NEW YORK UNIVERSITY

Propagation-of-Chaos in Shallow NNs beyond Logarithmic time

Margalit Glasgow
Denny Wu
Joan Bruna

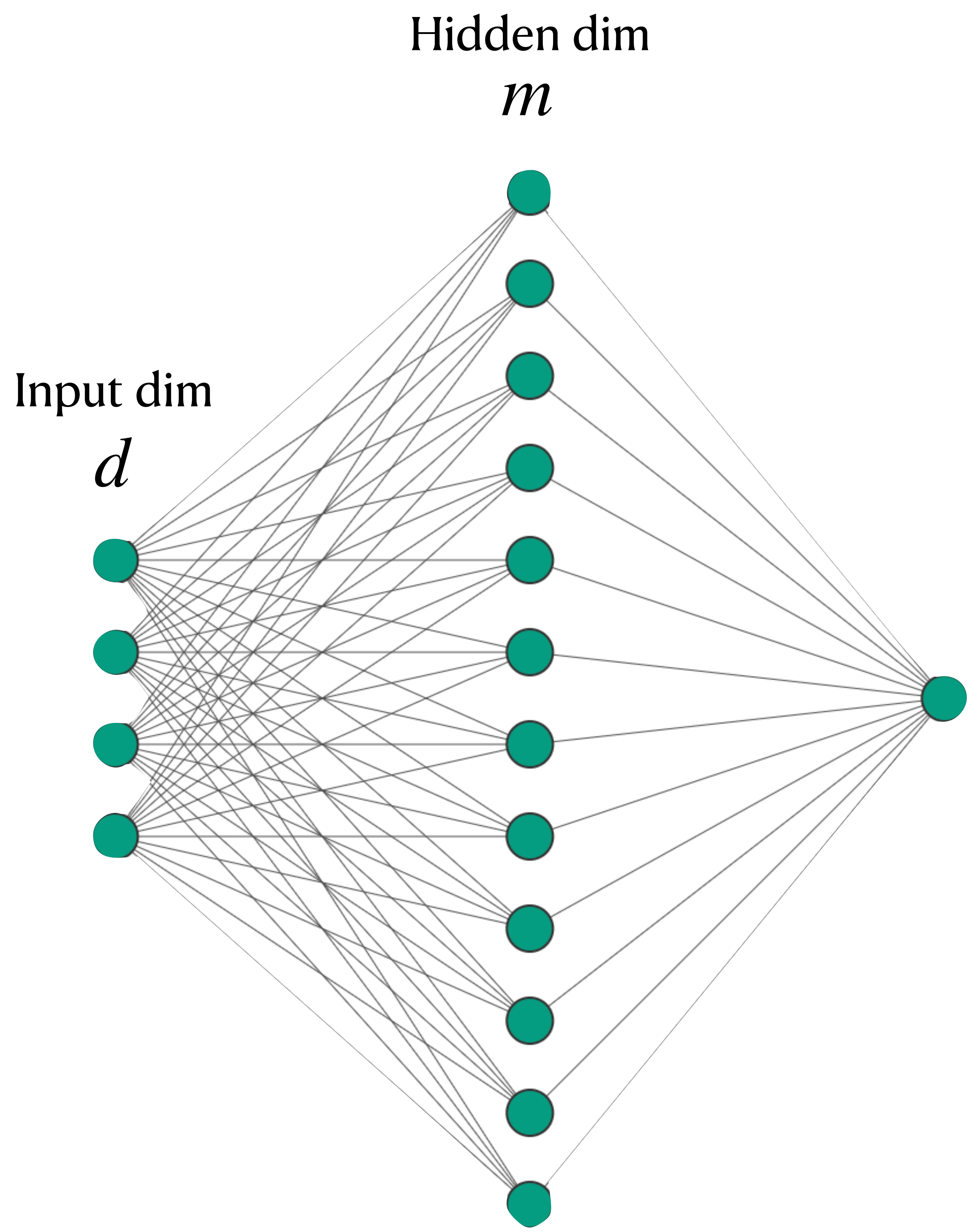
Joint Work with

Margalit Glasgow
(MIT)

Denny Wu
(NYU/Flatiron)

Overparametrized Shallow Nets

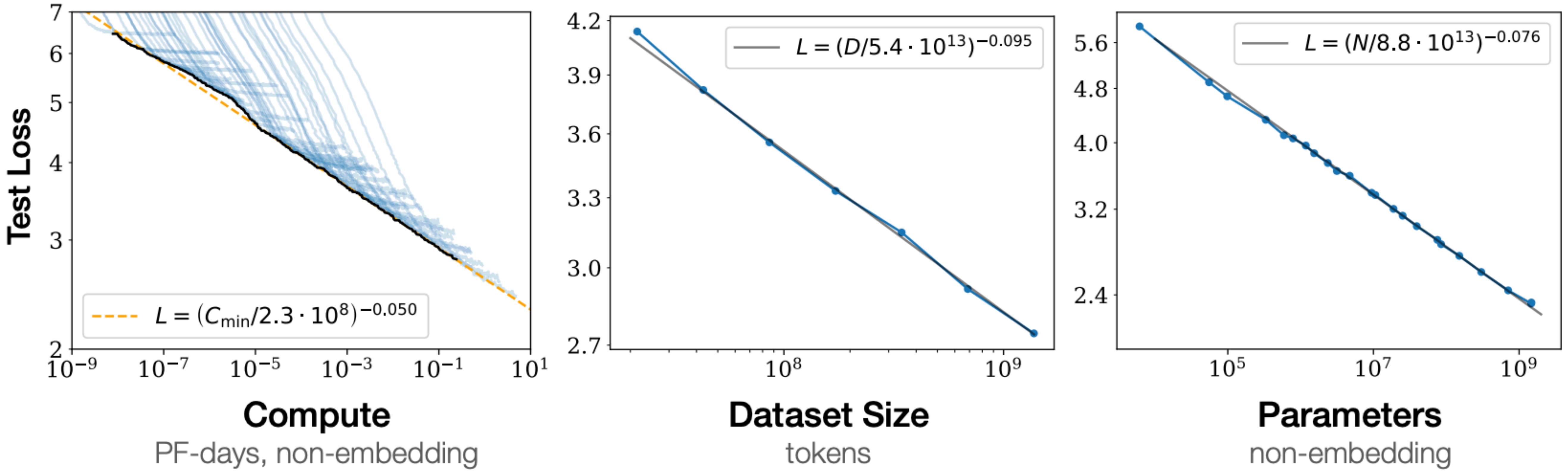
- Simplest non-linear model enabling feature learning.
- Approximation and statistical advantage over linear methods [Barron,'90s, Bach'17].



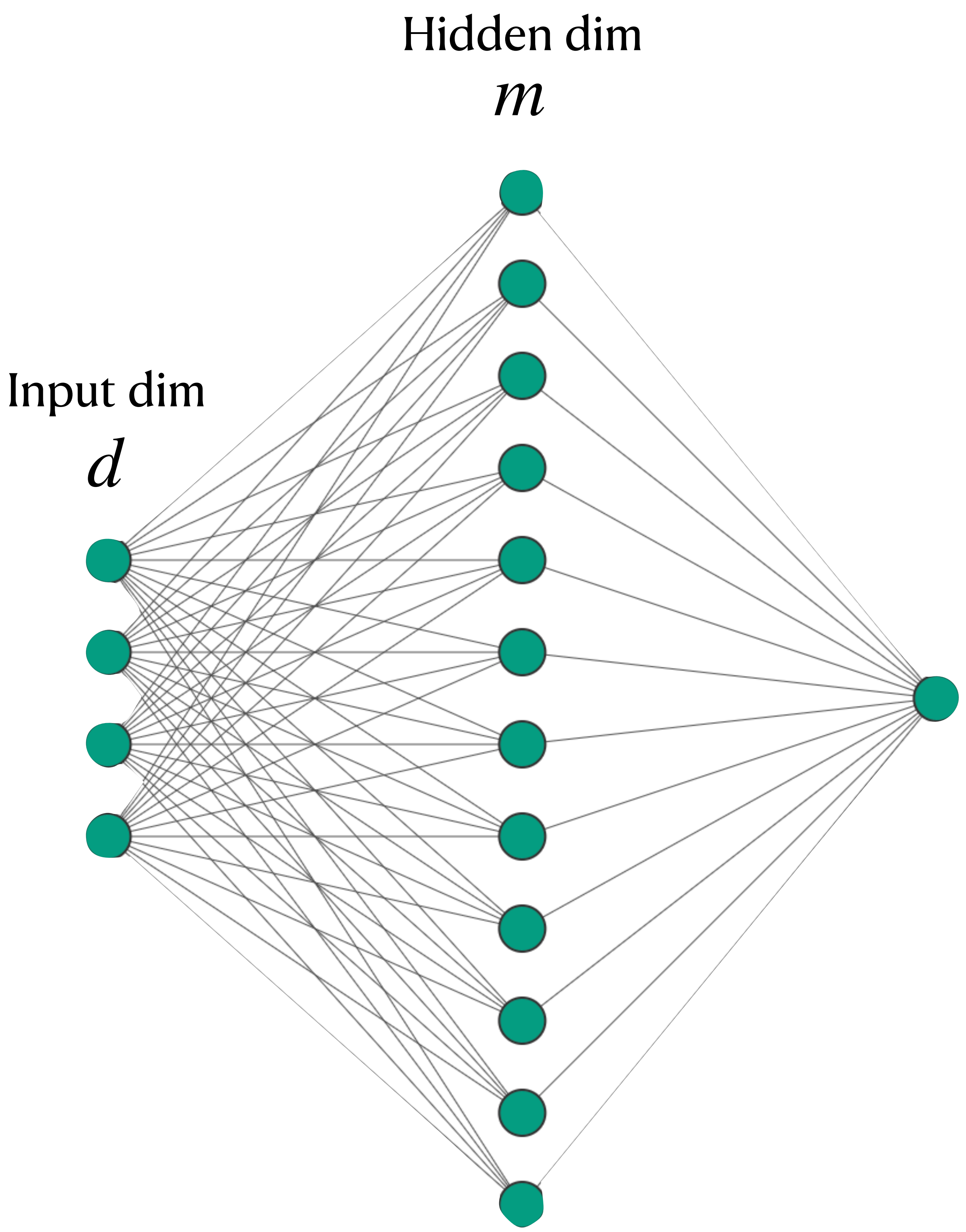
$$f(x) = \sum_{j=1}^{m} a_j \rho(\theta_j^\top x + b_j)$$

Overparametrized Shallow Nets

- Simplest non-linear model enabling feature learning.
- Approximation and statistical advantage over linear methods [Barron,'90s, Bach'17].
- *Folklore*: Wide NNs provide best learning tradeoffs in practice [Neyshabour et al, Yang, Hanin, Bartlett, many more]



[Kaplan et al]



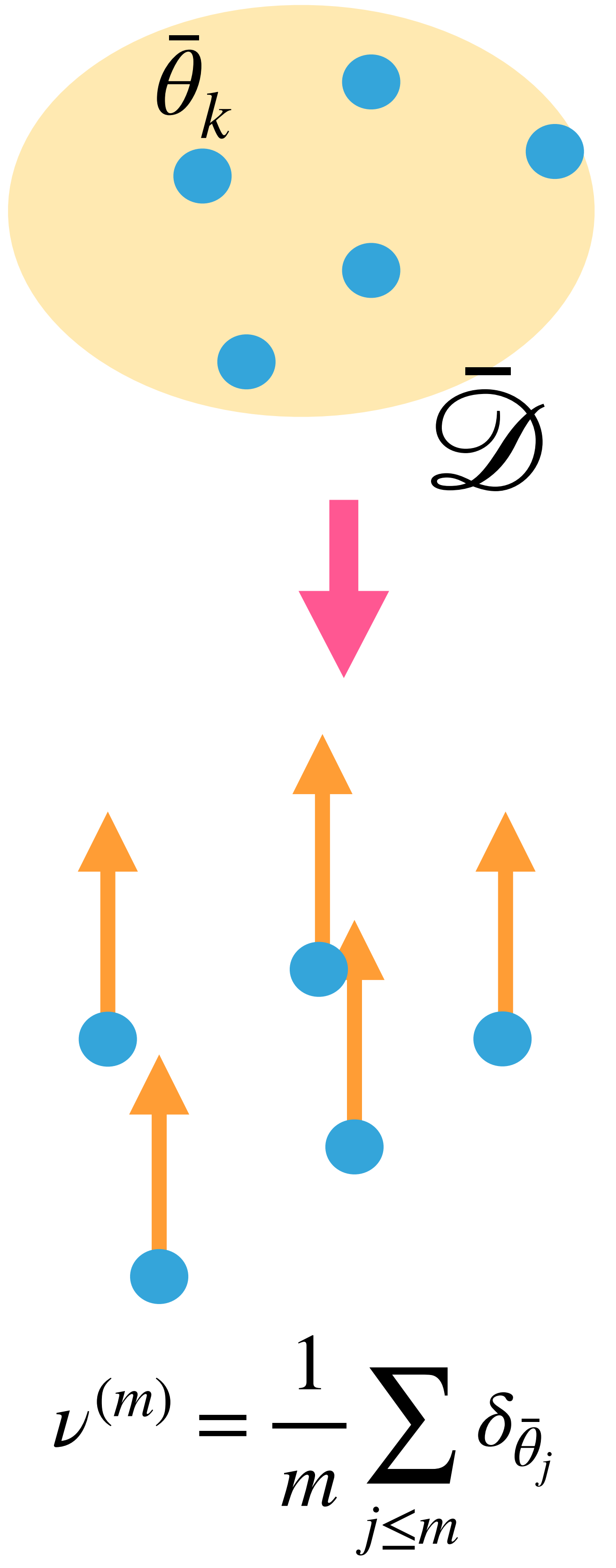
$$f(x) = \sum_{j=1}^{m} a_j \rho(\theta_j^\top x + b_j)$$

Eulerian view of Shallow NNs

[Bach, Rosset et al. Chizat et al., Nitanda et al, Mei et al, Rotskoff&EVE, Kurkova et al]

- Rewrite model $f(x) = \dfrac{1}{m}\sum_{j=1}^{m} \rho(x, \bar{\theta}_j) = \int_{\mathscr{D}} \rho(x, \bar{\theta}) d\nu^{(m)}(\bar{\theta}) := f_\nu(x)$

Squared-loss: System of **interacting** particles



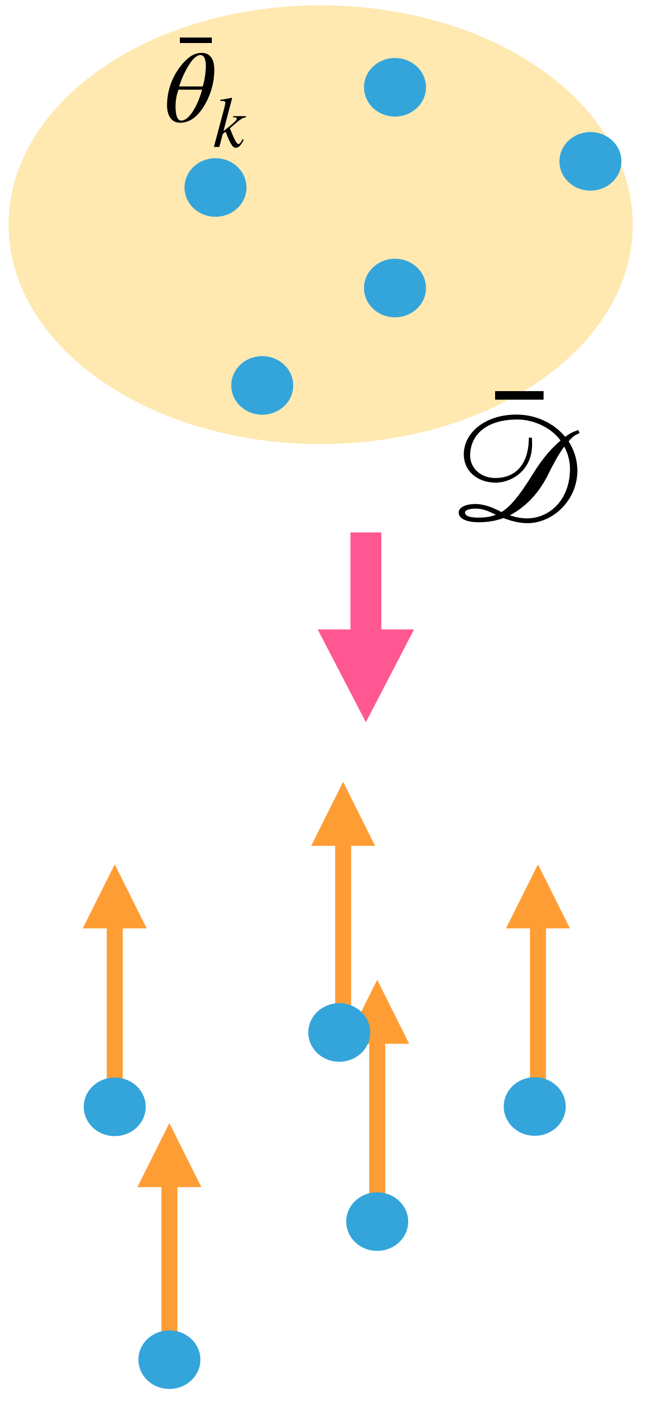
Eulerian view of Shallow NNs

[Bach, Rosset et al. Chizat et al., Nitanda et al, Mei et al, Rotskoff&EVE, Kurkova et al]

Squared-loss: System of **interacting** particles

- Rewrite model $f(x) = \frac{1}{m}\sum_{j=1}^{m} \rho(x, \bar{\theta}_j) = \int_{\mathscr{D}} \rho(x, \bar{\theta}) d\nu^{(m)}(\bar{\theta}) := f_\nu(x)$
- Regression loss becomes 'convex' in terms of ν:

$$\min_{\bar{\theta}_1, \ldots \bar{\theta}_m} L(\bar{\theta}) = \mathbb{E}\,|f(x) - y|^2 \leftrightarrow \min_{\nu} \mathbb{E}\left|\int \rho(\tilde{x}\cdot\theta)d\nu(\theta) - y\right|^2 := \mathscr{L}(\nu)\ .$$

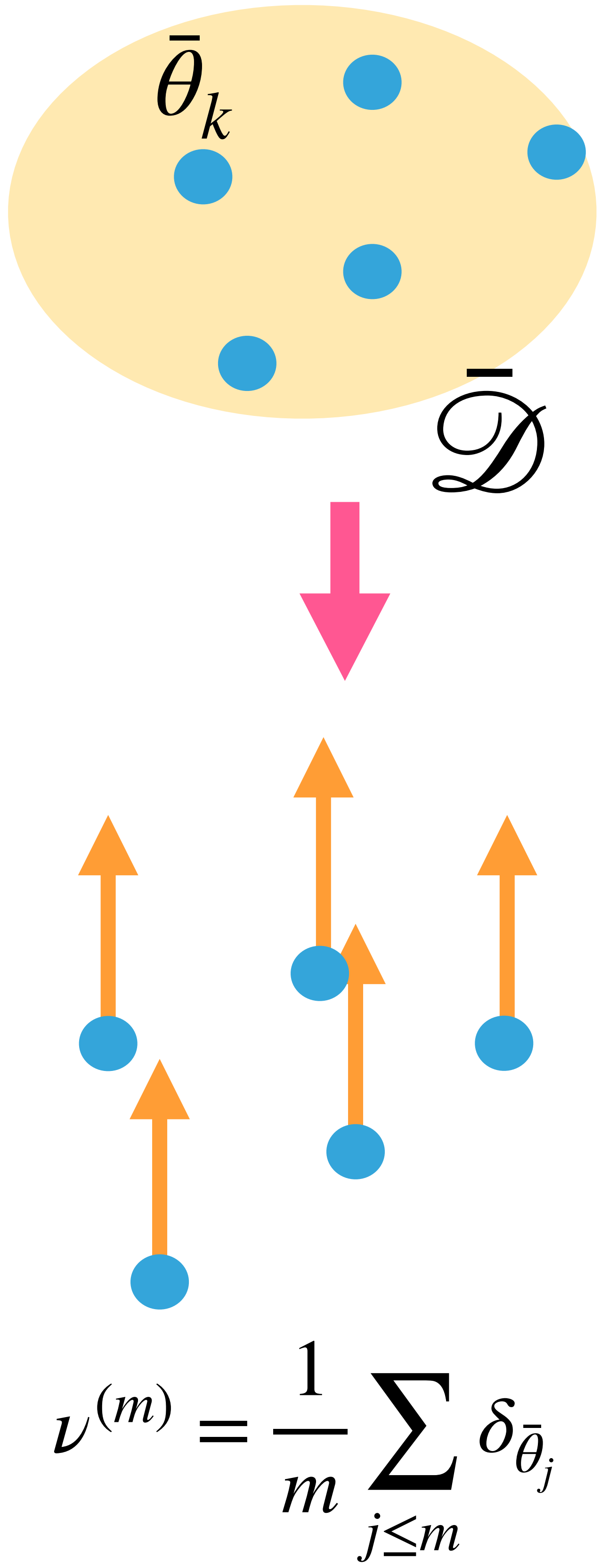


$$\nu^{(m)} = \frac{1}{m}\sum_{j\leq m} \delta_{\bar{\theta}_j}$$

Eulerian view of Shallow NNs

[Bach, Rosset et al. Chizat et al., Nitanda et al, Mei et al, Rotskoff&EVE, Kurkova et al]

Squared-loss: System of **interacting** particles



- Rewrite model $f(x) = \frac{1}{m}\sum_{j=1}^{m}\rho(x,\bar{\theta}_j) = \int_{\mathscr{D}}\rho(x,\bar{\theta})d\nu^{(m)}(\bar{\theta}) := f_\nu(x)$
- Regression loss becomes 'convex' in terms of ν:

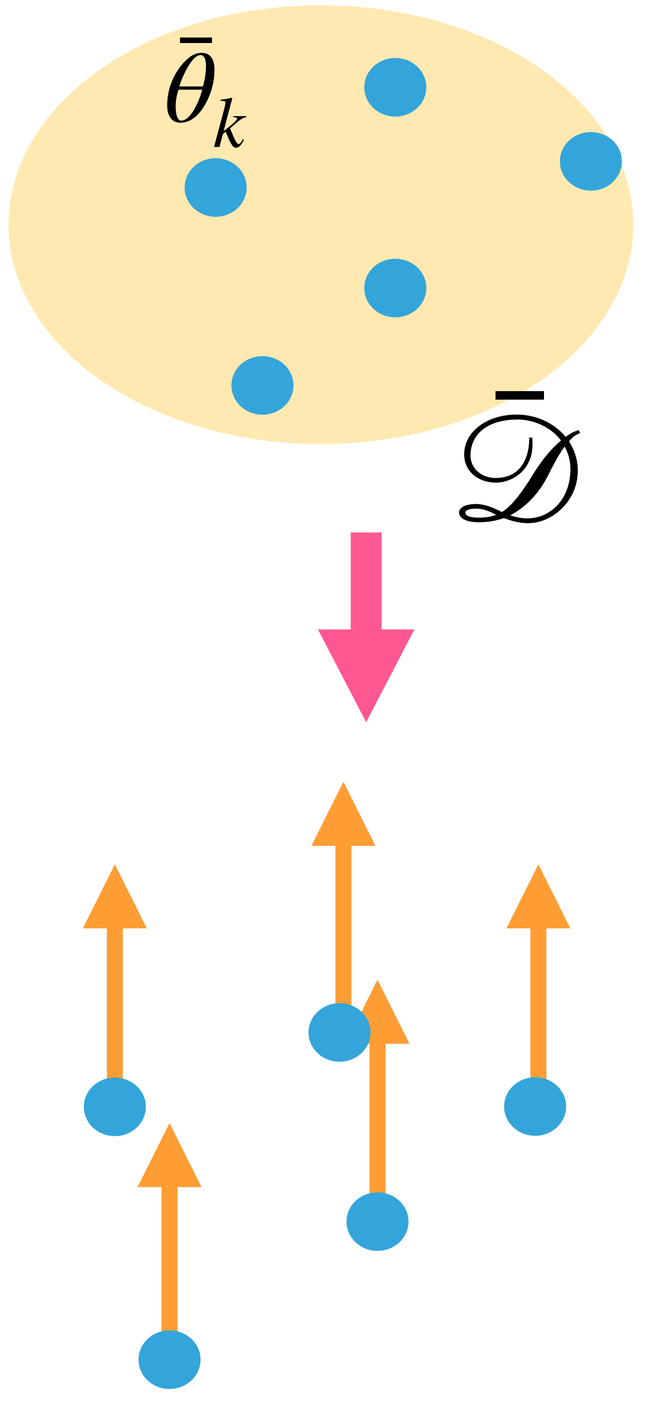
$$\min_{\bar{\theta}_1,\ldots\bar{\theta}_m} L(\bar{\theta}) = \mathbb{E}\,|f(x)-y|^2 \leftrightarrow \min_{\nu}\mathbb{E}\left|\int\rho(\tilde{x}\cdot\theta)d\nu(\theta)-y\right|^2 := \mathscr{L}(\nu)\ .$$

