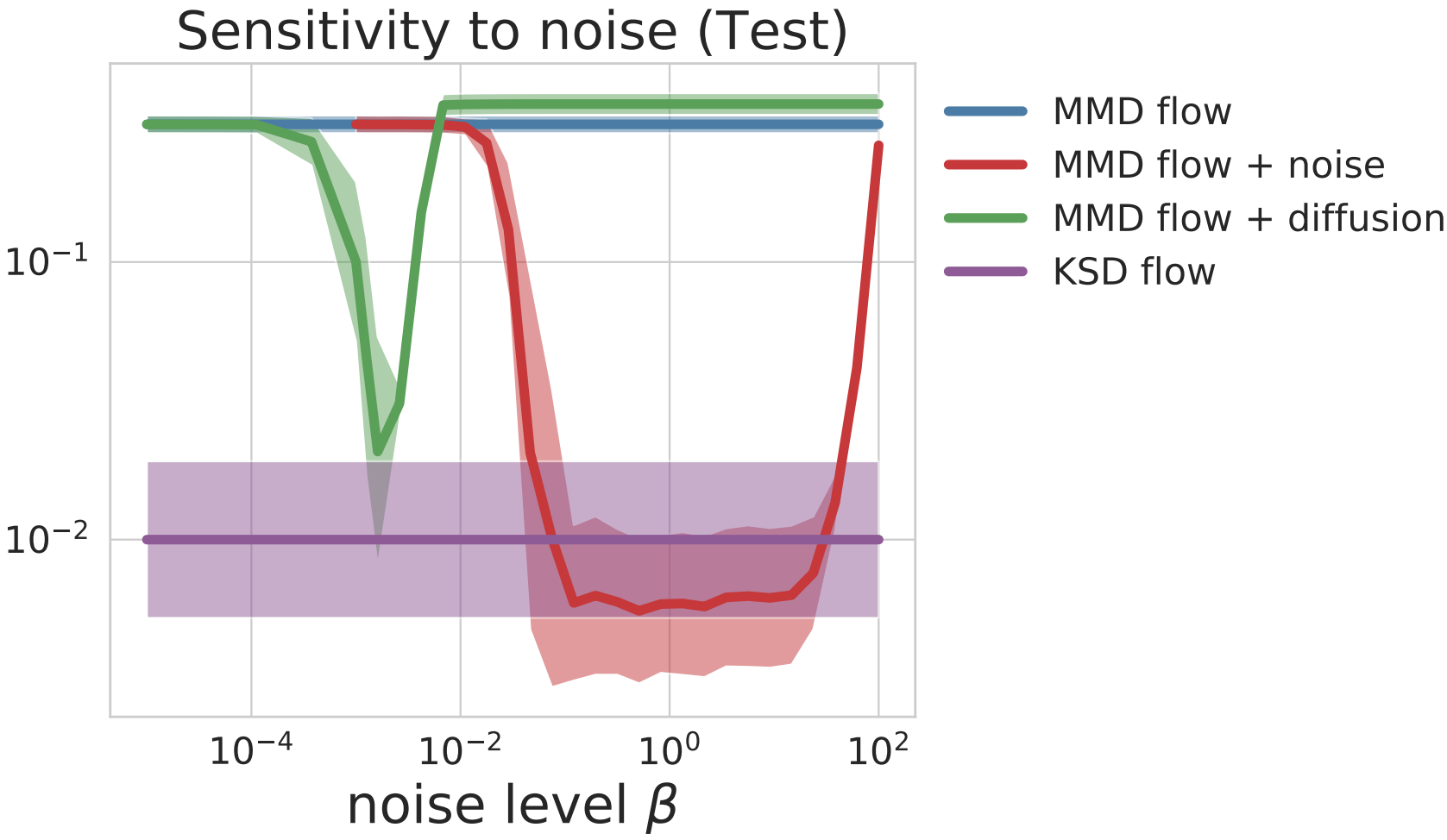
MMD + diffusion: [Mei et al., 2018]

KSD: [Mroueh, 2018]

Noise Injection: Experiments



Noise Injection: Experiments



Conclusion

Contributions:

- Analyzed a simple algorithm to optimize the MMD based on interacting particles.
- Provided a convergence criterion for the algorithm.
- Proposed an extension to the noise injection algorithm for interacting particles.

Future work:

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

Take home message

When you are stuck, add noise!!

Thank you!

Arbel, M., Sutherland, D. J., Bińkowski, M., and Gretton, A. (2018).
On gradient regularizers for MMD GANs.
arXiv:1805.11565 [cs, stat].
arXiv: 1805.11565.

Arjovsky, M., Chintala, S., and Bottou, L. (2017).
Wasserstein gan.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018).
Demystifying MMD GANs.

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