- Gradient Flow dynamics $\dot{\bar{\theta}}_j = -\nabla_{\bar{\theta}_j}L(\bar{\theta})$ in $\bar{\mathscr{D}}^m$ lift to a Wasserstein Gradient Flow in $\mathscr{P}(\bar{\mathscr{D}})$: $\partial_t\nu_t = \mathrm{div}\left(\nabla\frac{\delta\mathscr{L}}{\delta\nu}\nu_t\right)$ $\qquad \frac{\delta\mathscr{L}}{\delta\nu}(\theta) = U(\theta;\nu)$: instantaneous potential

Eulerian view of Shallow NNs

[Bach, Rosset et al. Chizat et al., Nitanda et al, Mei et al, Rotskoff&EVE, Kurkova et al]

Squared-loss: System of **interacting** particles



$\nu^{(m)} = \frac{1}{m}\sum_{j\le m} \delta_{\bar{\theta}_j}$

- Rewrite model $f(x) = \frac{1}{m}\sum_{j=1}^{m} \rho(x, \bar{\theta}_j) = \int_{\mathscr{D}} \rho(x, \bar{\theta}) d\nu^{(m)}(\bar{\theta}) := f_\nu(x)$
- Regression loss becomes 'convex' in terms of ν:

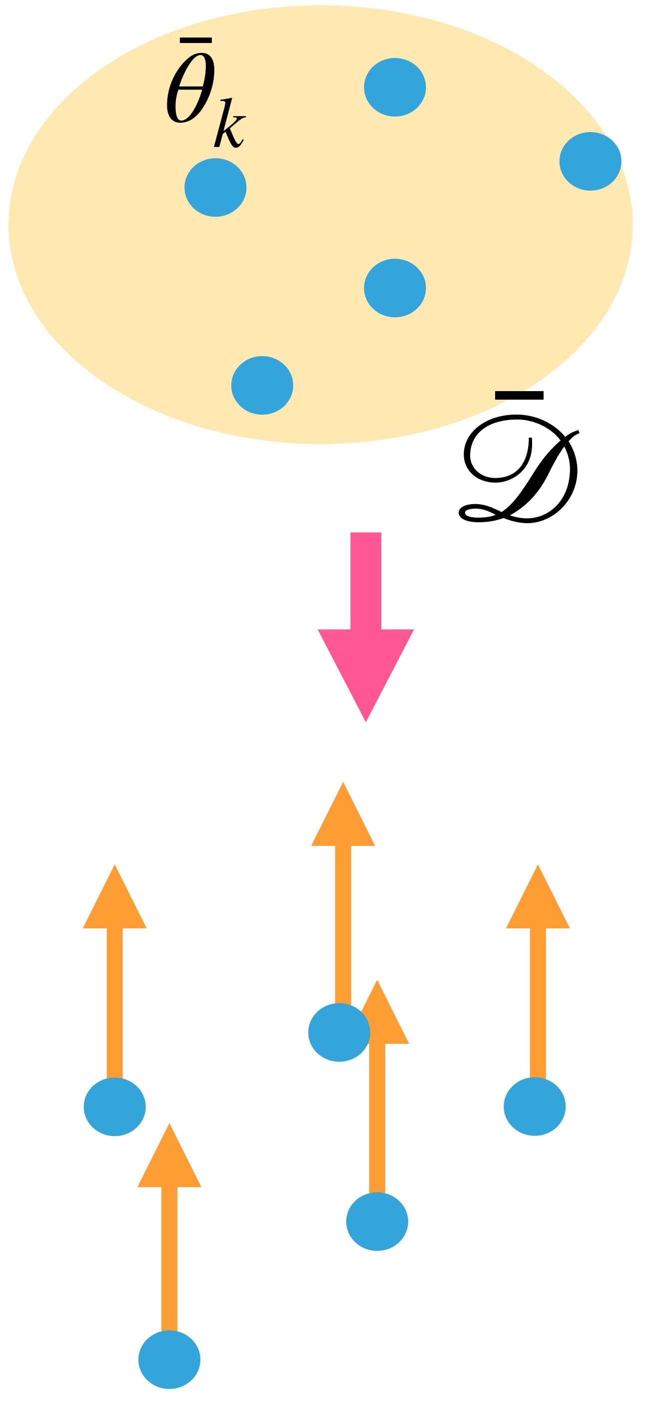
$$\min_{\bar{\theta}_1, \ldots \bar{\theta}_m} L(\bar{\theta}) = \mathbb{E}\,|f(x) - y|^2 \leftrightarrow \min_{\nu} \mathbb{E} \left| \int \rho(\tilde{x} \cdot \theta) d\nu(\theta) - y \right|^2 := \mathscr{L}(\nu) \ .$$

- Gradient Flow dynamics $\dot{\bar{\theta}}_j = -\nabla_{\bar{\theta}_j} L(\bar{\theta})$ in $\bar{\mathscr{D}}^m$ lift to a Wasserstein Gradient Flow in $\mathscr{P}(\bar{\mathscr{D}})$: $\partial_t \nu_t = \mathrm{div}\left(\nabla \frac{\delta \mathscr{L}}{\delta \nu} \nu_t\right)$ $\quad \frac{\delta \mathscr{L}}{\delta \nu}(\theta) = U(\theta; \nu)$: instantaneous potential
- Analysis of associated Wasserstein Gradient Flow: **qualitative** convergence to global minima in the thermodynamic limit $m \to \infty$.

Eulerian view of Shallow NNs

[Bach, Rosset et al. Chizat et al., Nitanda et al, Mei et al, Rotskoff&EVE, Kurkova et al]

Squared-loss: System of **interacting** particles



$\nu^{(m)} = \frac{1}{m}\sum_{j \le m} \delta_{\bar{\theta}_j}$

- Rewrite model $f(x) = \frac{1}{m}\sum_{j=1}^{m} \rho(x, \bar{\theta}_j) = \int_{\mathscr{D}} \rho(x, \bar{\theta}) d\nu^{(m)}(\bar{\theta}) := f_\nu(x)$
- Regression loss becomes 'convex' in terms of ν:

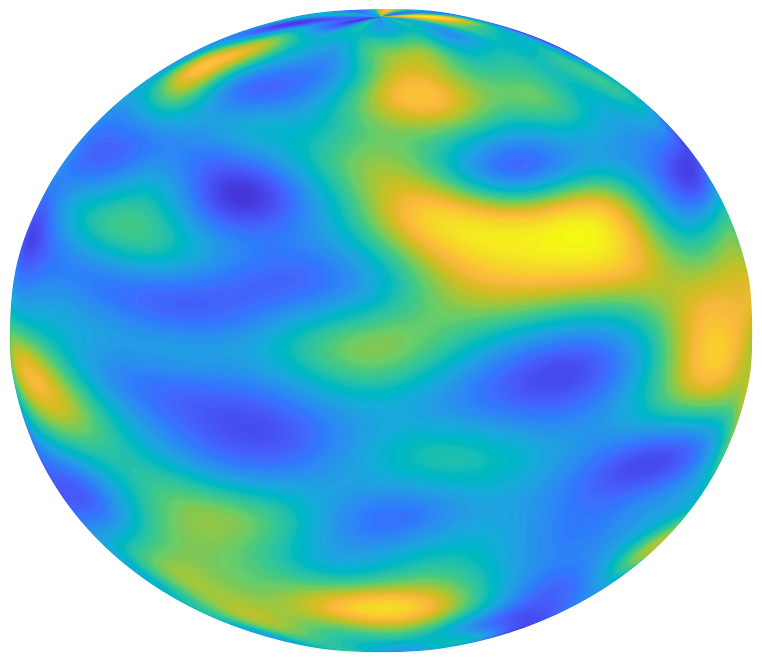
$$\min_{\bar{\theta}_1, \ldots \bar{\theta}_m} L(\bar{\theta}) = \mathbb{E}\,|f(x) - y|^2 \leftrightarrow \min_{\nu} \mathbb{E}\left|\int \rho(\tilde{x}\cdot\theta)d\nu(\theta) - y\right|^2 := \mathscr{L}(\nu)\ .$$

- Gradient Flow dynamics $\dot{\bar{\theta}}_j = -\nabla_{\bar{\theta}_j} L(\bar{\theta})$ in $\bar{\mathscr{D}}^m$ lift to a Wasserstein Gradient Flow in $\mathscr{P}(\bar{\mathscr{D}})$: $\partial_t \nu_t = \text{div}\left(\nabla \frac{\delta \mathscr{L}}{\delta \nu} \nu_t\right)$ $\quad \frac{\delta \mathscr{L}}{\delta \nu}(\theta) = U(\theta; \nu)$: instantaneous potential
- Analysis of associated Wasserstein Gradient Flow: **qualitative** convergence to global minima in the thermodynamic limit $m \to \infty$.

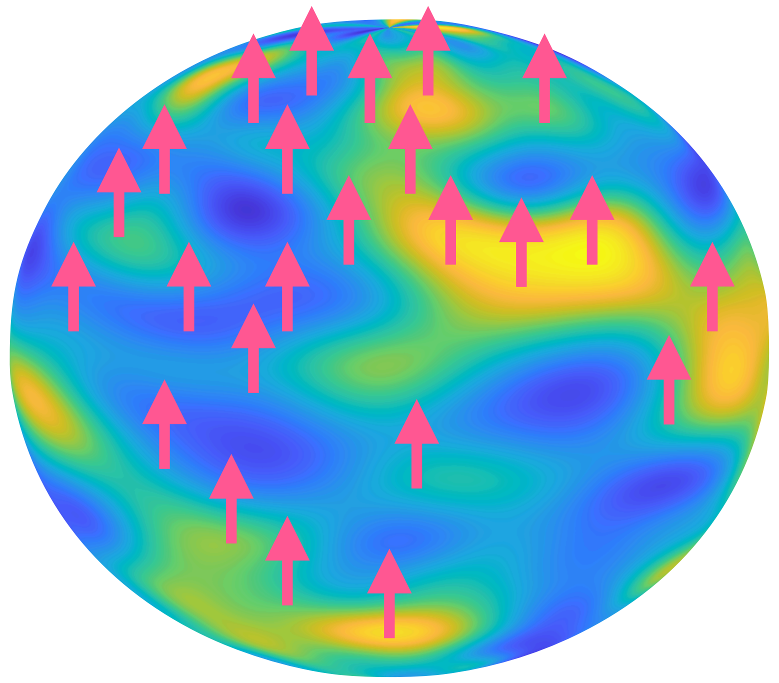
Towards quantitative (non-asymptotic) guarantees?

Quantitative Guarantees

- **Remark**: Not possible in all generality: existence of computational lower bounds under restricted algorithmic classes (SQ, LDP) [Goel et al, Diakonikolas et al.], or cryptographic assumptions [Song et al, Chen et al., Vardi et al.].



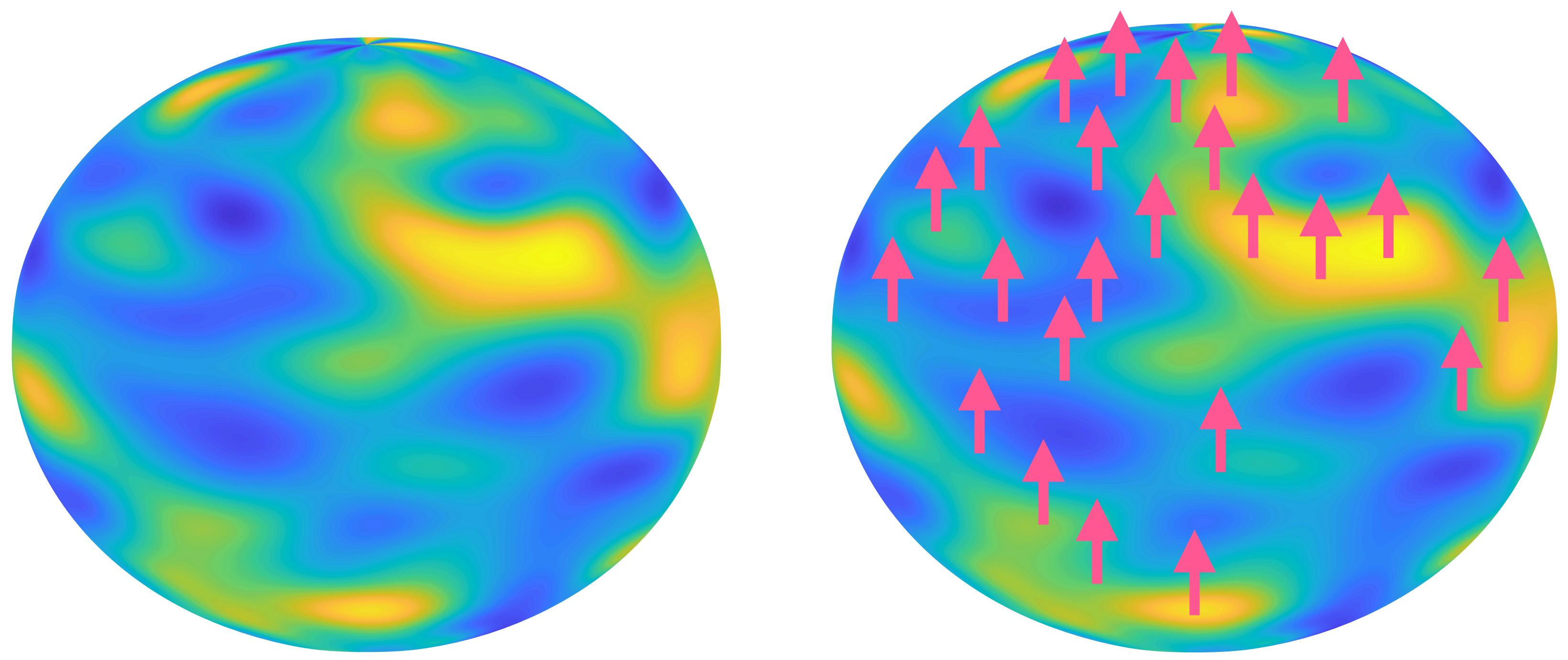
Mean-field
benign initial condition



Initial empirical
measure

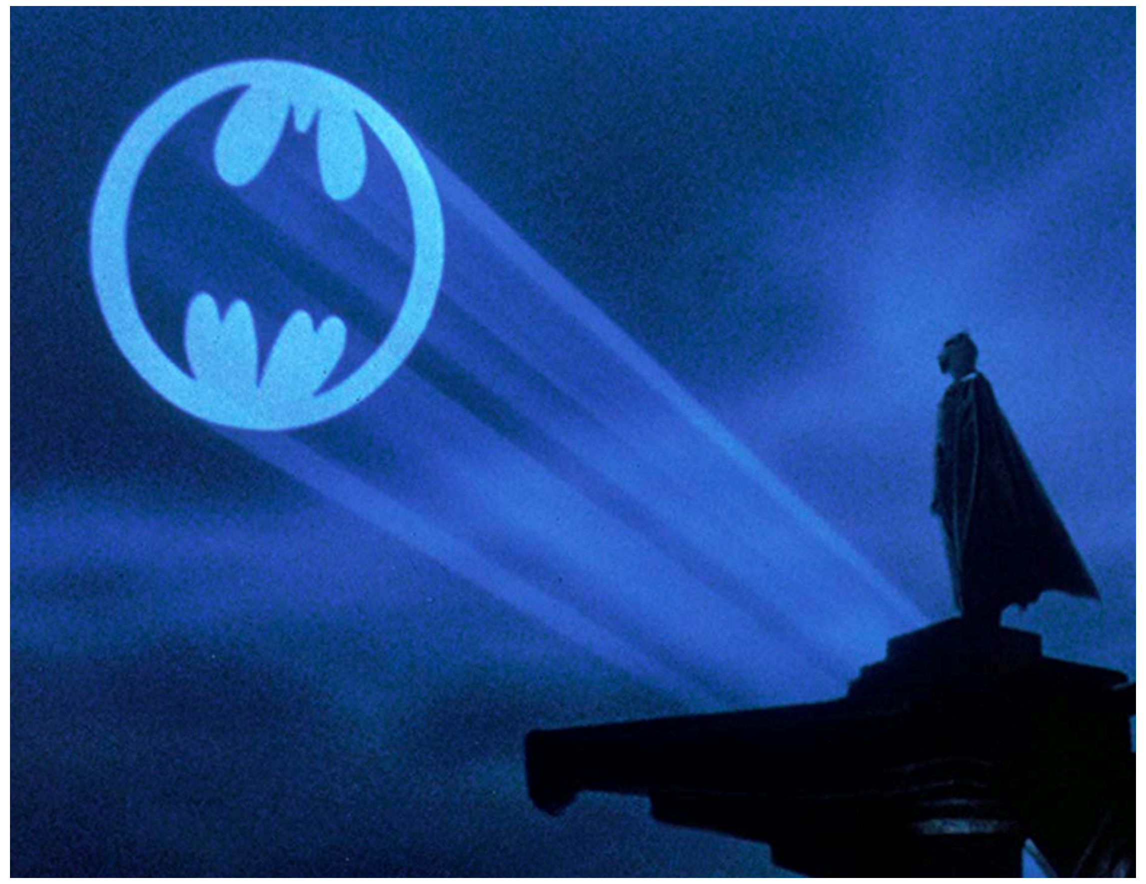
Quantitative Guarantees

- **Remark**: Not possible in all generality: existence of computational lower bounds under restricted algorithmic classes (SQ, LDP) [Goel et al, Diakonikolas et al.], or cryptographic assumptions [Song et al, Chen et al., Vardi et al.].
- **First option**: exploit structural assumptions with *dedicated* architectures / algorithms, e.g. multi-index models [Abbe et al, Dandi et al, Damian et al, Diakonikolas et al, Bietti et al, Wu et al, …] (cf Bruno's talk tomorrow)



Mean-field benign initial condition

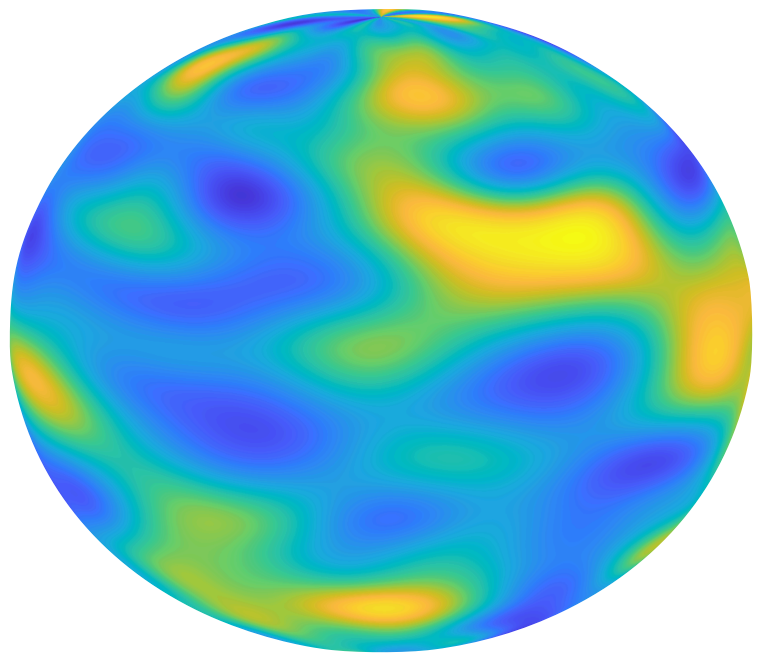
Initial empirical measure



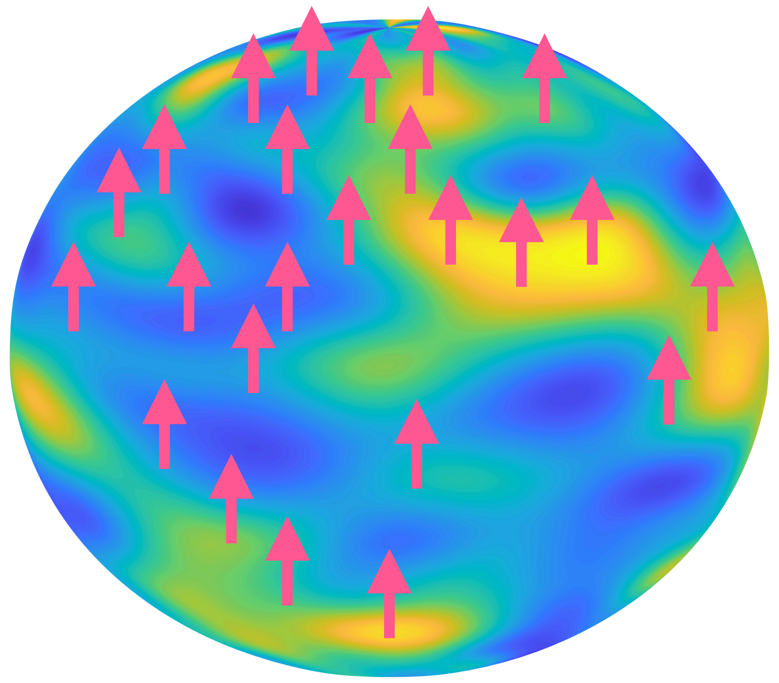
Quantitative Guarantees

- **Remark**: Not possible in all generality: existence of computational lower bounds under restricted algorithmic classes (SQ, LDP) [Goel et al, Diakonikolas et al.], or cryptographic assumptions [Song et al, Chen et al., Vardi et al.].
- **First option**: exploit structural assumptions with *dedicated* architectures / algorithms, e.g. multi-index models [Abbe et al, Dandi et al, Damian et al, Diakonikolas et al, Bietti et al, Wu et al, …] (cf Bruno's talk tomorrow)

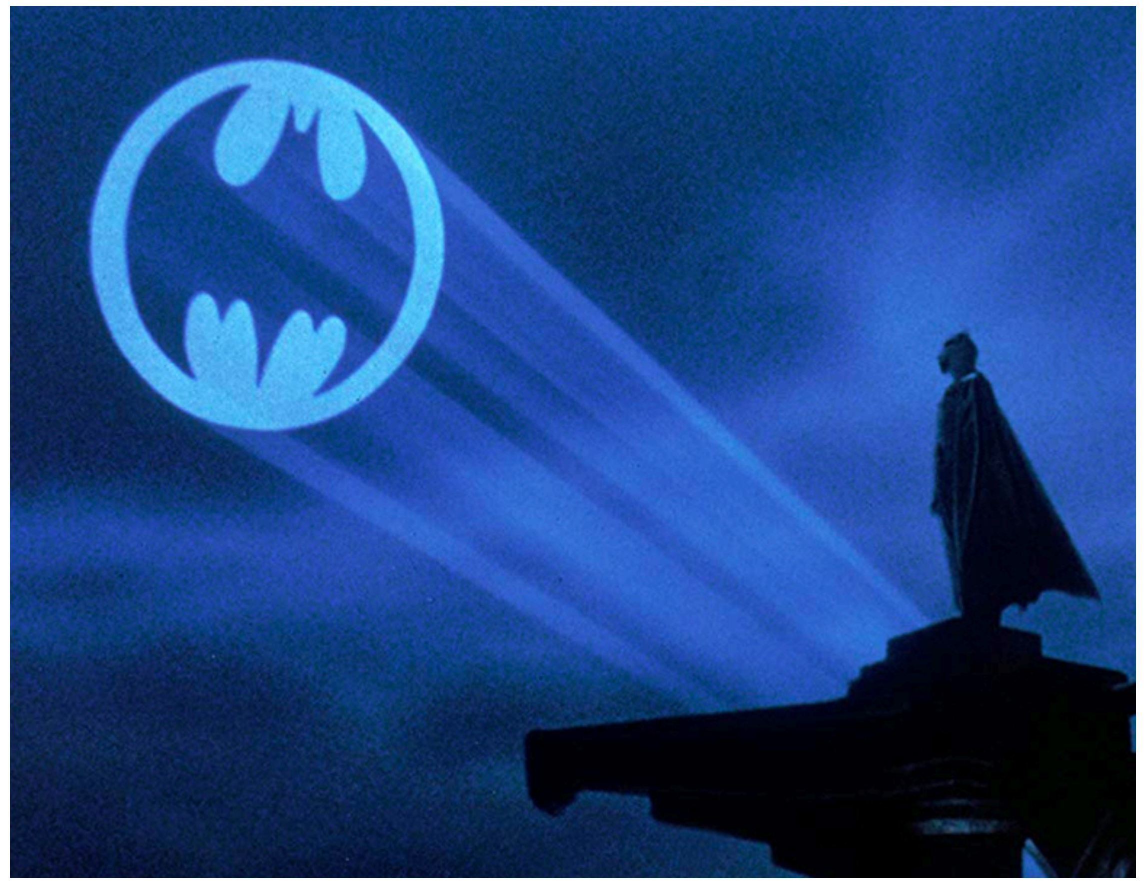
How about GD/GF on vanilla shallow NN?



Mean-field
benign initial condition



Initial empirical
measure

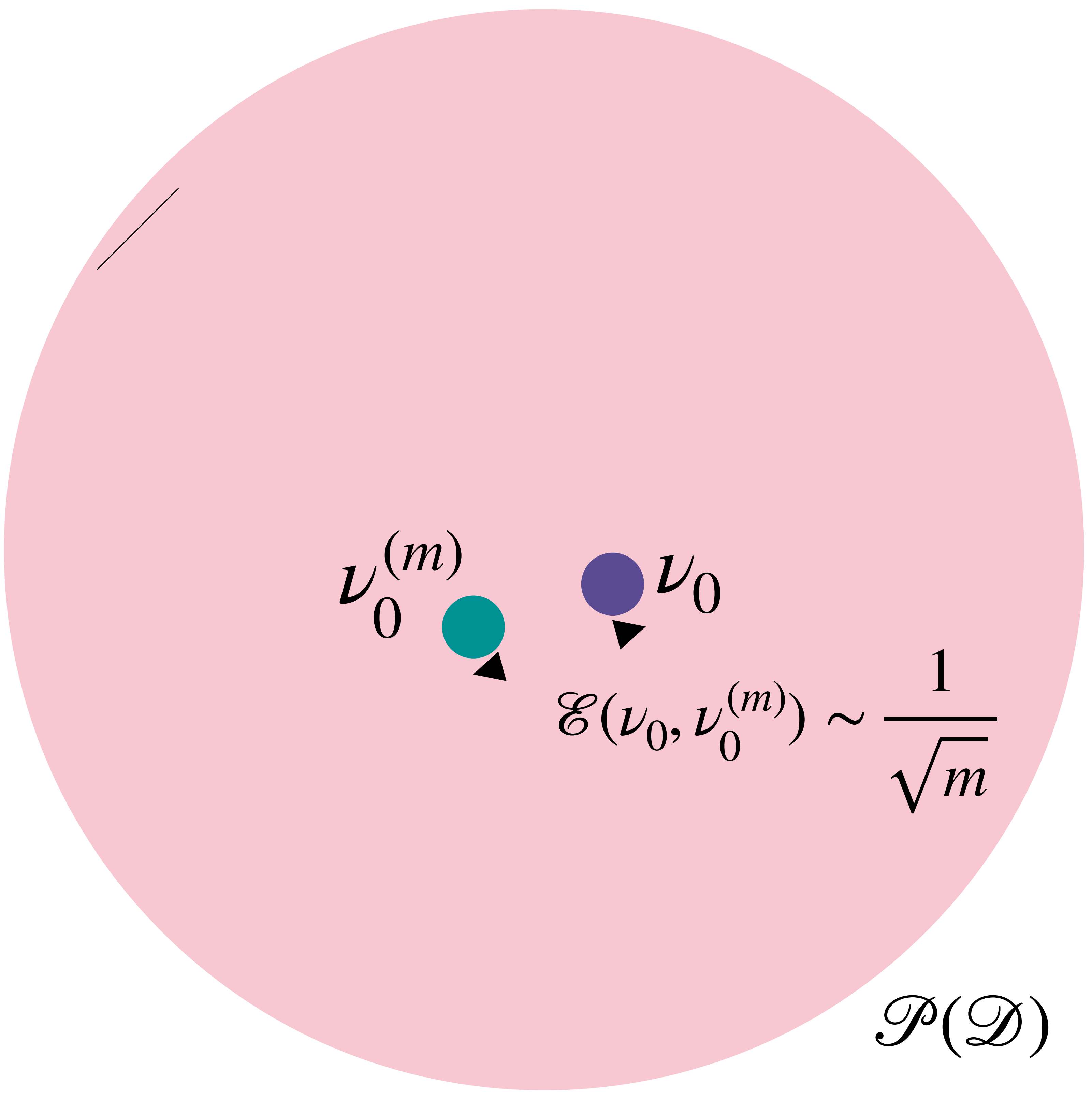


Finite-width Fluctuations

- **Static picture**: Monte-Carlo Approximation

$$\theta_1, \ldots, \theta_m \sim_{iid} \nu, f_{\nu^{(m)}}(x) = \frac{1}{m}\sum_{j \le m} \rho(\theta_j, x) \text{ satisfies}$$

$$\mathcal{E}(\nu, \nu^{(m)}) := (\mathbb{E}_x[\,|f_\nu(x) - f_{\nu^{(m)}}(x)|^2)^{1/2} \lesssim \frac{\sqrt{\mathbb{E}[\,|\rho(x,\theta)|^2]}}{\sqrt{m}}$$



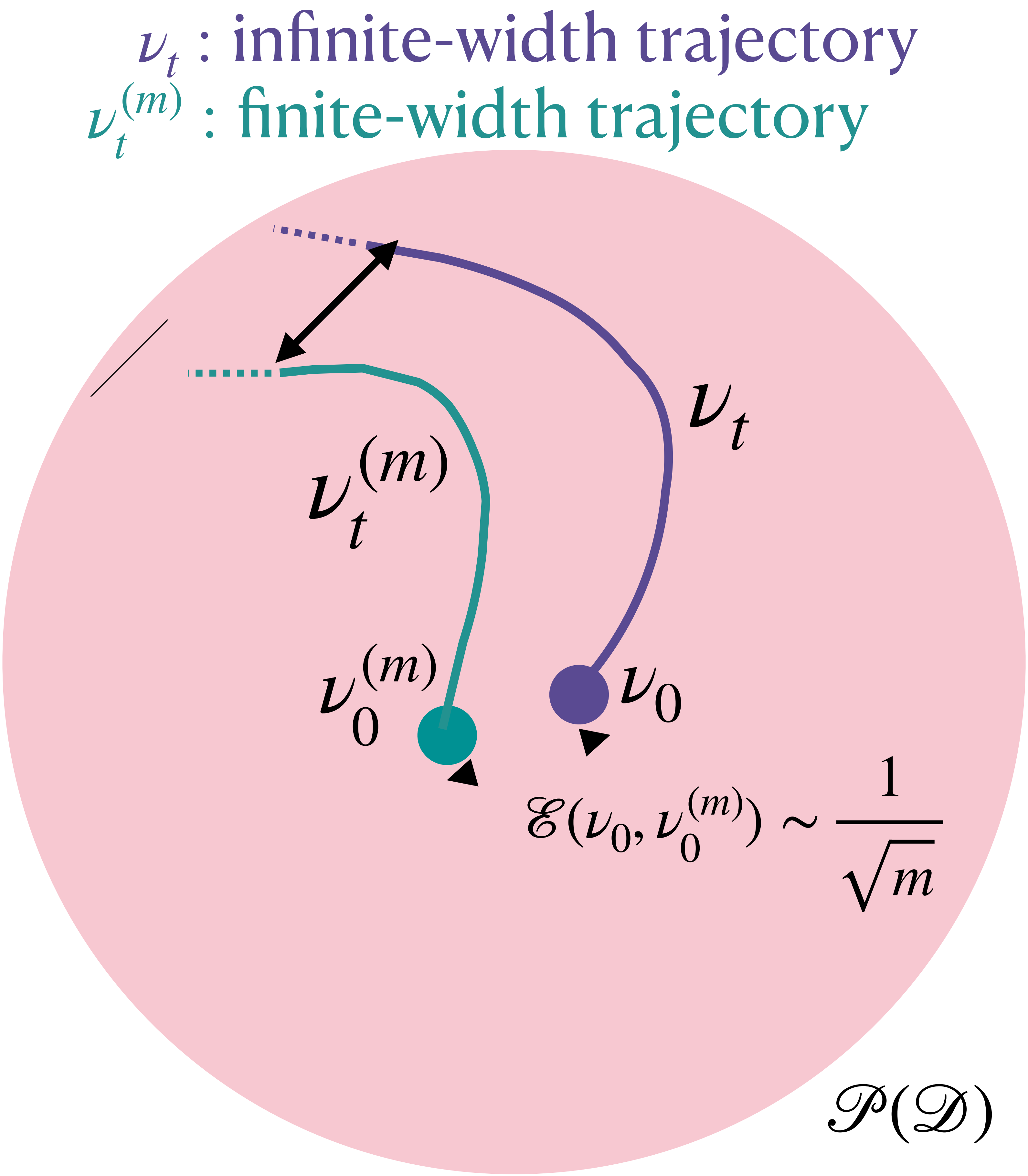
Finite-width Fluctuations

- **Static picture**: Monte-Carlo Approximation
  $\theta_1, \dots, \theta_m \sim_{iid} \nu, f_{\nu^{(m)}}(x) = \frac{1}{m} \sum_{j \le m} \rho(\theta_j, x)$ satisfies
  $$\mathcal{E}(\nu, \nu^{(m)}) := (\mathbb{E}_x[\,|f_\nu(x) - f_{\nu^{(m)}}(x)|^2)^{1/2} \lesssim \frac{\sqrt{\mathbb{E}[\,|\rho(x,\theta)|^2]}}{\sqrt{m}}$$

- **Dynamic picture**, aka *Propagation-of-Chaos* [Kac,Sznitman]:

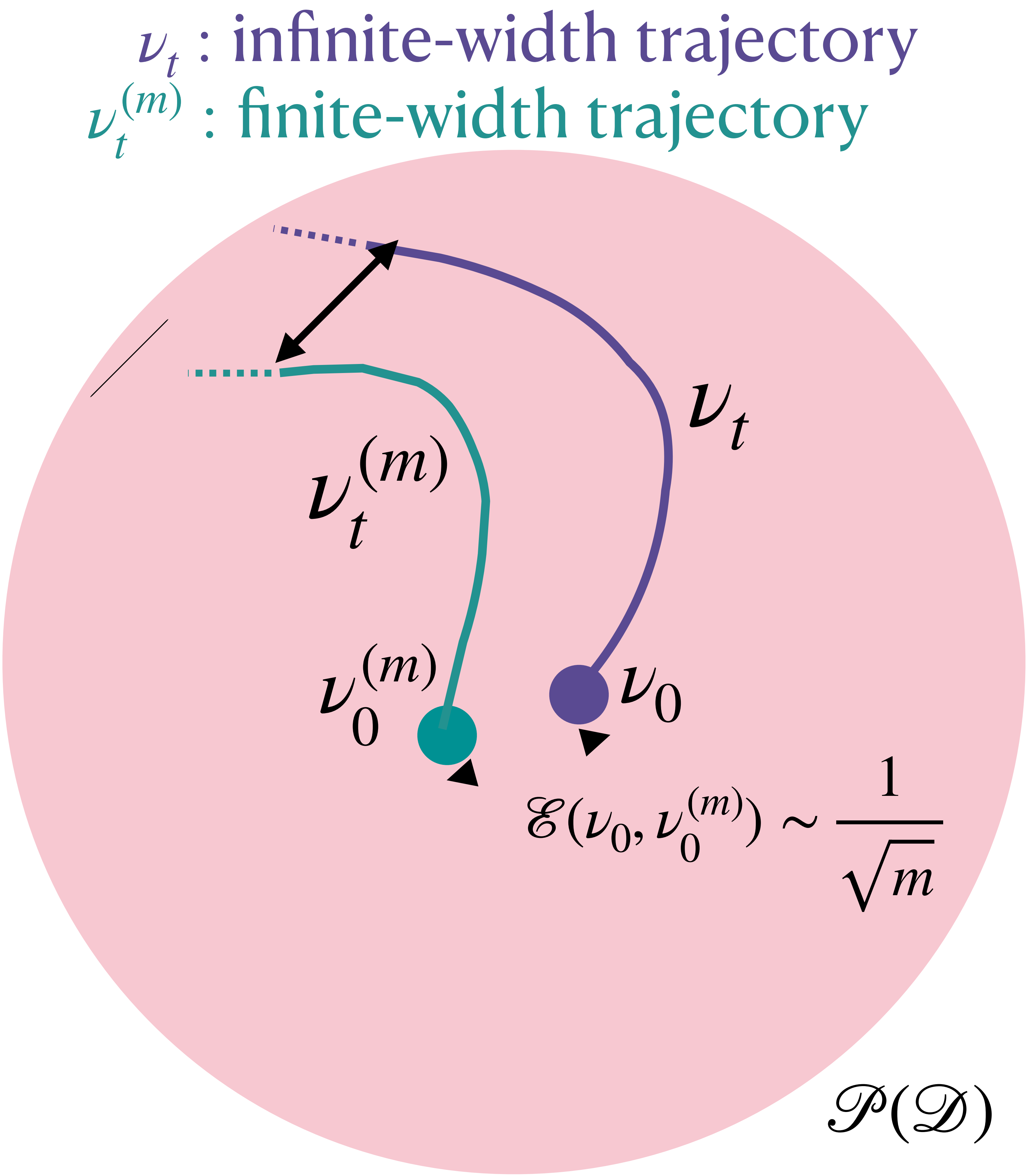
Does error remain at scale $1/\sqrt{m}$? Expand? Contract?

For how long?



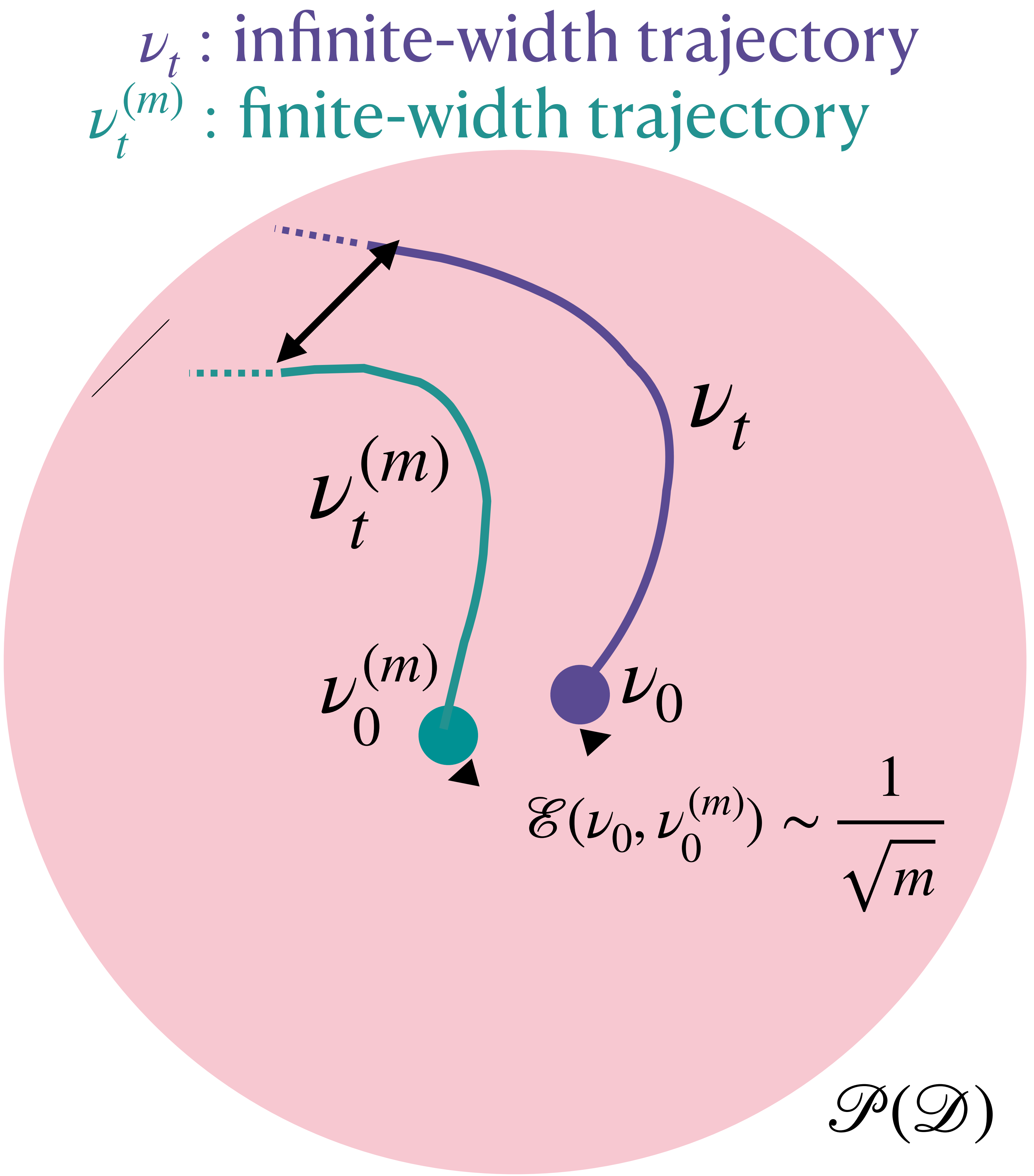
Finite-width Fluctuations

- **Goal**: For time horizon T s.t. mean-field dynamics converge, establish polynomial PoC:
  $\mathscr{E}(\nu_T, \nu_T^{(m)}) \lesssim \dfrac{\mathrm{poly}(d, T)}{\sqrt{m}}$, thus $\mathscr{L}(\nu_T^{(m)}) \lesssim \dfrac{\mathrm{poly}(d, T)}{\sqrt{m}}$.
- Generally, tension between MF-convergence rate and PoC expansion rate.



Finite-width Fluctuations

- **Goal**: For time horizon T s.t. mean-field dynamics converge, establish polynomial PoC:
$\mathscr{E}(\nu_T, \nu_T^{(m)}) \lesssim \dfrac{\mathrm{poly}(d, T)}{\sqrt{m}}$, thus $\mathscr{L}(\nu_T^{(m)}) \lesssim \dfrac{\mathrm{poly}(d, T)}{\sqrt{m}}$.
- Generally, tension between MF-convergence rate and PoC expansion rate.

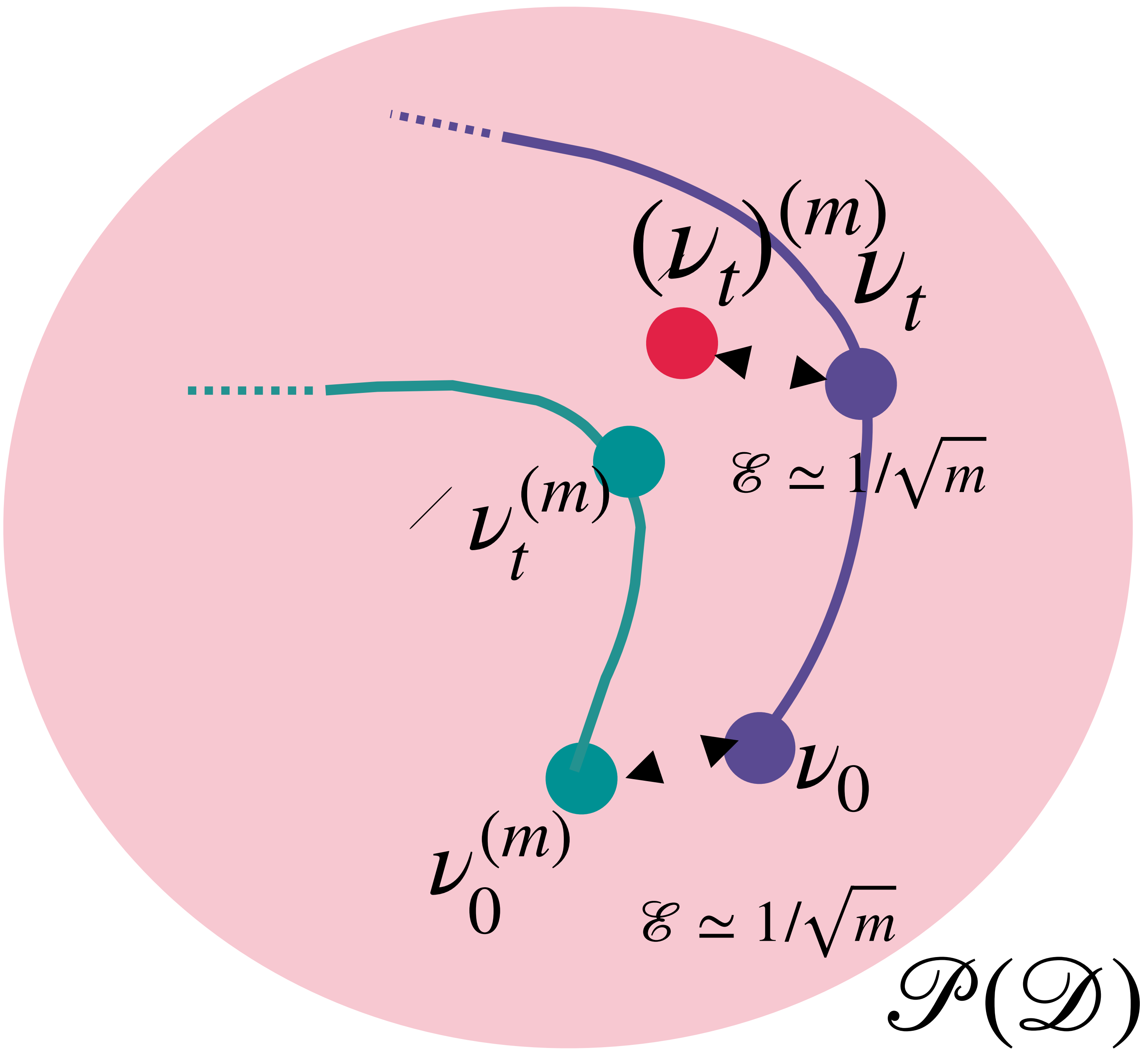


What assumptions to enable such polynomial control?

Coupling Dynamics

$\partial_t \nu_t = \mathrm{div}\left(\nabla U(\theta;\nu_t)\nu_t\right)$, $U(\,\cdot\,,\nu)$: instantaneous potential.

- Given ν_t, its empirical measure $(\nu_t)^{(m)}$ satisfies $\mathscr{E}(\nu_t,(\nu_t)^{(m)}) = O(1/\sqrt{m})$ whp.

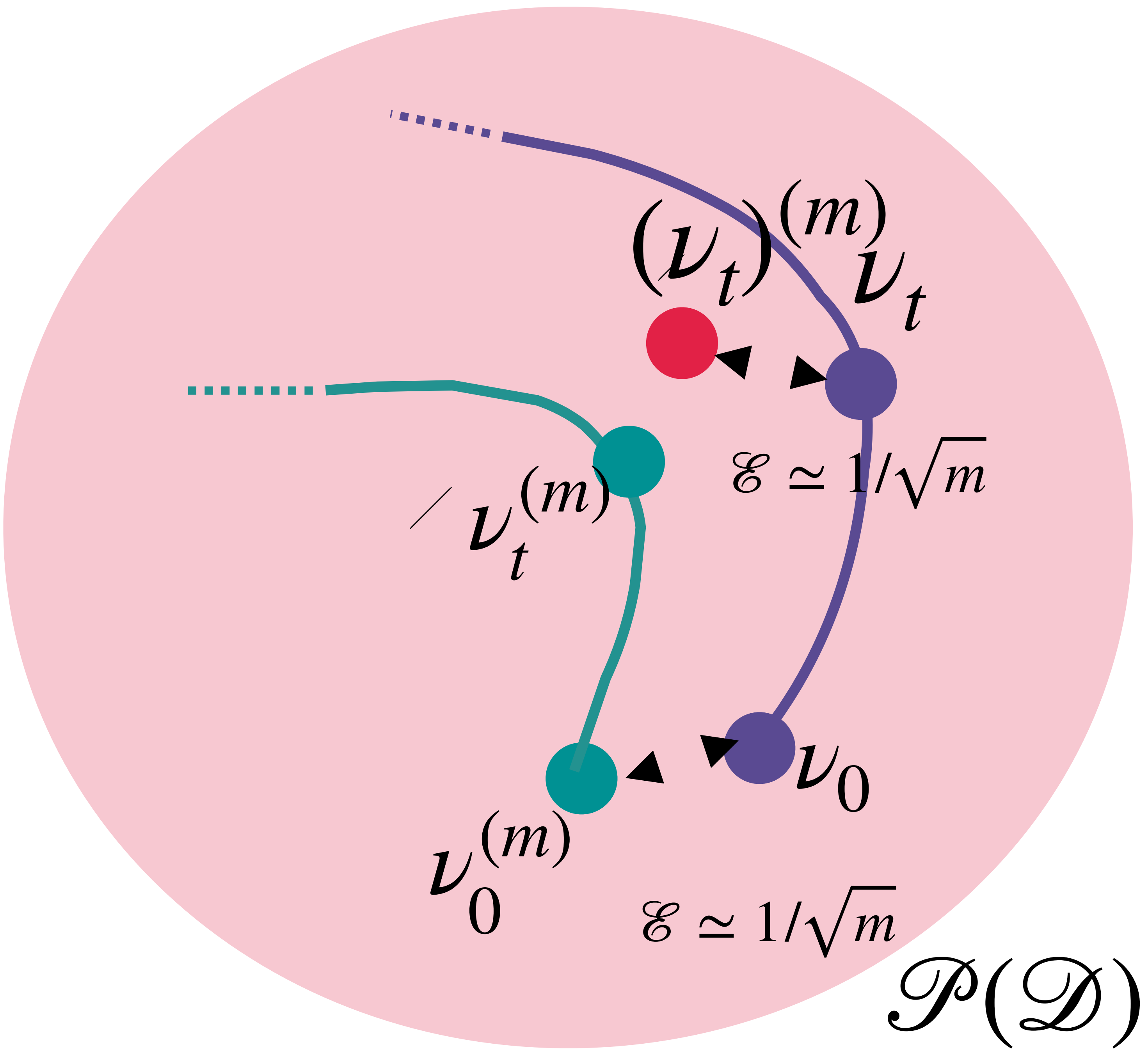


Coupling Dynamics

$\partial_t \nu_t = \mathrm{div}\left(\nabla U(\theta; \nu_t)\nu_t\right)$, $U(\,\cdot\,, \nu)$: instantaneous potential.

- Given ν_t, its empirical measure $(\nu_t)^{(m)}$ satisfies $\mathscr{E}(\nu_t, (\nu_t)^{(m)}) = \tilde{O}(1/\sqrt{m})$ whp.

How to choose a 'good' $(\nu_t)^{(m)}$, coupled with $\nu_t^{(m)}$?



$\nu_t^{(m)}$: sample, then evolve

$(\nu_t)^{(m)}$: evolve, then sample

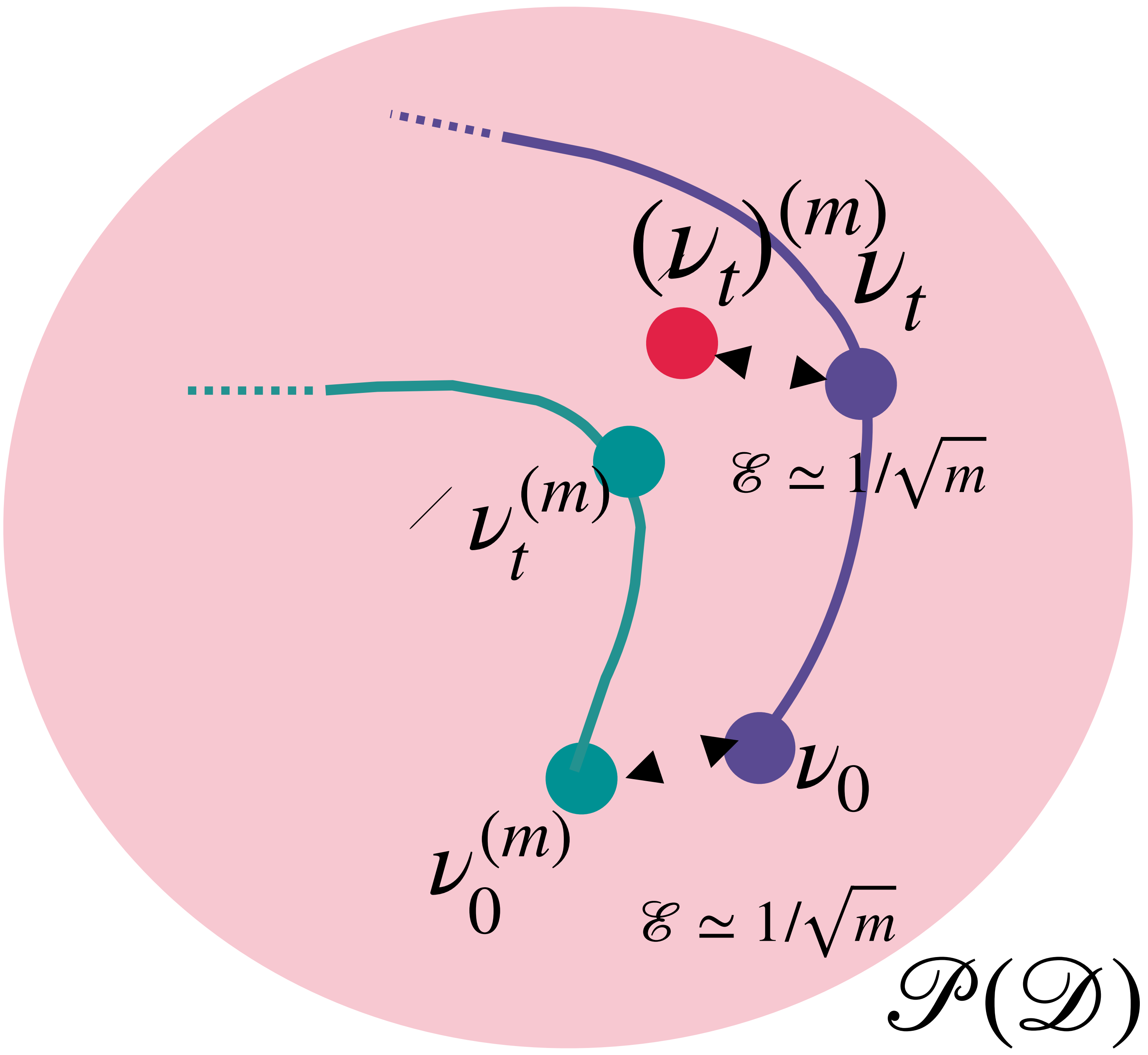
Coupling Dynamics

$$\partial_t \nu_t = \mathrm{div}\left(\nabla U(\theta;\nu_t)\nu_t\right)$$, $U(\,\cdot\,,\nu)$: instantaneous potential.

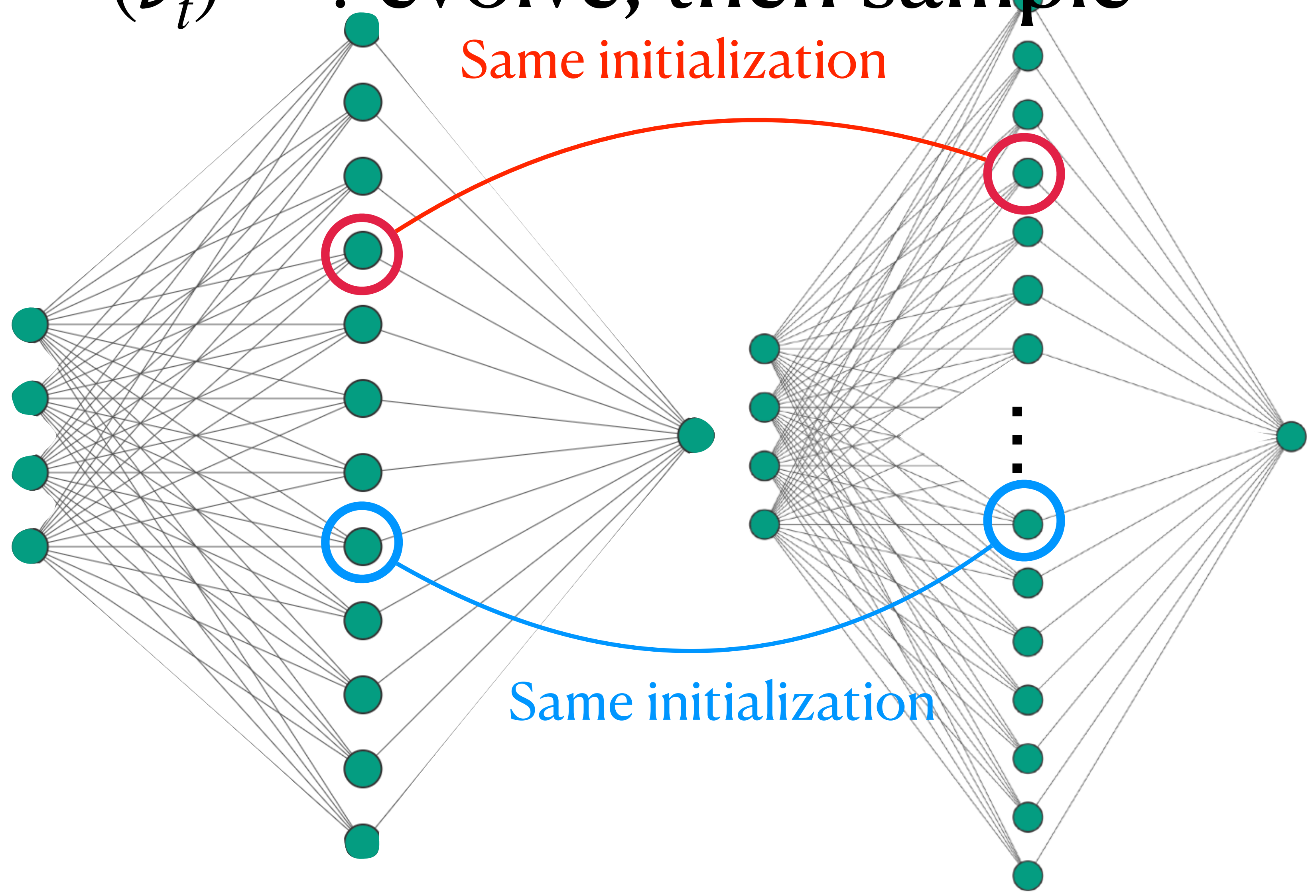
- Given ν_t, its empirical measure $(\nu_t)^{(m)}$ satisfies $\mathscr{E}(\nu_t, (\nu_t)^{(m)}) = \tilde{O}(1/\sqrt{m})$ whp.

How to choose a 'good' $(\nu_t)^{(m)}$, coupled with $\nu_t^{(m)}$?

- Solve transport eq via method of characteristics:
  Mean-Field particle evolution: $\dot{\bar{\theta}}_j = -\nabla U(\bar{\theta}_j(t);\nu_t),\ \bar{\theta}_j(0) \sim \nu_0$
  Finite-Net evolution: $\dot{\theta}_j = -\nabla U(\theta_j(t);\nu_t^{(m)}),\ \boxed{\theta_j(0) = \bar{\theta}_j(0)}.$



$\nu_t^{(m)}$: sample, then evolve
$(\nu_t)^{(m)}$: evolve, then sample



Coupling Dynamics

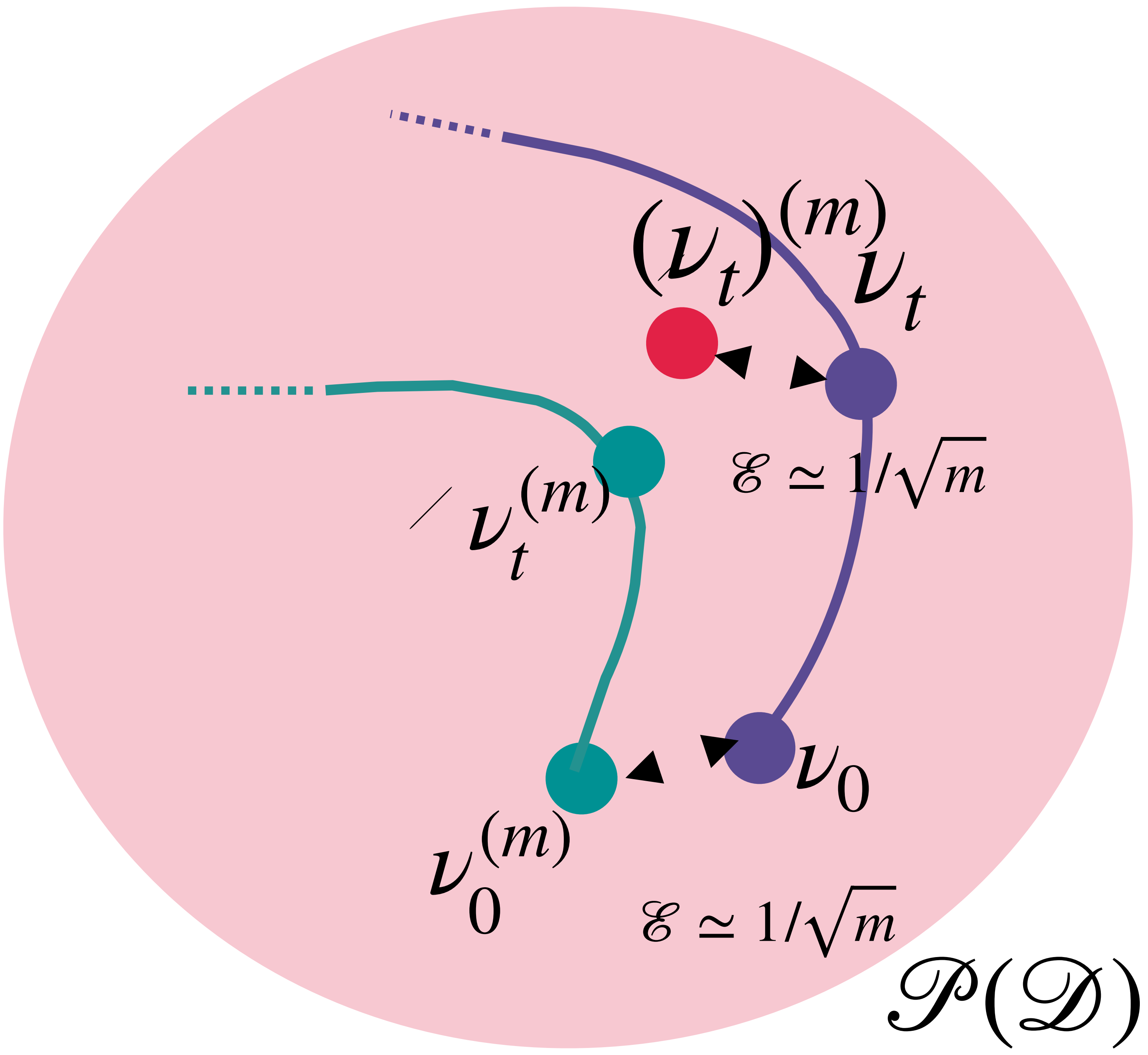
$\partial_t \nu_t = \mathrm{div}\left(\nabla U(\theta;\nu_t)\nu_t\right)$, $U(\cdot,\nu)$: instantaneous potential.

- Given ν_t, its empirical measure $(\nu_t)^{(m)}$ satisfies $\mathscr{E}(\nu_t, (\nu_t)^{(m)}) = \tilde{O}(1/\sqrt{m})$ whp.

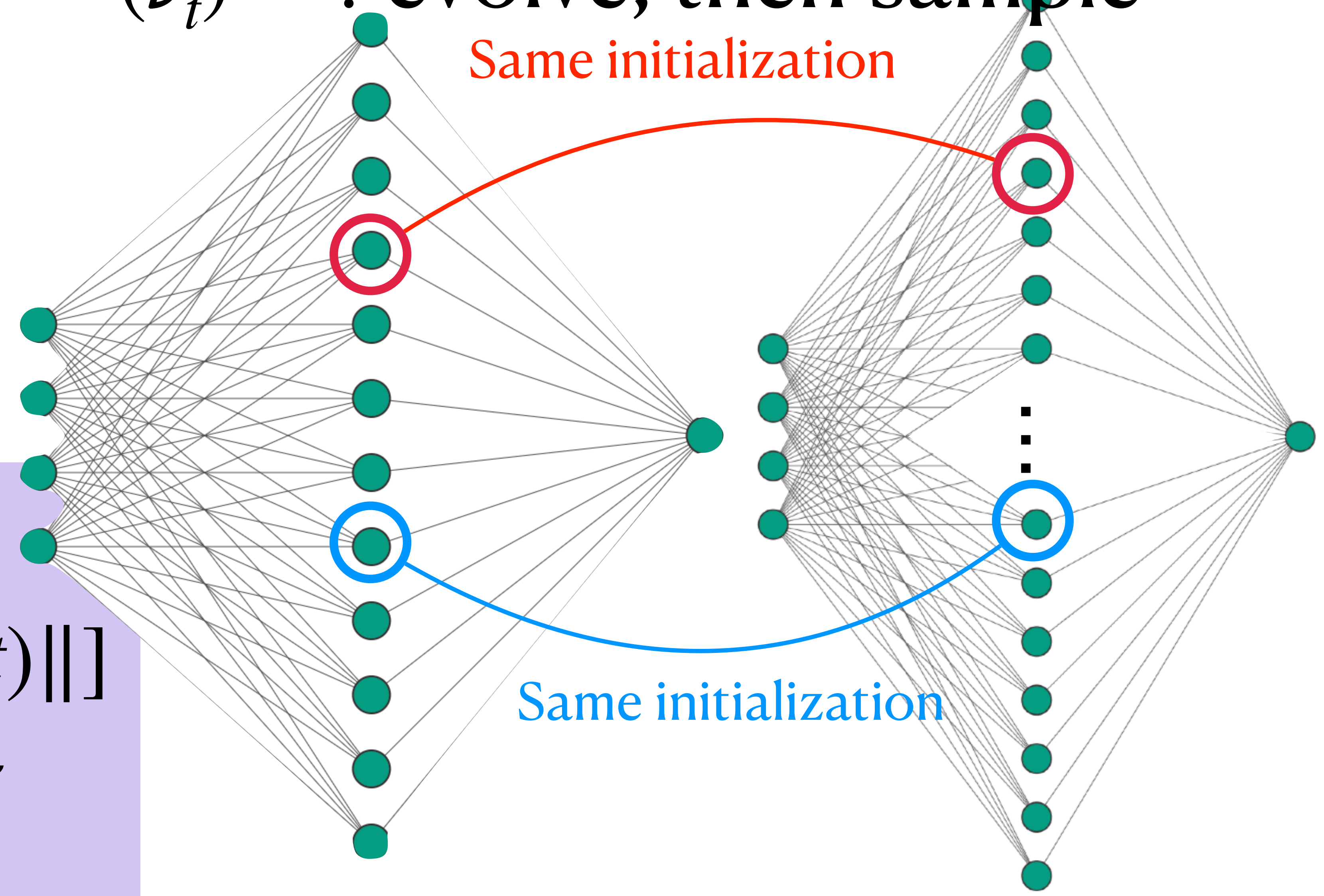
How to choose a 'good' $(\nu_t)^{(m)}$, coupled with $\nu_t^{(m)}$?

- Solve transport eq via method of characteristics:
  Mean-Field particle evolution: $\dot{\bar{\theta}}_j = -\nabla U(\bar{\theta}_j(t);\nu_t), \bar{\theta}_j(0) \sim \nu_0$
  Finite-Net evolution: $\dot{\theta}_j = -\nabla U(\theta_j(t);\nu_t^{(m)}), \boxed{\theta_j(0) = \bar{\theta}_j(0)}.$

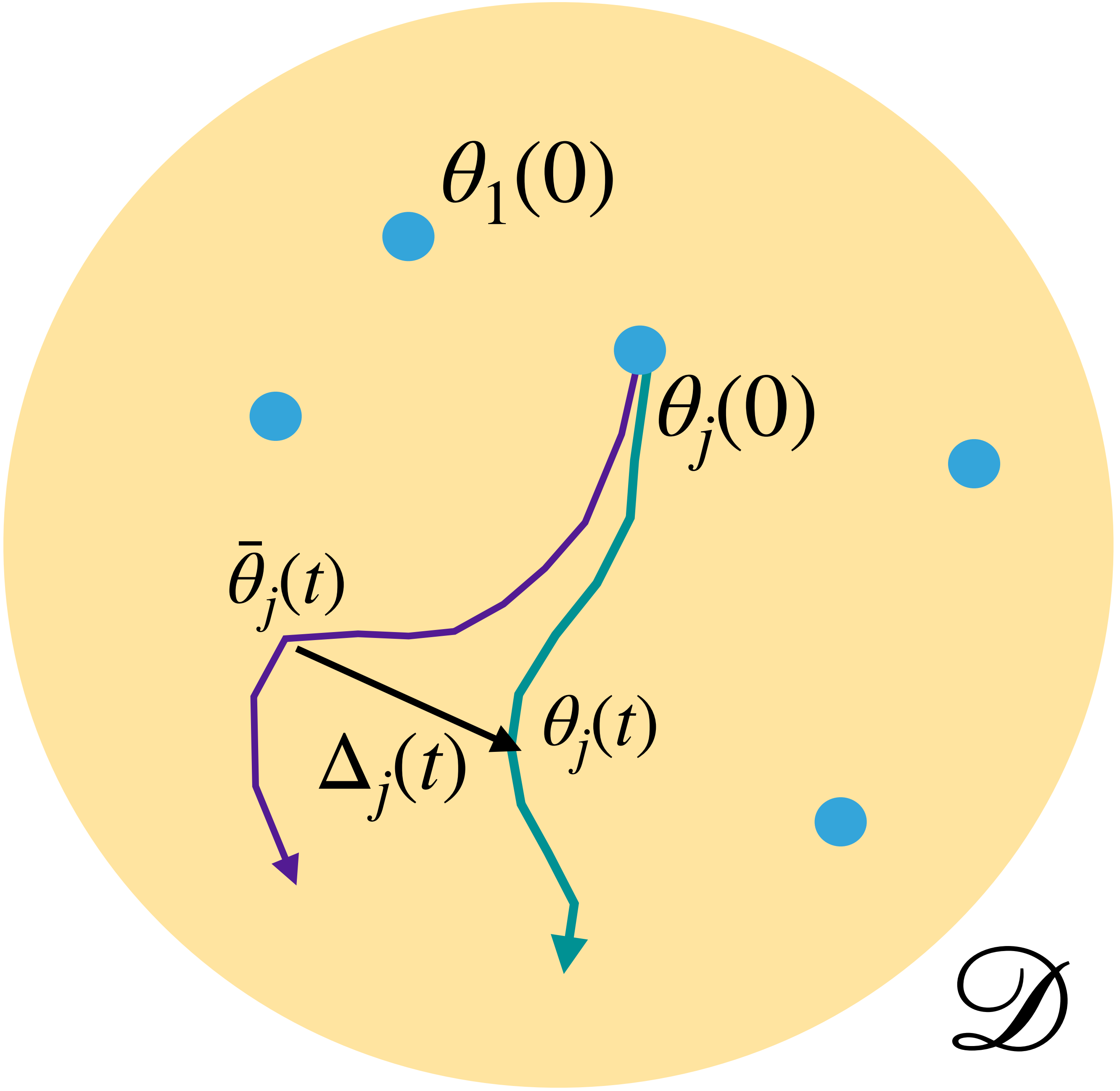
- **Proposition**: under mild regularity, we have
  $\mathscr{E}(\nu_t^{(m)},\nu_t) \lesssim O(1\sqrt{m}) + W_1(\nu_t^{(m)},(\nu_t)^{(m)}) \leq O(1/\sqrt{m}) + \mathbb{E}_j[\underbrace{\|\theta_j(t) - \bar{\theta}_j(t)\|}_{\Delta_j(t)}]$



$\nu_t^{(m)}$: sample, then evolve
$(\nu_t)^{(m)}$: evolve, then sample



Coupling Dynamics and Gronwall



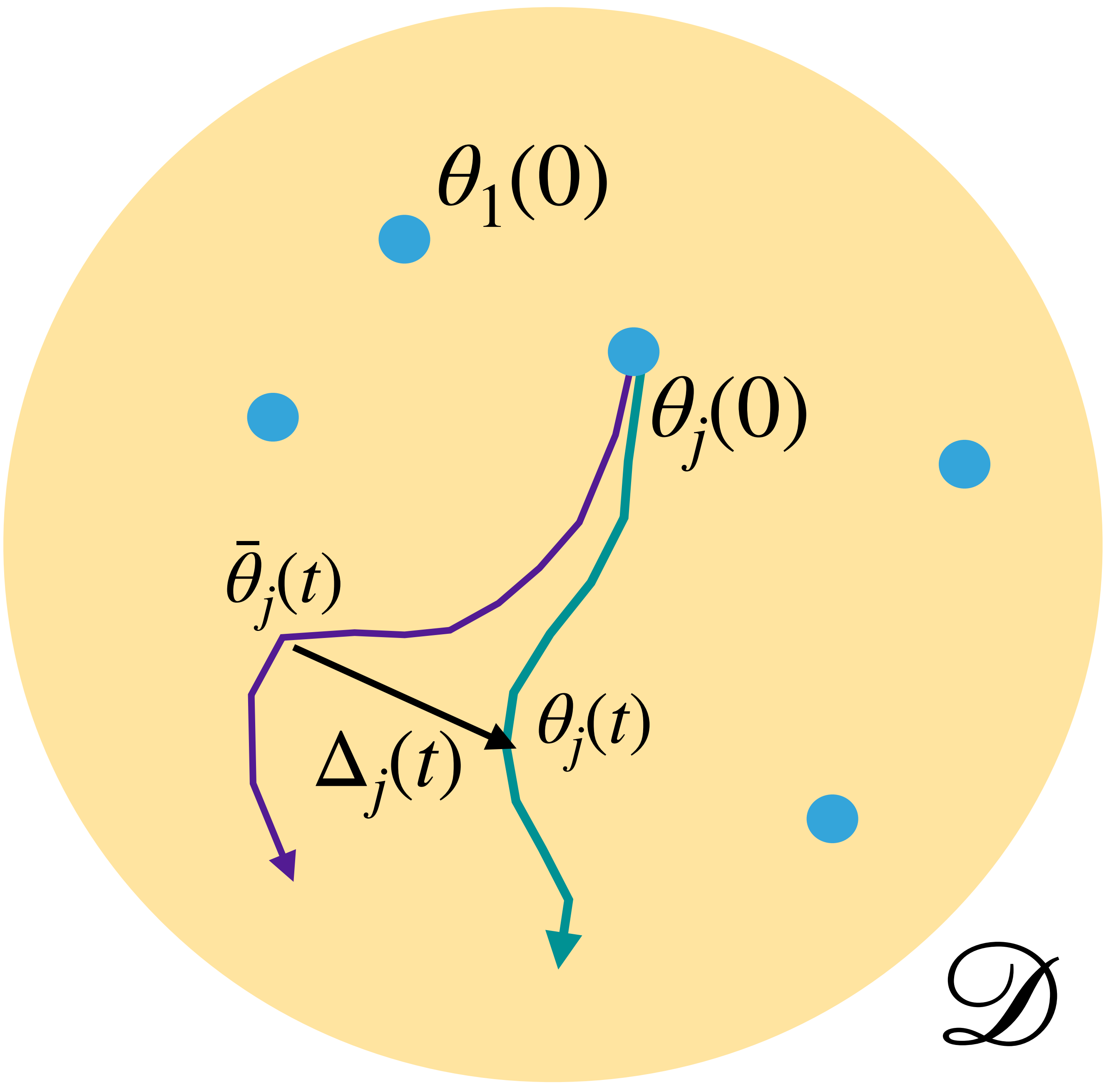
$\Delta_j(t) = \theta_j(t) - \bar{\theta}_j(t)$: Coupling errors.

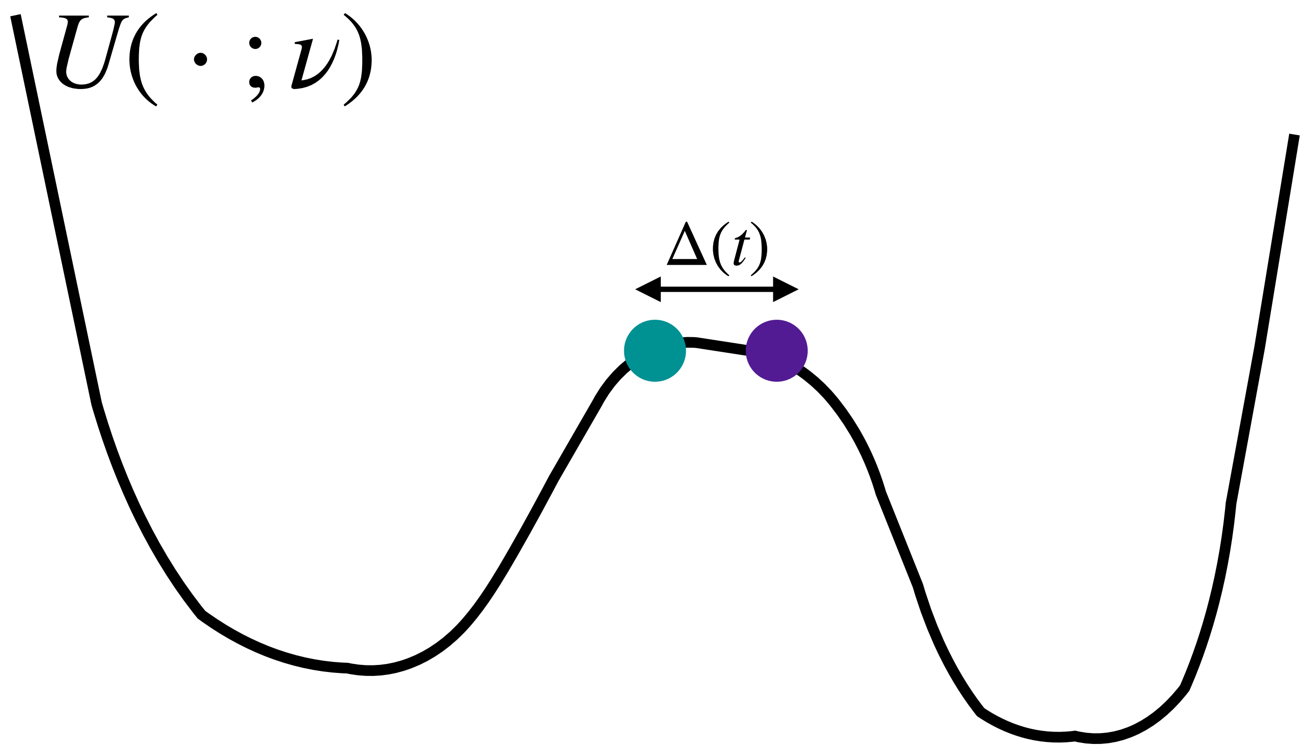
- How does this error evolve over time?

Coupling Dynamics and Gronwall

$\Delta_j(t) = \theta_j(t) - \bar{\theta}_j(t)$: Coupling errors.

- How does this error evolve over time?
- $\dot{\Delta}_j = \dot{\theta}_j - \dot{\bar{\theta}}_j = \nabla U(\theta_j; \nu_t^{(m)}) - \nabla U(\bar{\theta}_j; \nu_t)$
- **Key difficulty**: non-convex potential expands trajectories.

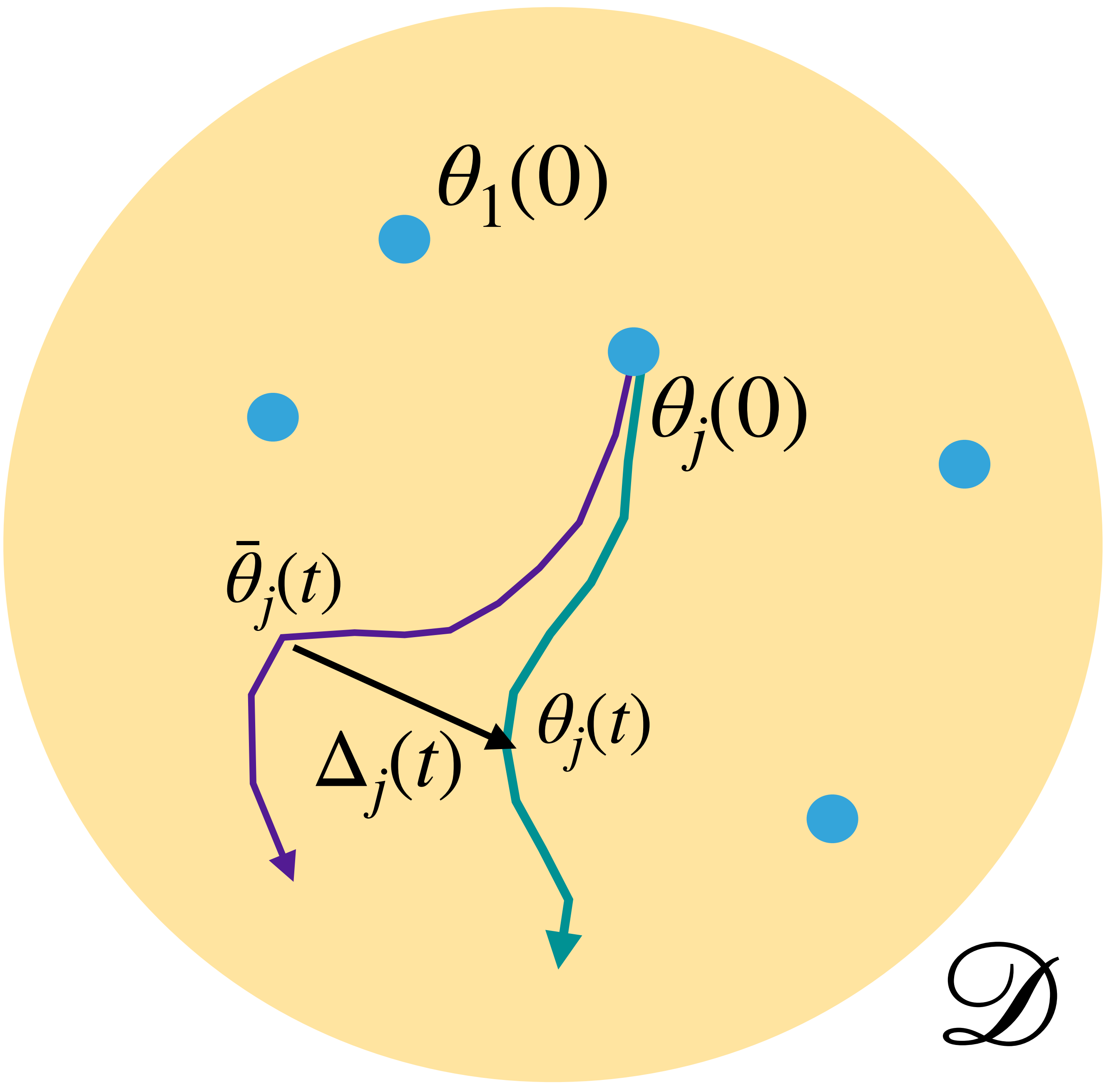


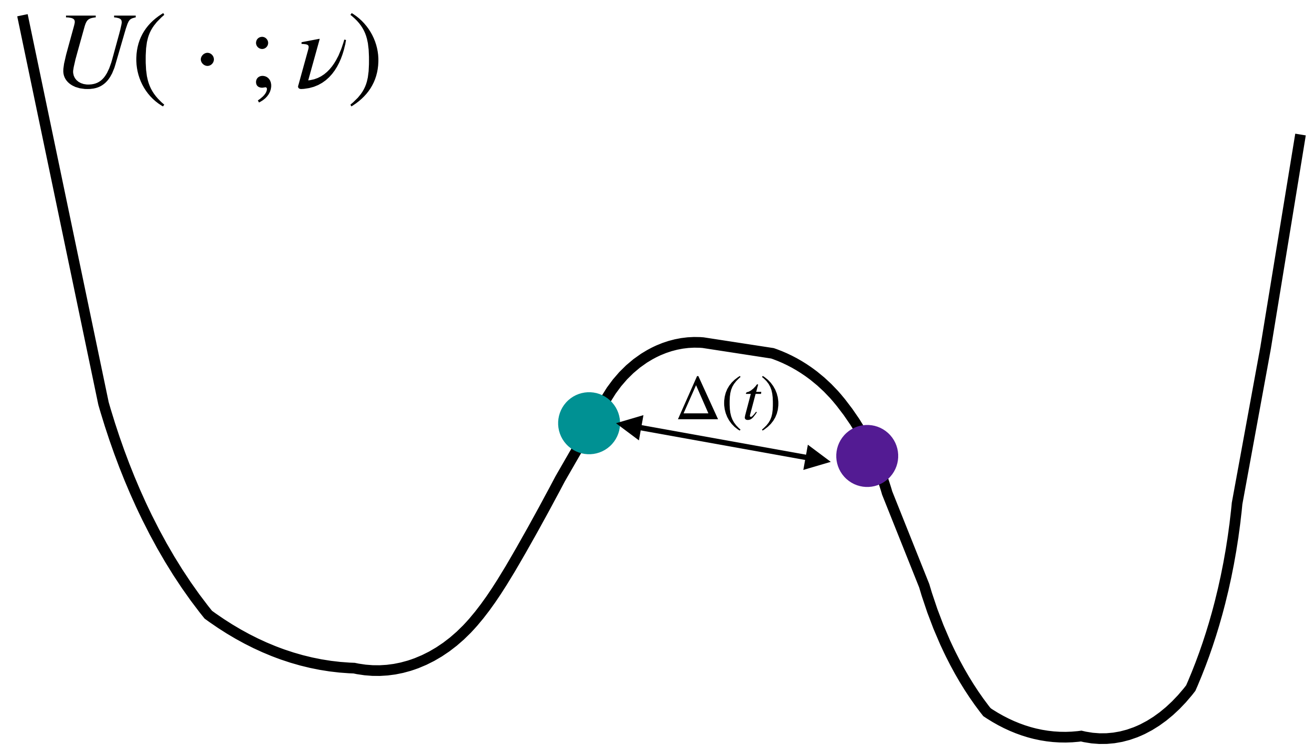


Coupling Dynamics and Gronwall

$\Delta_j(t) = \theta_j(t) - \bar{\theta}_j(t)$: Coupling errors.

- How does this error evolve over time?
- $\dot{\Delta}_j = \dot{\theta}_j - \dot{\bar{\theta}}_j = \nabla U(\theta_j; \nu_t^{(m)}) - \nabla U(\bar{\theta}_j; \nu_t)$
- **Key difficulty**: non-convex potential expands trajectories.

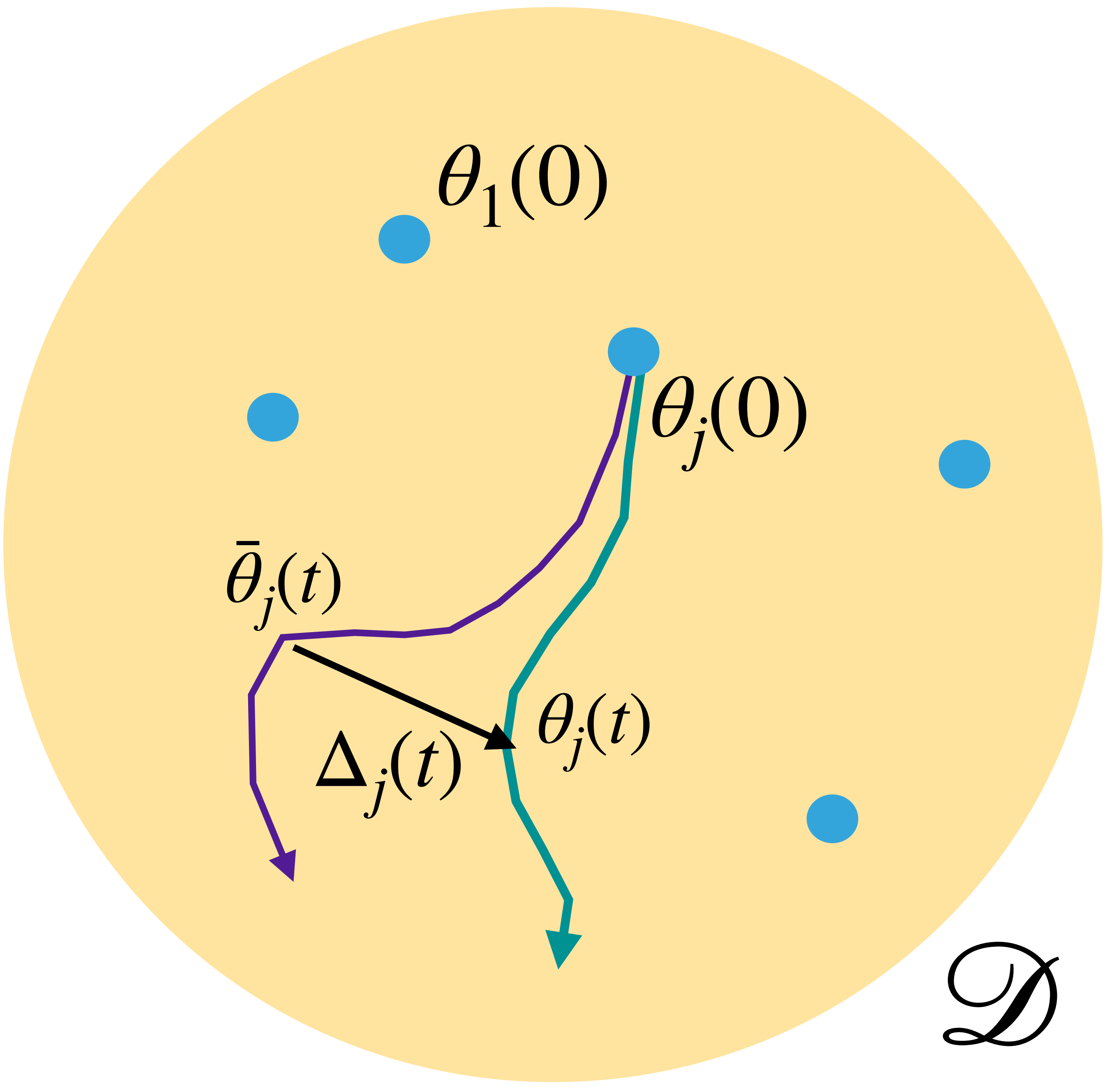


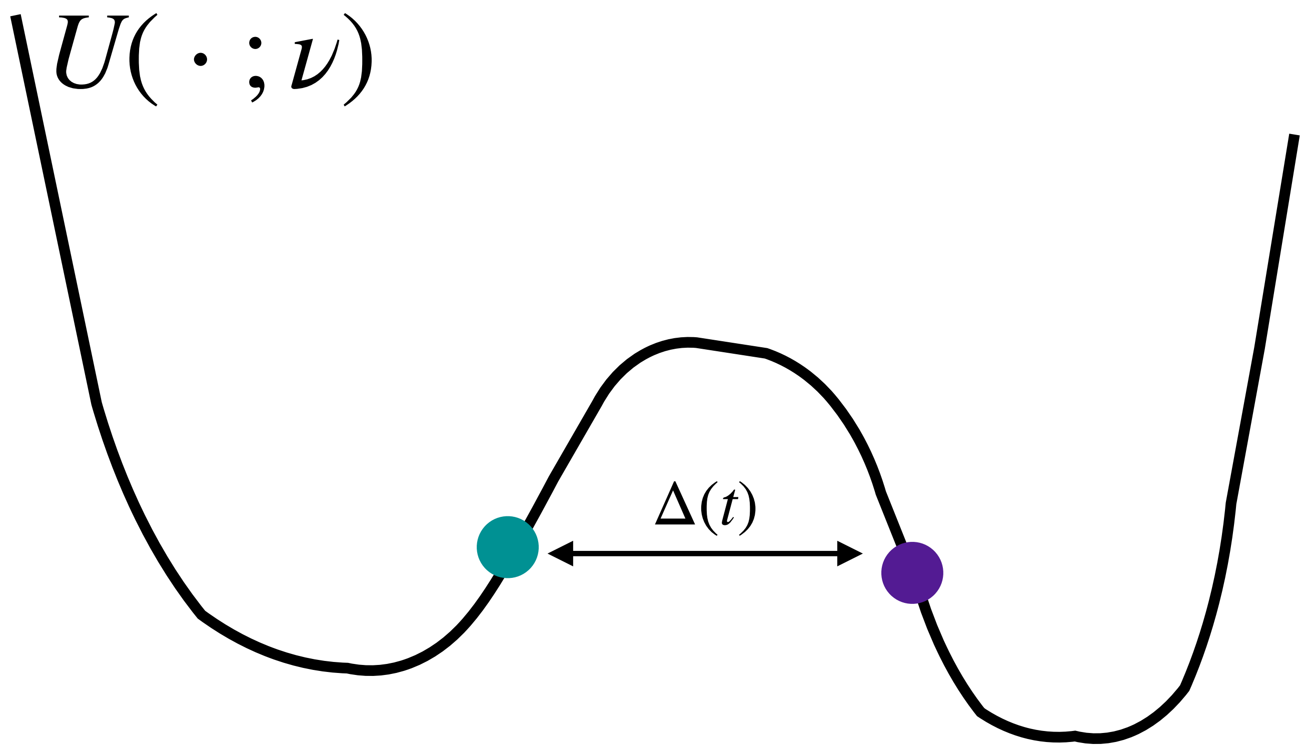


Coupling Dynamics and Gronwall

$\Delta_j(t) = \theta_j(t) - \bar{\theta}_j(t)$: Coupling errors.

- How does this error evolve over time?
- $\dot{\Delta}_j = \dot{\theta}_j - \dot{\bar{\theta}}_j = \nabla U(\theta_j; \nu_t^{(m)}) - \nabla U(\bar{\theta}_j; \nu_t)$
- **Key difficulty**: non-convex potential expands trajectories.



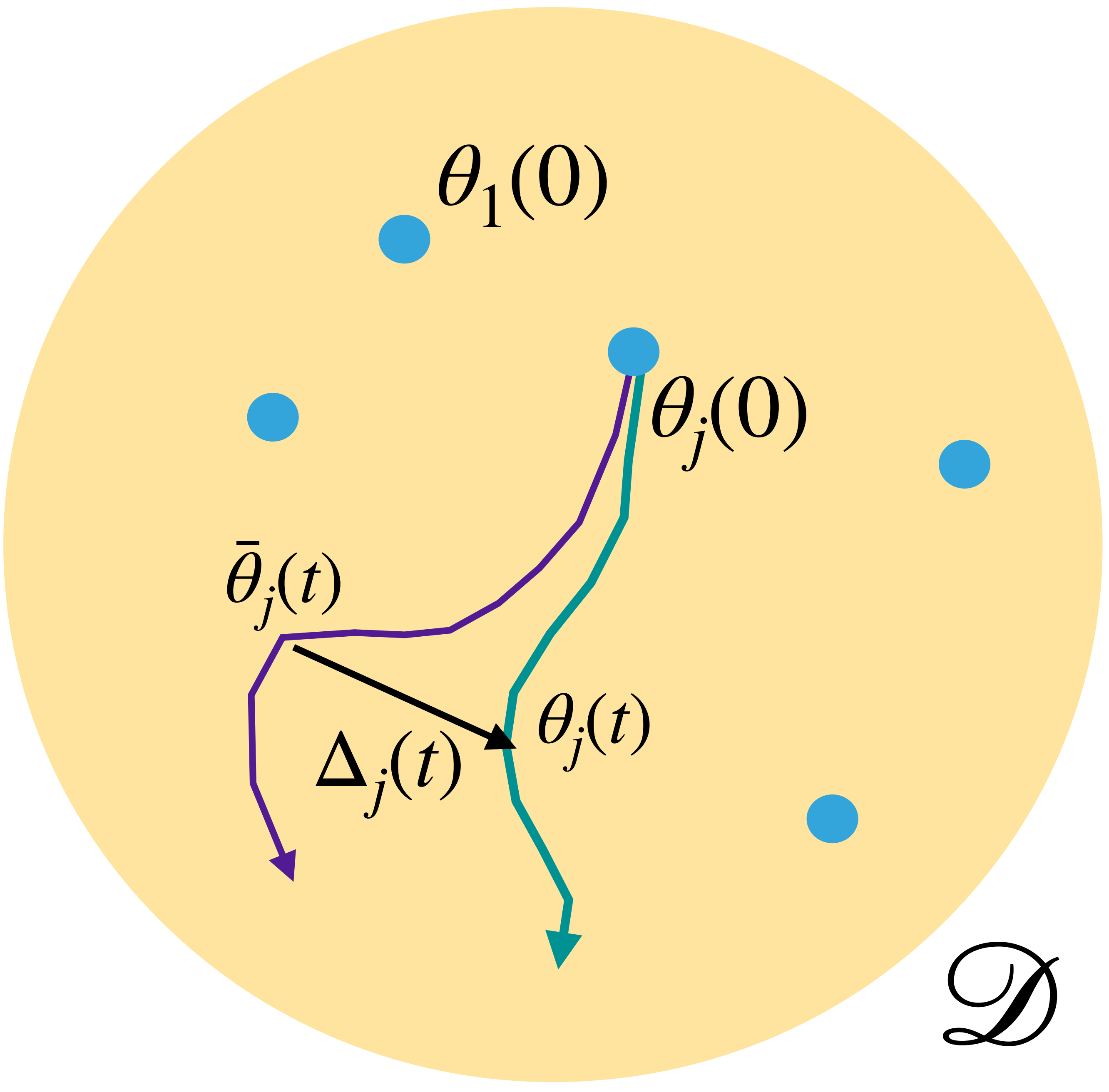


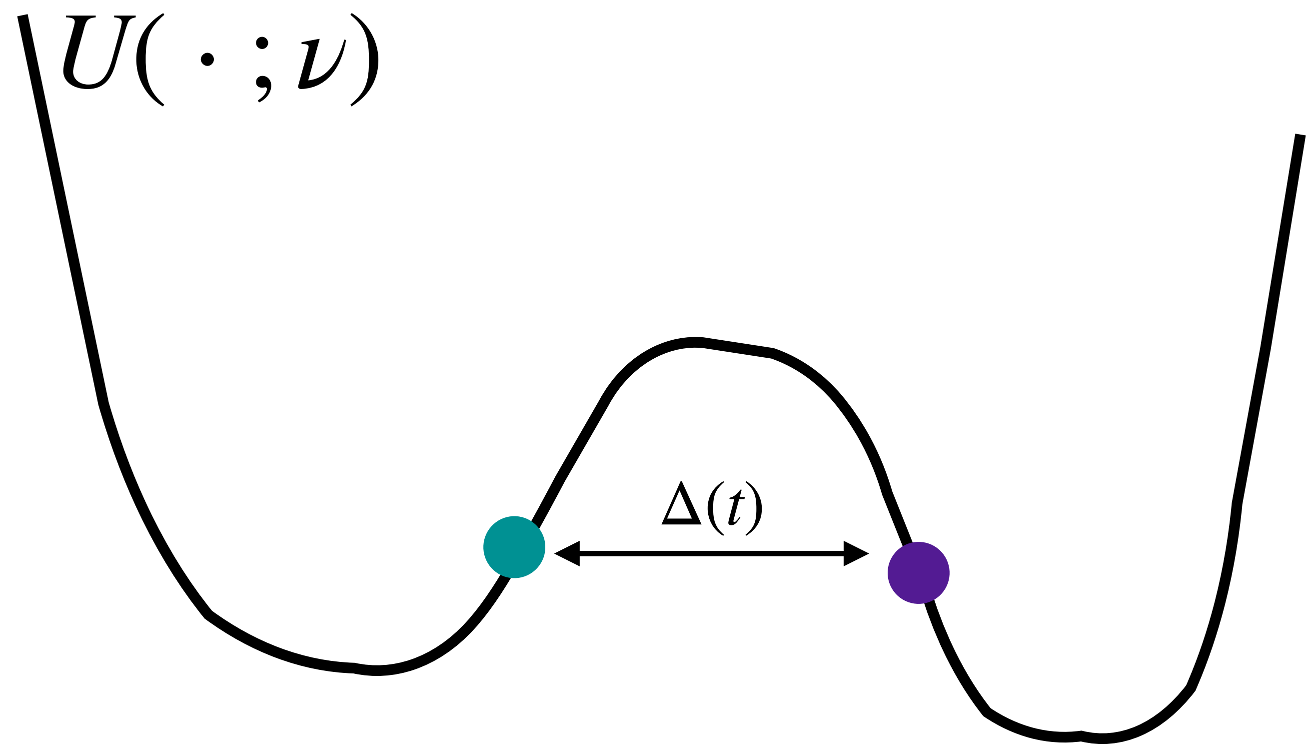
Coupling Dynamics and Gronwall

$\Delta_j(t) = \theta_j(t) - \bar{\theta}_j(t)$: Coupling errors.

- How does this error evolve over time?
- $\dot{\Delta}_j = \dot{\theta}_j - \dot{\bar{\theta}}_j = \nabla U(\theta_j; \nu_t^{(m)}) - \nabla U(\bar{\theta}_j; \nu_t)$
- **Key difficulty**: non-convex potential expands trajectories.
- Leveraging uniform Lipschitz smoothness:

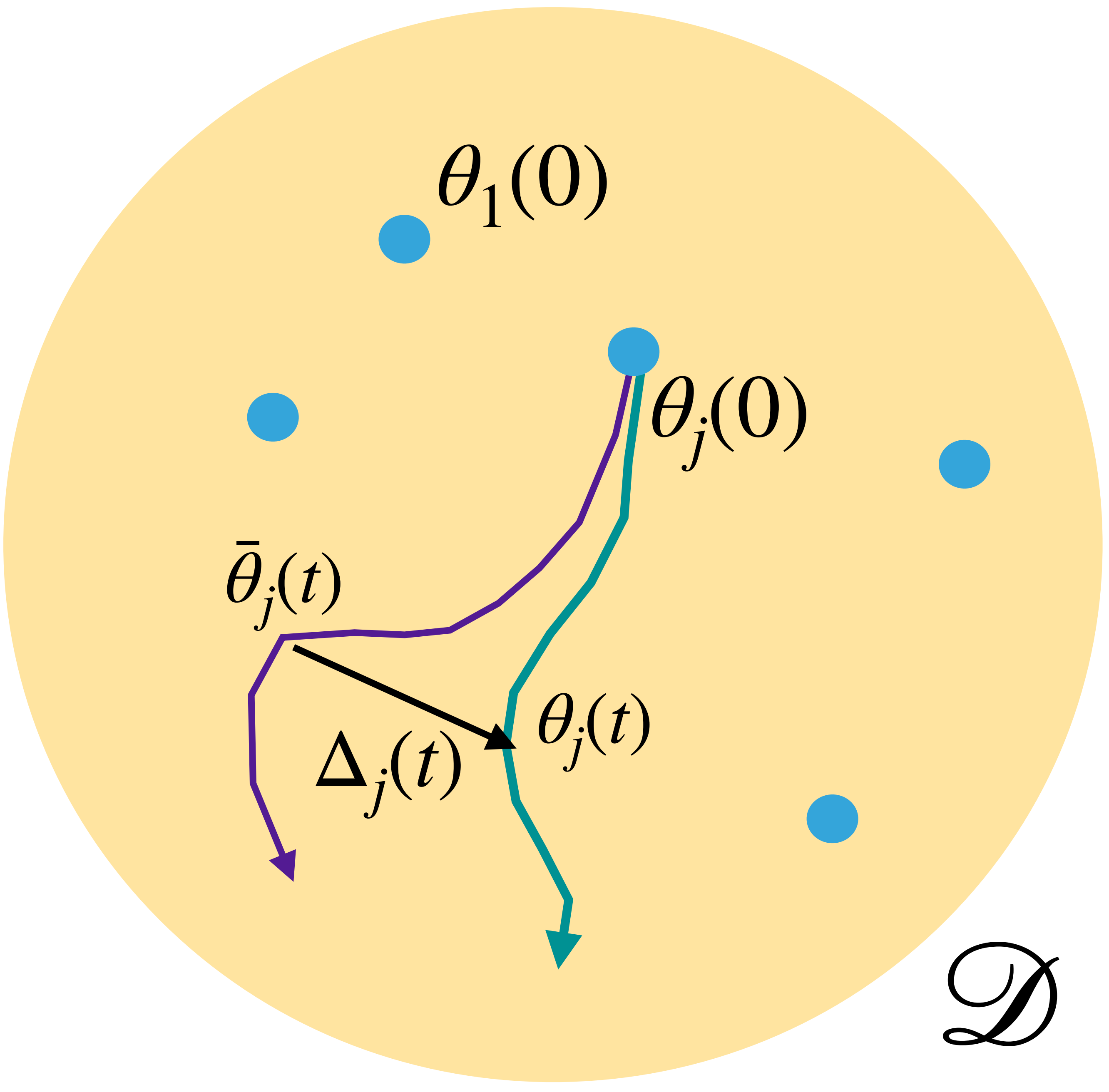
$$\frac{d}{dt}\mathbb{E}_j\|\Delta_j(t)\| \le L_\theta \mathbb{E}_j\|\Delta_j(t)\| + L_\nu W_1(\nu_t^{(m)}, (\nu_t)^{(m)}) + O(1/\sqrt{m})$$
$$\le (L_\theta \vee L_\nu)\mathbb{E}_j\|\Delta_j(t)\| + O(1/\sqrt{m})\,.$$





$L = L_\theta \vee L_\nu$

Coupling Dynamics and Gronwall

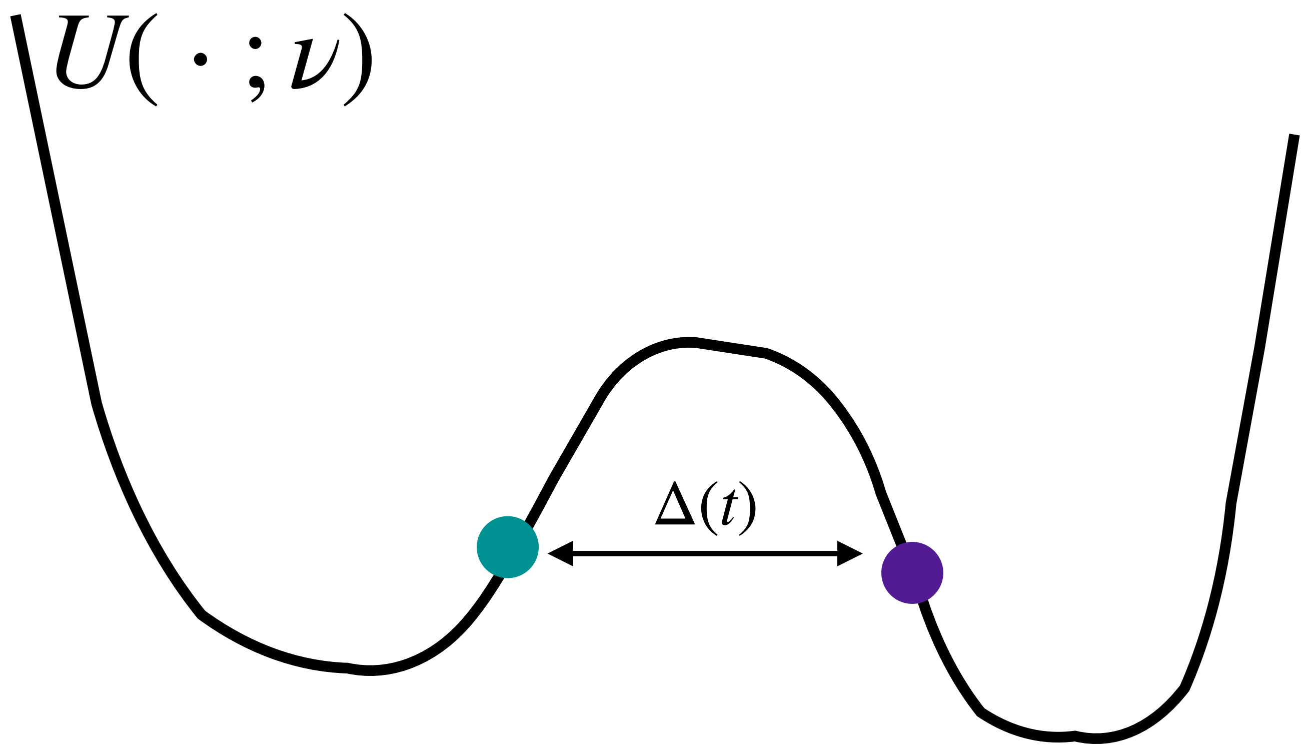


$\Delta_j(t) = \theta_j(t) - \bar{\theta}_j(t)$: Coupling errors.

- How does this error evolve over time?
- $\dot{\Delta}_j = \dot{\theta}_j - \dot{\bar{\theta}}_j = \nabla U(\theta_j; \nu_t^{(m)}) - \nabla U(\bar{\theta}_j; \nu_t)$
- **Key difficulty**: non-convex potential expands trajectories.
- Leveraging uniform Lipschitz smoothness:

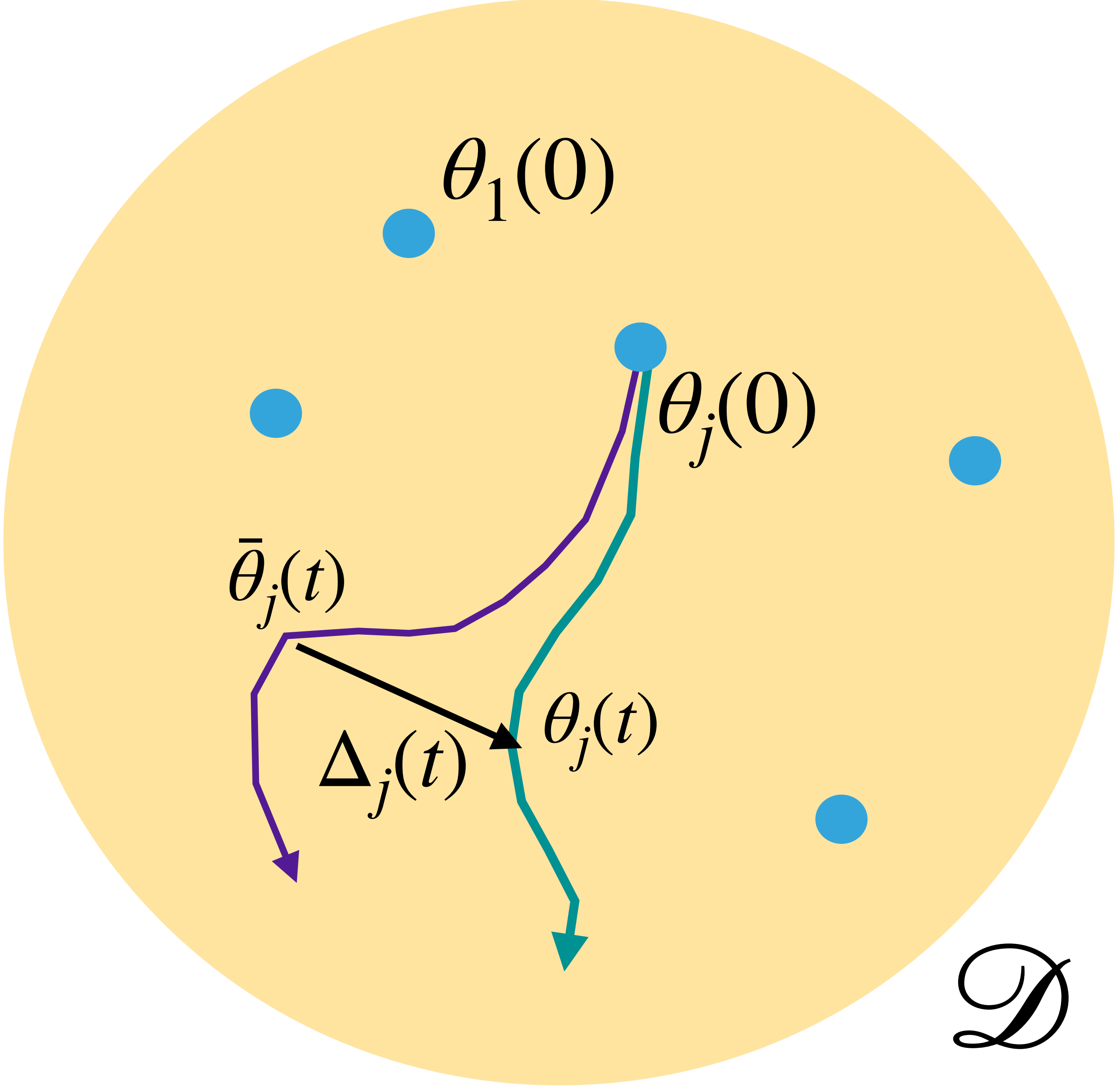
$$\frac{d}{dt}\mathbb{E}_j\|\Delta_j(t)\| \leq L_\theta \mathbb{E}_j\|\Delta_j(t)\| + L_\nu W_1(\nu_t^{(m)}, (\nu_t)^{(m)}) + O(1/\sqrt{m})$$
$$\leq (L_\theta \vee L_\nu)\mathbb{E}_j\|\Delta_j(t)\| + O(1/\sqrt{m}).$$

- PoC via Gronwall's inequality: $\mathscr{E}(\nu_t^{(m)}, \nu_t) \lesssim \dfrac{\exp(Lt)}{\sqrt{m}}$.
  - Exploited in [Mei et al, Misiakiewicz et al, Mahankali et al] for **short** time-horizons, e.g $T = O(1)$ or $T = O(\log d)$. Morally IE ≤ 2 'type' problems.



$L = L_\theta \vee L_\nu$

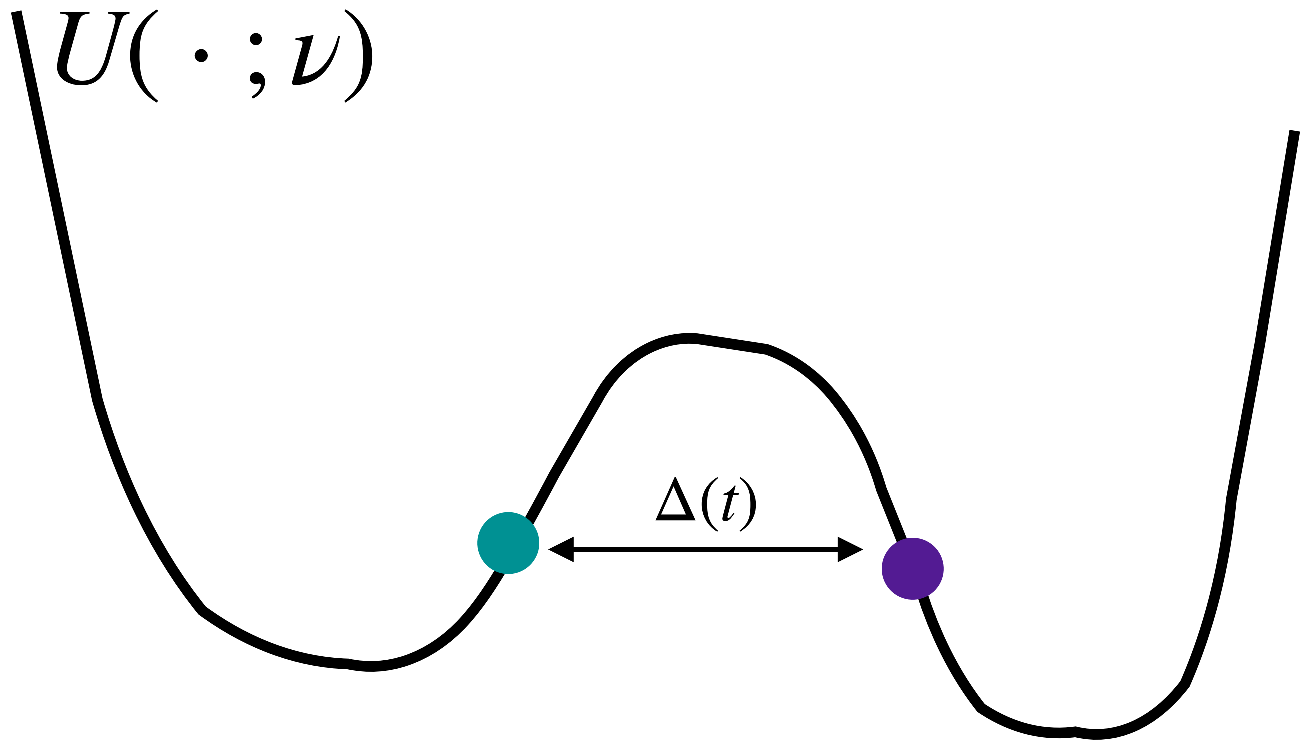
Coupling Dynamics and Gronwall



$\Delta_j(t) = \theta_j(t) - \bar{\theta}_j(t)$: Coupling errors.

- How does this error evolve over time?
- $\dot{\Delta}_j = \dot{\theta}_j - \dot{\bar{\theta}}_j = \nabla U(\theta_j; \nu_t^{(m)}) - \nabla U(\bar{\theta}_j; \nu_t)$
- **Key difficulty**: non-convex potential expands trajectories.
- Leveraging uniform Lipschitz smoothness:

$$\begin{aligned}\frac{d}{dt}\mathbb{E}_j\|\Delta_j(t)\| &\leq L_\theta \mathbb{E}_j\|\Delta_j(t)\| + L_\nu W_1(\nu_t^{(m)}, (\nu_t)^{(m)}) + O(1/\sqrt{m}) \\ &\leq (L_\theta \vee L_\nu)\mathbb{E}_j\|\Delta_j(t)\| + O(1/\sqrt{m})\,.\end{aligned}$$



- PoC via Gronwall's inequality: $\mathscr{E}(\nu_t^{(m)}, \nu_t) \lesssim \dfrac{\exp(Lt)}{\sqrt{m}}$.
  - Exploited in [Mei et al, Misiakiewicz et al, Mahankali et al] for **short** time-horizons, e.g $T = O(1)$ or $T = O(\log d)$. Morally IE ≤ 2 'type' problems.

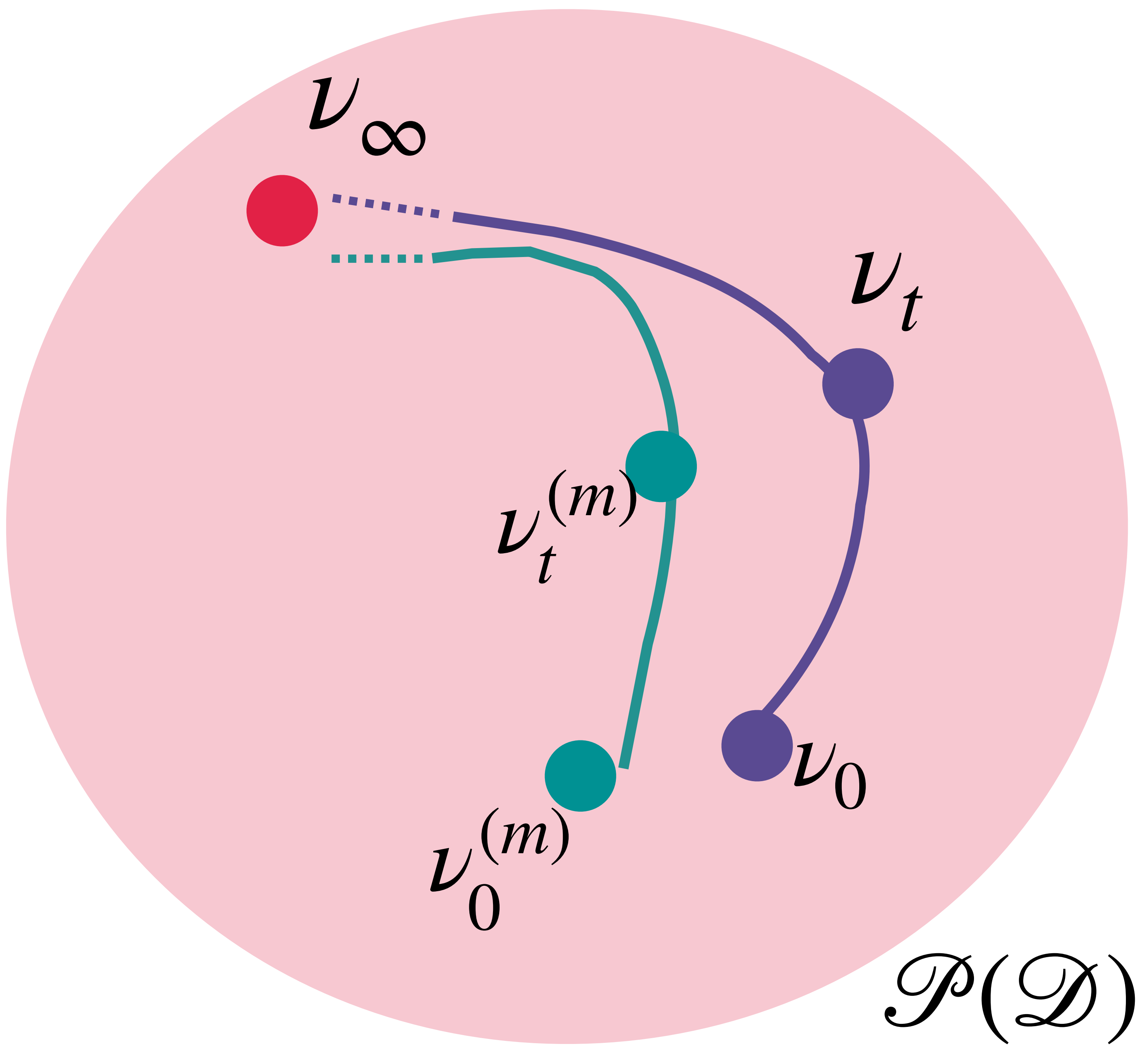
Excludes many situations of interest!

$L = L_\theta \vee L_\nu$

PoC via Mean-Field Langevin Contraction

[Chizat et al., Suzuki et al., Nitanda]

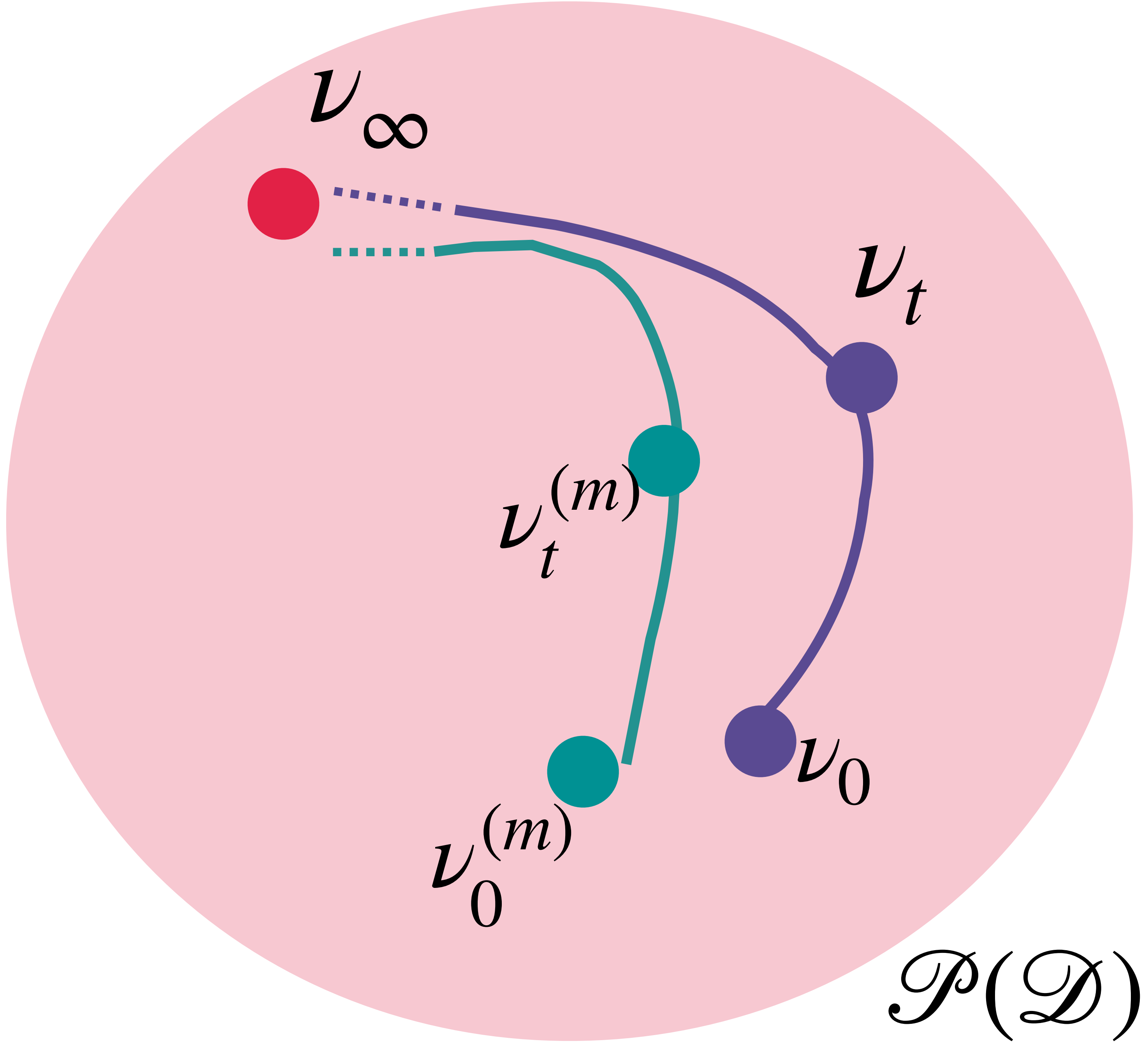
- Alternatively, one can regularize using entropic term:
  $\tilde{\mathscr{L}}(\nu) = \mathscr{L}(\nu) + \lambda H(\nu)$.



PoC via Mean-Field Langevin Contraction

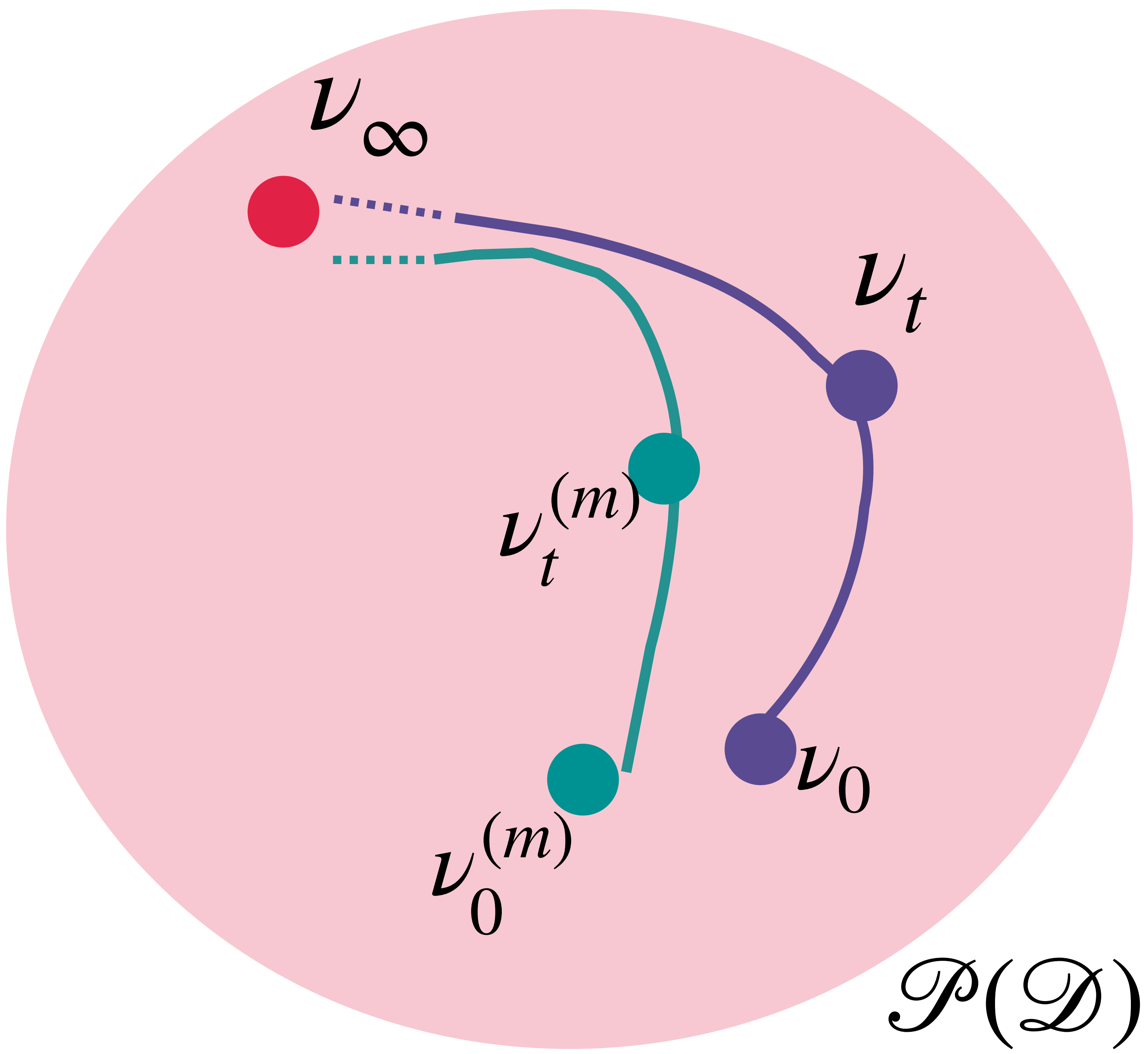
[Chizat et al., Suzuki et al., Nitanda]

- Alternatively, one can regularize using entropic term:
  $\tilde{\mathscr{L}}(\nu) = \mathscr{L}(\nu) + \lambda H(\nu)$.
- Noisy dynamics creates Wasserstein contraction via Log-Sobolev Inequality, leading to [Nitanda'24]:
  $\mathscr{E}(\nu_t^{(m)}, \nu_t)^2 \lesssim \frac{1}{m} + \exp(-2\alpha_m \lambda t)\mathscr{E}(\nu_0^{(m)}, \nu_*)$
  - Here, α_m is the LSI of minimiser, of order $\alpha_m \simeq \exp(-\Theta(m/\lambda))$.



PoC via Mean-Field Langevin Contraction

[Chizat et al., Suzuki et al., Nitanda]

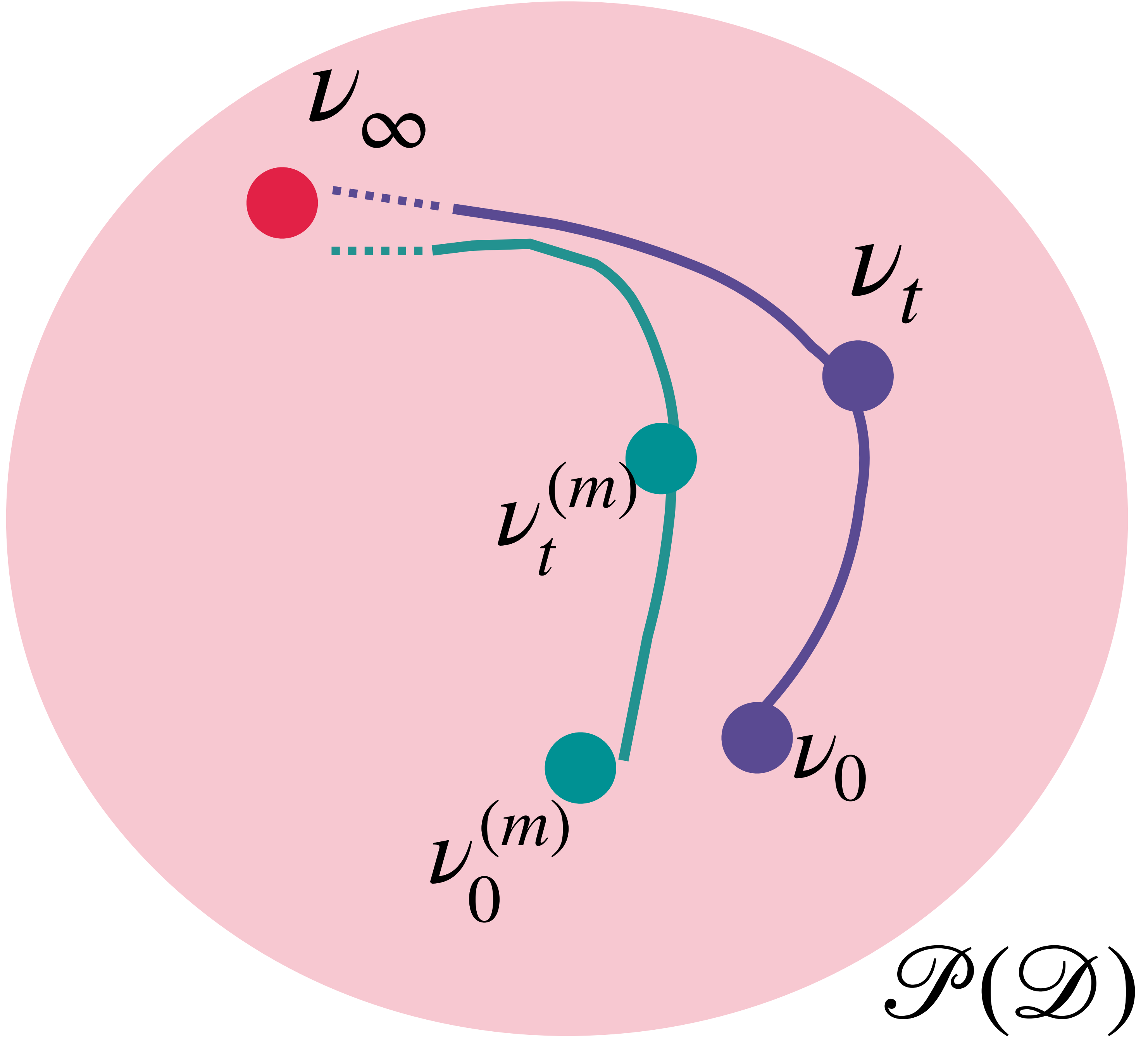


- Alternatively, one can regularize using entropic term: $\tilde{\mathscr{L}}(\nu) = \mathscr{L}(\nu) + \lambda H(\nu)$.
- Noisy dynamics creates Wasserstein contraction via Log-Sobolev Inequality, leading to [Nitanda'24]: $\mathscr{E}(\nu_t^{(m)}, \nu_t)^2 \lesssim \frac{1}{m} + \exp(-2\alpha_m \lambda t)\mathscr{E}(\nu_0^{(m)}, \nu_*)$
  - Here, α_m is the LSI of minimiser, of order $\alpha_m \simeq \exp(-\Theta(m/\lambda))$.
  - Efficient particle approximation, but cursed iteration complexity $T = O\left(\frac{\log \epsilon^{-1}}{\alpha_m^2 \lambda \epsilon}\right)$.

PoC via Mean-Field Langevin Contraction

[Chizat et al., Suzuki et al., Nitanda]

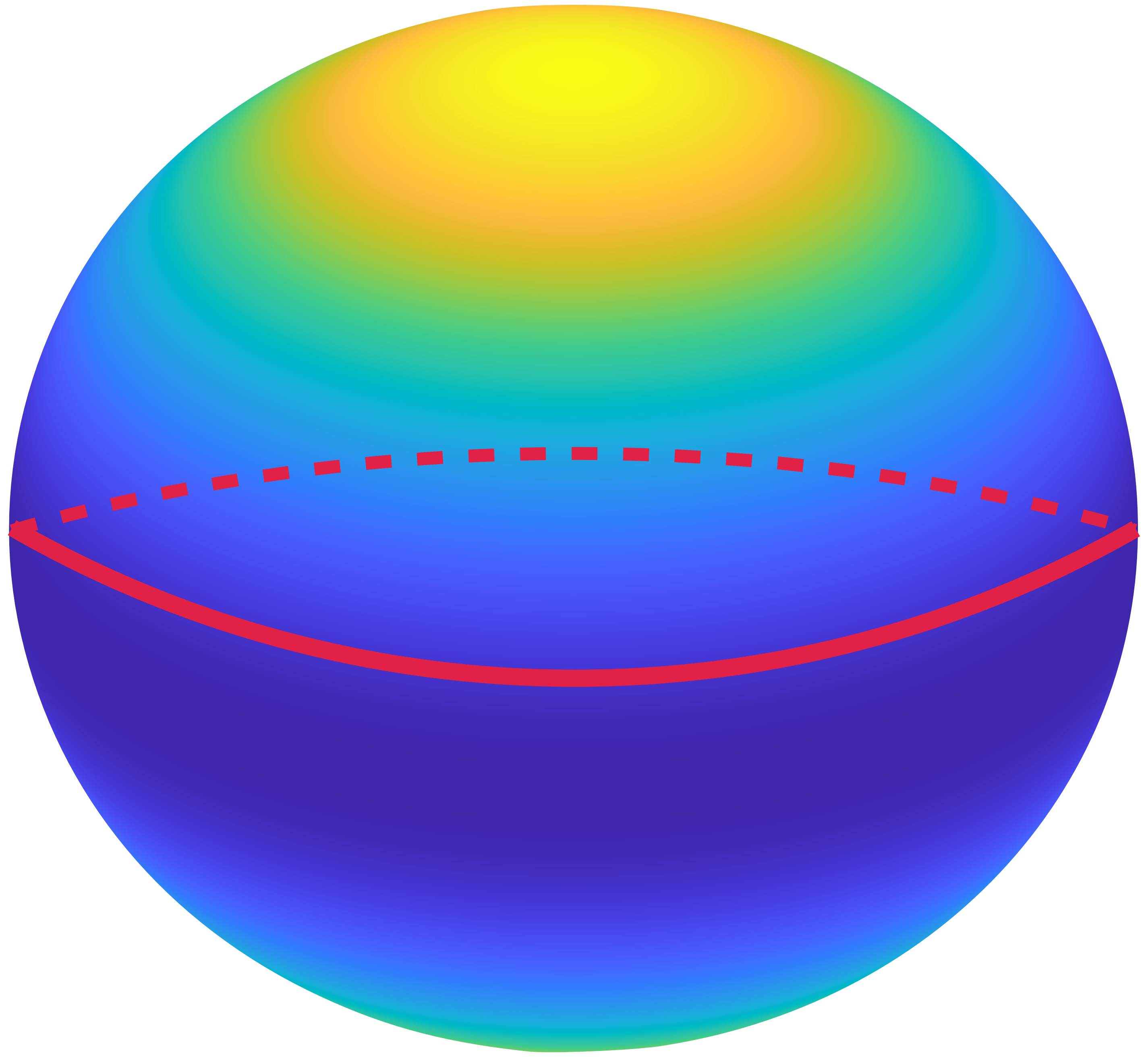
- Alternatively, one can regularize using entropic term: $\tilde{\mathscr{L}}(\nu) = \mathscr{L}(\nu) + \lambda H(\nu)$.
- Noisy dynamics creates Wasserstein contraction via Log-Sobolev Inequality, leading to [Nitanda'24]:
$$\mathscr{E}(\nu_t^{(m)}, \nu_t)^2 \lesssim \frac{1}{m} + \exp(-2\alpha_m \lambda t)\mathscr{E}(\nu_0^{(m)}, \nu_*)$$
  - Here, α_m is the LSI of minimiser, of order $\alpha_m \simeq \exp(-\Theta(m/\lambda))$.
  - Efficient particle approximation, but cursed iteration complexity $T = O\left(\frac{\log \epsilon^{-1}}{\alpha_m^2 \lambda \epsilon}\right)$.



Log-concavity is 'artificial' — noiseless alternative?

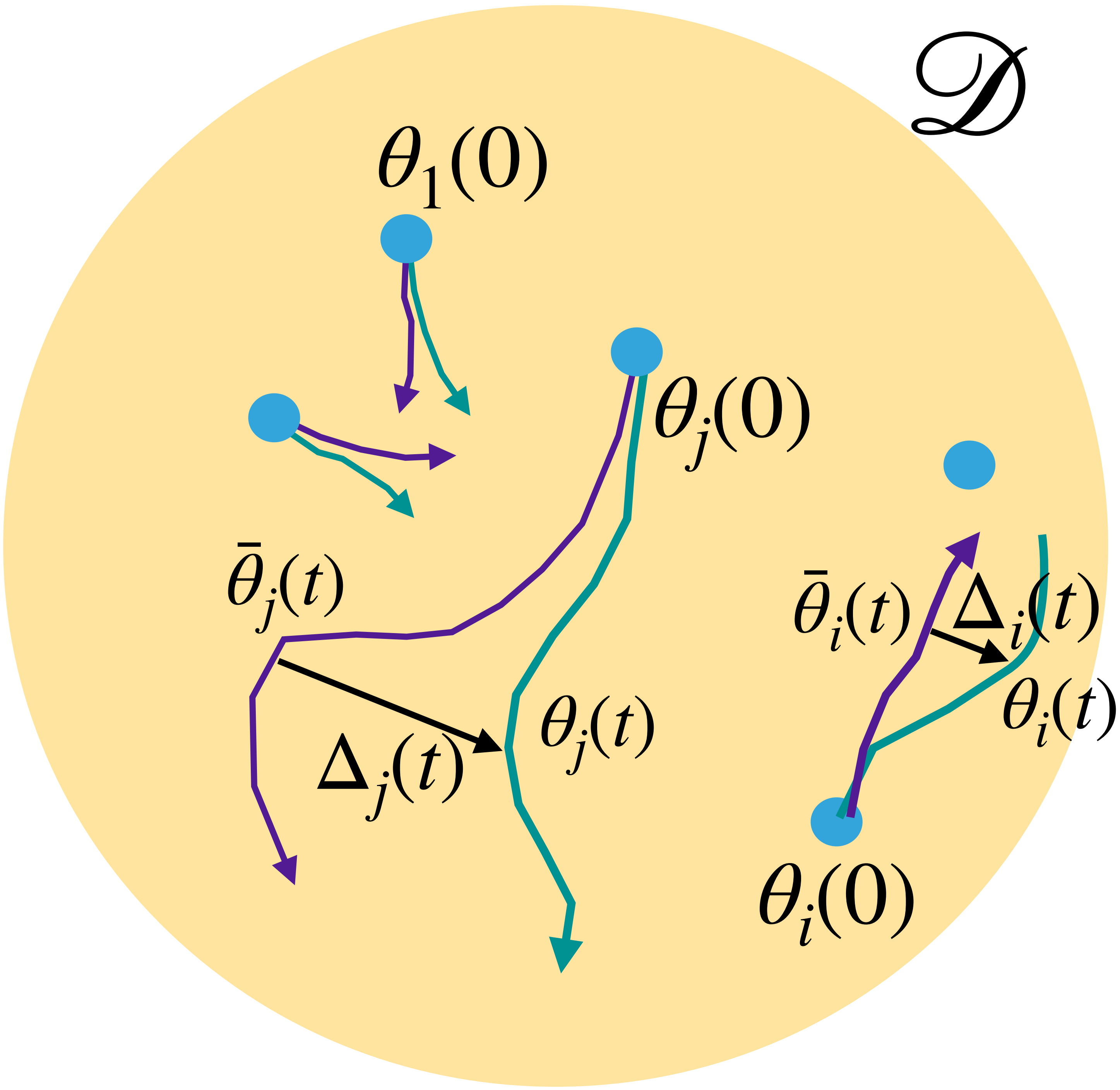
Setup from now on

- Shallow NN architecture with unit-norm 1st-layer weights and fixed 2nd-layer weights: $f(x) = \frac{1}{m} \sum_{j \le m} \rho(\theta_j \cdot x)$, $\theta_j \in \mathbb{S}^{d-1}$.
- Planted setting: $y = f_{\nu^*}(x)$ for some $\nu^* \in \mathscr{P}(\mathbb{S}^{d-1})$.
- Training by Spherical Gradient Flow:
$$\frac{d}{dt}\theta_j = (I - \theta_j \theta_j^\top) \nabla_{\theta_j} \mathbb{E}_x[\,|f(x) - f_{\nu^*}(x)|^2]$$



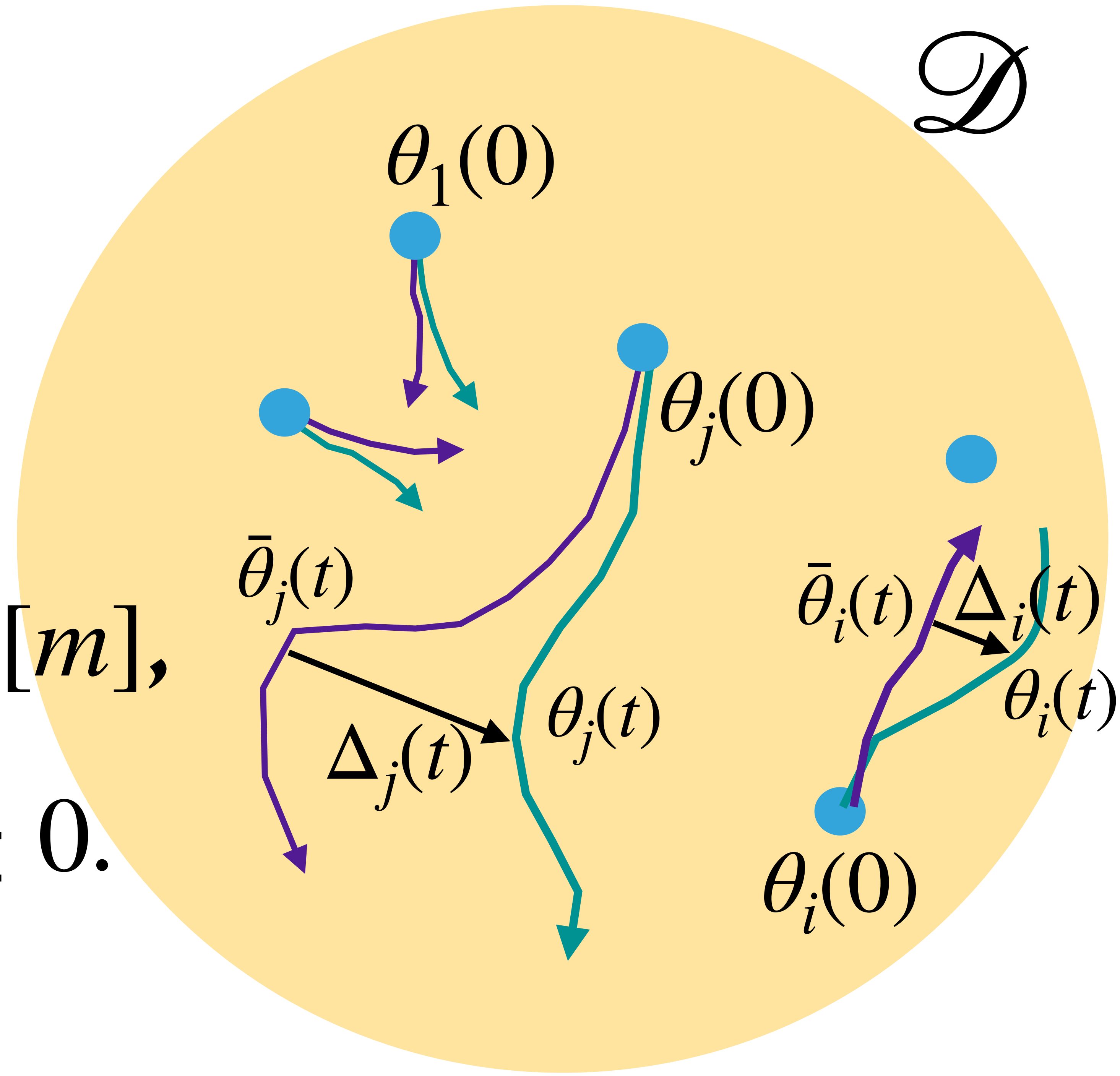
Dissecting the Coupling Dynamics

- In regression, instantaneous potential writes
$U(\theta;\rho) = -F(\theta) + \int K(\theta,\theta')d\rho(\theta')$, with
$F(\theta) = \mathbb{E}[y\rho(\theta\cdot x)]$, $K(\theta,\theta') = \mathbb{E}[\rho(\theta\cdot x)\rho(\theta'\cdot x)]$.



Dissecting the Coupling Dynamics

- In regression, instantaneous potential writes $U(\theta;\rho) = -F(\theta) + \int K(\theta,\theta')d\rho(\theta')$, with $F(\theta) = \mathbb{E}[y\rho(\theta\cdot x)]$, $K(\theta,\theta') = \mathbb{E}[\rho(\theta\cdot x)\rho(\theta'\cdot x)]$.
- Define the **local Hessians** $D_i(t) = \nabla^2_\theta U(\bar\theta_i(t);\nu_t) \in \mathbb{R}^{d\times d}$**,** $i \in [m]$**,**
- and the **interaction Hessians** $H_{i,j}(t) = \nabla_\theta \nabla_{\theta'} K(\bar\theta_i(t), \bar\theta_j(t)) \succeq 0.$

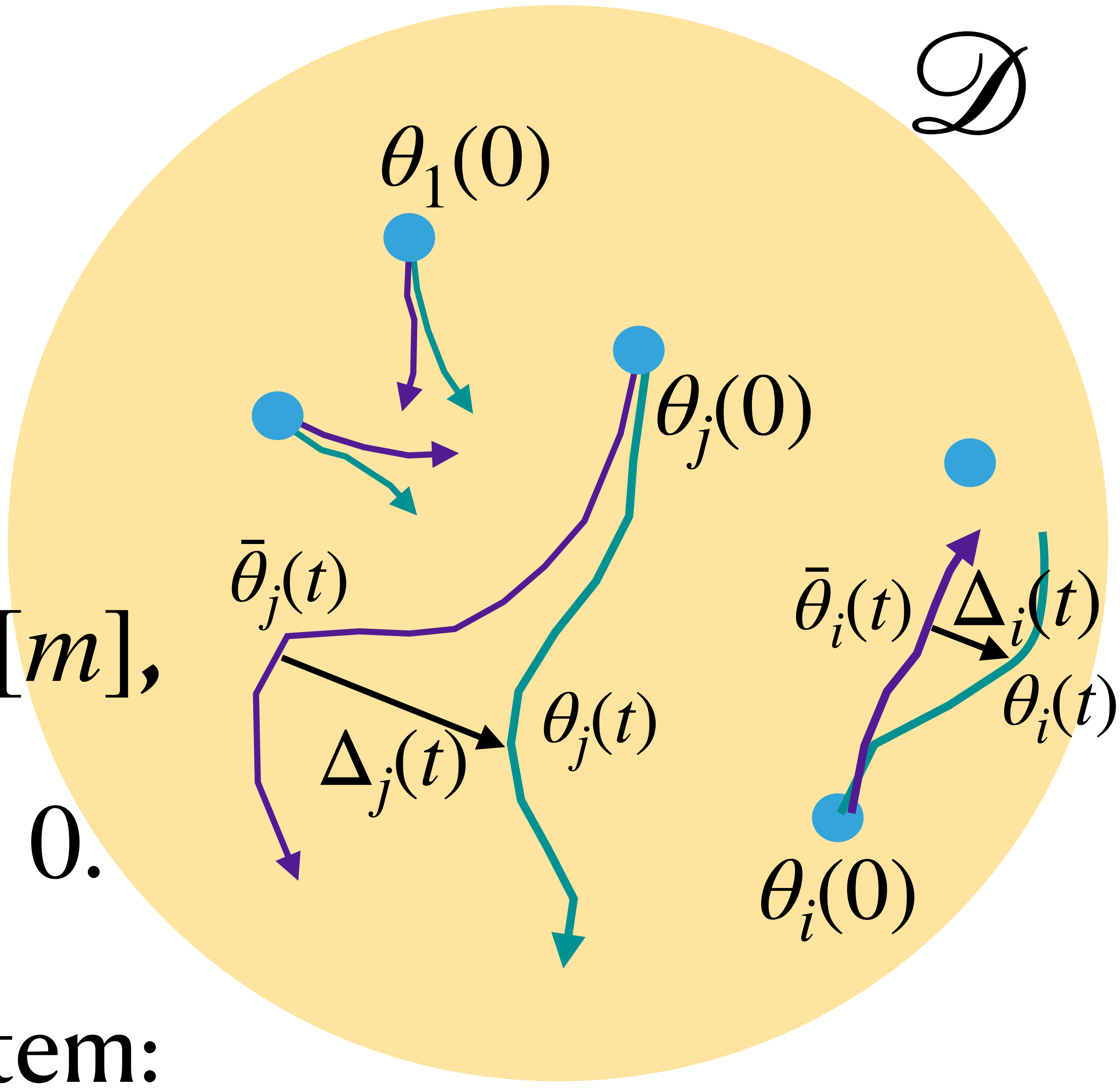


∇, ∇^2 : Spherical Gradient/ Hessian

Dissecting the Coupling Dynamics

- In regression, instantaneous potential writes $U(\theta;\rho) = -F(\theta) + \int K(\theta,\theta')d\rho(\theta')$, with $F(\theta) = \mathbb{E}[y\rho(\theta\cdot x)]$, $K(\theta,\theta') = \mathbb{E}[\rho(\theta\cdot x)\rho(\theta'\cdot x)]$.
- Define the **local Hessians** $D_i(t) = \nabla^2_\theta U(\bar{\theta}_i(t);\nu_t) \in \mathbb{R}^{d\times d}$, $i \in [m]$,
- and the **interaction Hessians** $H_{i,j}(t) = \nabla_\theta \nabla_{\theta'} K(\bar{\theta}_i(t), \bar{\theta}_j(t)) \succeq 0$.
- Coupling errors $\Delta_i(t)$ follow their own particle interaction system:

$$\frac{d}{dt}\Delta_i(t) = D_i(t)\Delta_i(t) - \mathbb{E}_j[H_{i,j}\Delta_j(t)] + O(\|\Delta_i\|^2) + O(1/\sqrt{m})\,.$$



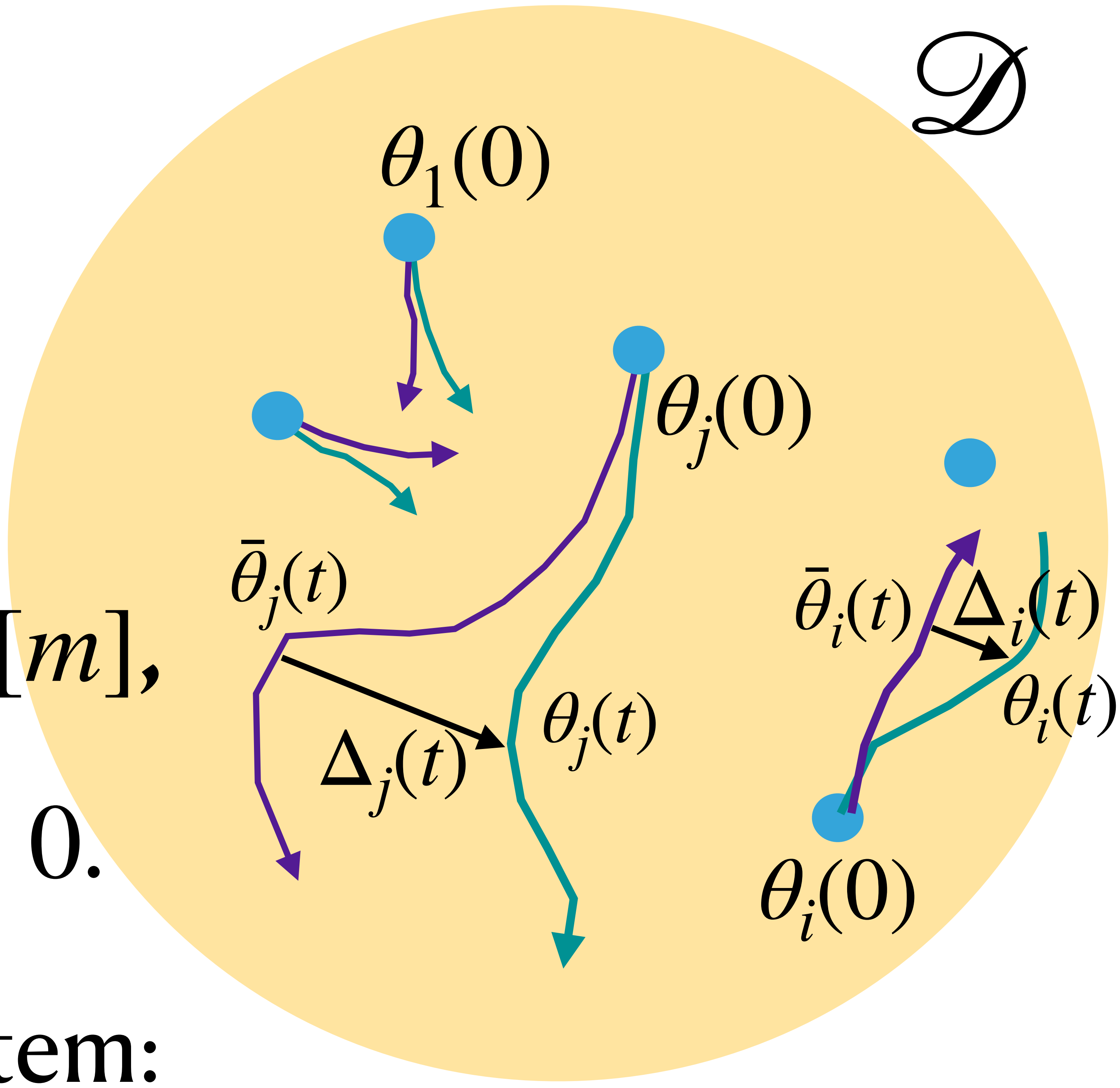
∇, ∇^2 : Spherical Gradient/ Hessian

Dissecting the Coupling Dynamics

- In regression, instantaneous potential writes
  $U(\theta;\rho) = -F(\theta) + \int K(\theta,\theta')d\rho(\theta')$, with
  $F(\theta) = \mathbb{E}[y\rho(\theta\cdot x)]$, $K(\theta,\theta') = \mathbb{E}[\rho(\theta\cdot x)\rho(\theta'\cdot x)]$.
- Define the **local Hessians** $D_i(t) = \nabla^2_\theta U(\bar{\theta}_i(t);\nu_t) \in \mathbb{R}^{d\times d}$**,** $i \in [m]$**,**
- and the **interaction Hessians** $H_{i,j}(t) = \nabla_\theta \nabla_{\theta'} K(\bar{\theta}_i(t), \bar{\theta}_j(t)) \succeq 0$.
- Coupling errors $\Delta_i(t)$ follow their own particle interaction system:

$$\frac{d}{dt}\Delta_i(t) = \boxed{D_i(t)\Delta_i(t)} - \boxed{\mathbb{E}_j[H_{i,j}\Delta_j(t)]} + \boxed{O(\|\Delta_i\|^2) + O(1/\sqrt{m})}.$$

"External field" Interaction term Source term



∇, ∇^2 : Spherical Gradient/ Hessian

Dissecting the Coupling Dynamics

$$\frac{d}{dt}\Delta_i(t) - D_i(t)\Delta_i(t) = -\mathbb{E}_j[H_{i,j}\Delta_j(t)] + O(\|\Delta_i\|^2) + O(1/\sqrt{m}) := -\mathbb{E}_j[H_{i,j}\Delta_j(t)] + \epsilon_i(t)$$

- **Key challenge**: Local and interaction Hessians do not commute.

Dissecting the Coupling Dynamics

$$\frac{d}{dt}\Delta_i(t) - D_i(t)\Delta_i(t) = -\mathbb{E}_j[H_{i,j}\Delta_j(t)] + O(\|\Delta_i\|^2) + O(1/\sqrt{m}) := -\mathbb{E}_j[H_{i,j}\Delta_j(t)] + \epsilon_i(t)$$

- **Key challenge**: Local and interaction Hessians do not commute.
- Viewing the RHS as the source, from Duhamel we have
  $\Delta_i(t) = \int_0^t J_i(t,s)(-\mathbb{E}_j[H_{i,j}\Delta_j(s)] + \epsilon_i(s))ds$, where $J_i(t,s)$ solves
  $\frac{d}{dt}J_i(t,s) = D_i(t)J_i(t,s)\ ,\ J_i(s,s) = \mathsf{P}^{\mathbb{S}}_{\bar{\theta}_i(s)} \Rightarrow J_i(t,s) = \exp\left(\int_s^t D_i(u)du\right).$

Dissecting the Coupling Dynamics

$$\frac{d}{dt}\Delta_i(t) - D_i(t)\Delta_i(t) = -\mathbb{E}_j[H_{i,j}\Delta_j(t)] + O(\|\Delta_i\|^2) + O(1/\sqrt{m}) := -\mathbb{E}_j[H_{i,j}\Delta_j(t)] + \epsilon_i(t)$$

- **Key challenge**: Local and interaction Hessians do not commute.
- Viewing the RHS as the source, from Duhamel we have $\Delta_i(t) = \int_0^t J_i(t,s)(-\mathbb{E}_j[H_{i,j}\Delta_j(s)] + \epsilon_i(s))ds$, where $J_i(t,s)$ solves $\frac{d}{dt}J_i(t,s) = D_i(t)J_i(t,s)$, $J_i(s,s) = \mathsf{P}^{\mathbb{S}}_{\bar{\theta}_i(s)} \Rightarrow J_i(t,s) = \exp\left(\int_s^t D_i(u)du\right)$.
- **Local stability** matrix $J_i(t,s)$: how a perturbation of $\bar{\theta}_i(s)$ (neuron i's position at time s) affects its position $\bar{\theta}_i(t)$ at future time t.

Dissecting the Coupling Dynamics

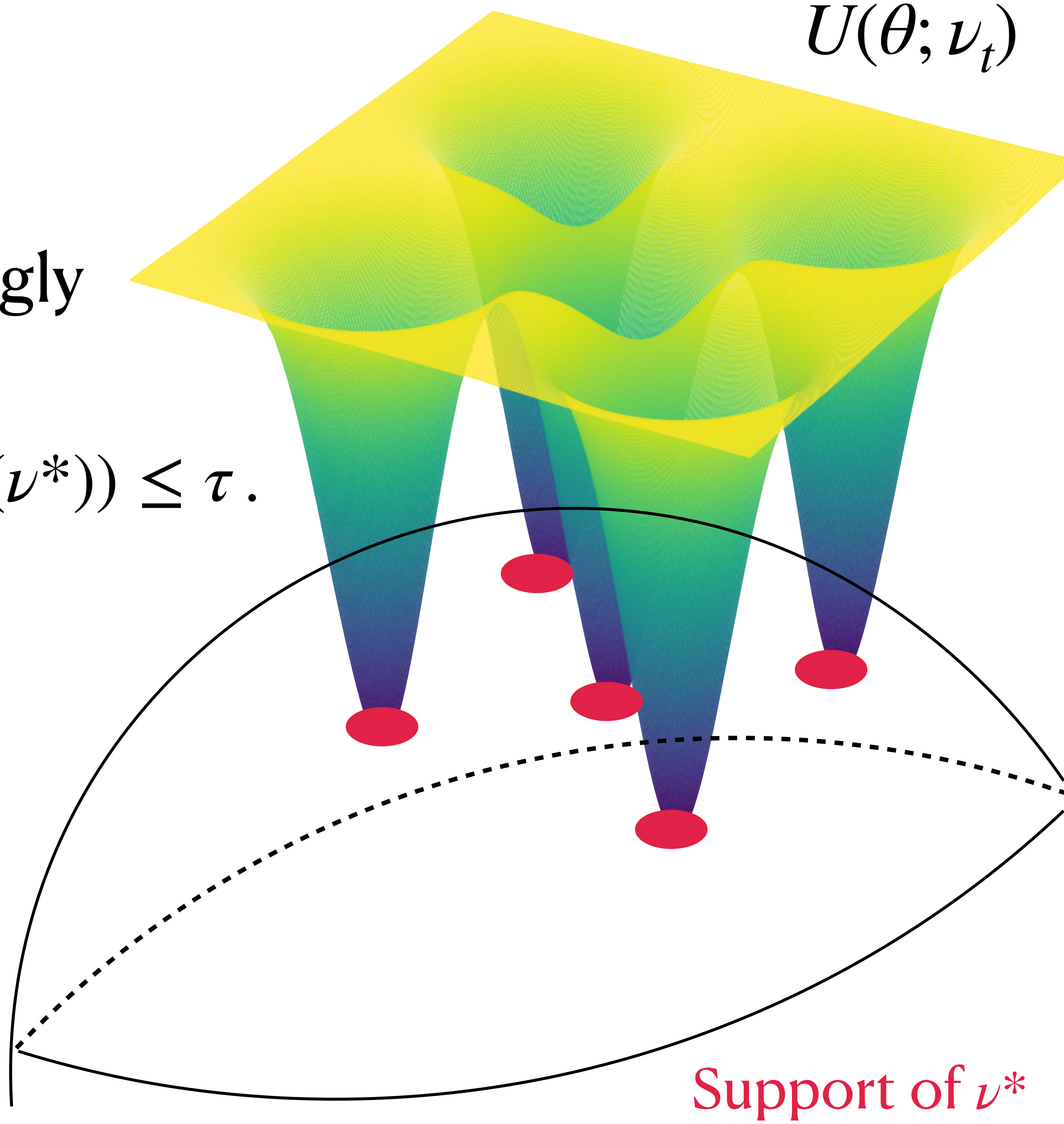
$$\frac{d}{dt}\Delta_i(t) - D_i(t)\Delta_i(t) = -\mathbb{E}_j[H_{i,j}\Delta_j(t)] + O(\|\Delta_i\|^2) + O(1/\sqrt{m}) := -\mathbb{E}_j[H_{i,j}\Delta_j(t)] + \epsilon_i(t)$$

- **Key challenge**: Local and interaction Hessians do not commute.
- Viewing the RHS as the source, from Duhamel we have
  $\Delta_i(t) = \int_0^t J_i(t,s)(-\mathbb{E}_j[H_{i,j}\Delta_j(s)] + \epsilon_i(s))ds$, where $J_i(t,s)$ solves
  $$\frac{d}{dt}J_i(t,s) = D_i(t)J_i(t,s)\ ,\ J_i(s,s) = \mathsf{P}^{\mathbb{S}}_{\bar{\theta}_i(s)} \Rightarrow J_i(t,s) = \exp\left(\int_s^t D_i(u)du\right).$$
- **Local stability** matrix $J_i(t,s)$: how a perturbation of $\bar{\theta}_i(s)$ (neuron i's position at time s) affects its position $\bar{\theta}_i(t)$ at future time t.

Key ingredients to control growth?

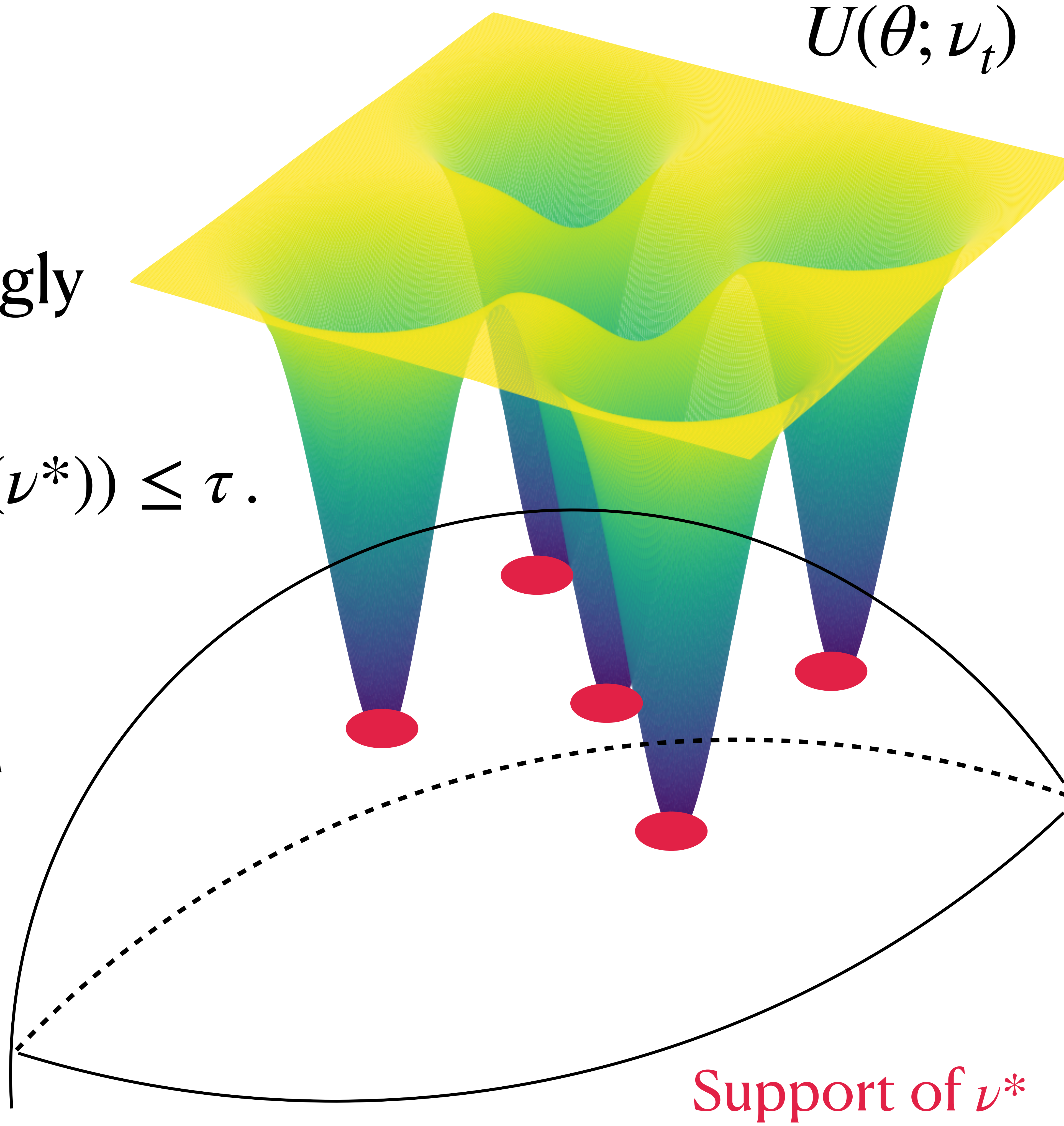
Ingredient 1: Local Strong Convexity

- Let $\xi_t(\theta)$ M-F flow map starting at θ: $\bar{\theta}_i(t) = \xi_t(\theta_i)$.
- Instantaneous potentials $U(\xi_t(\theta); \nu_t)$ are locally strongly convex in a neighborhood of supp(ν^*):
  $\exists \tau > 0;\ \nabla^2_\theta U(\xi_t(\theta); \nu_t) \succeq C\sqrt{\mathscr{L}(\nu_t)}\mathsf{P}^{\mathbb{S}}_\theta$ for $\mathrm{dist}(\xi_t(\theta), \mathrm{supp}(\nu^*)) \leq \tau$.



Ingredient 1: Local Strong Convexity

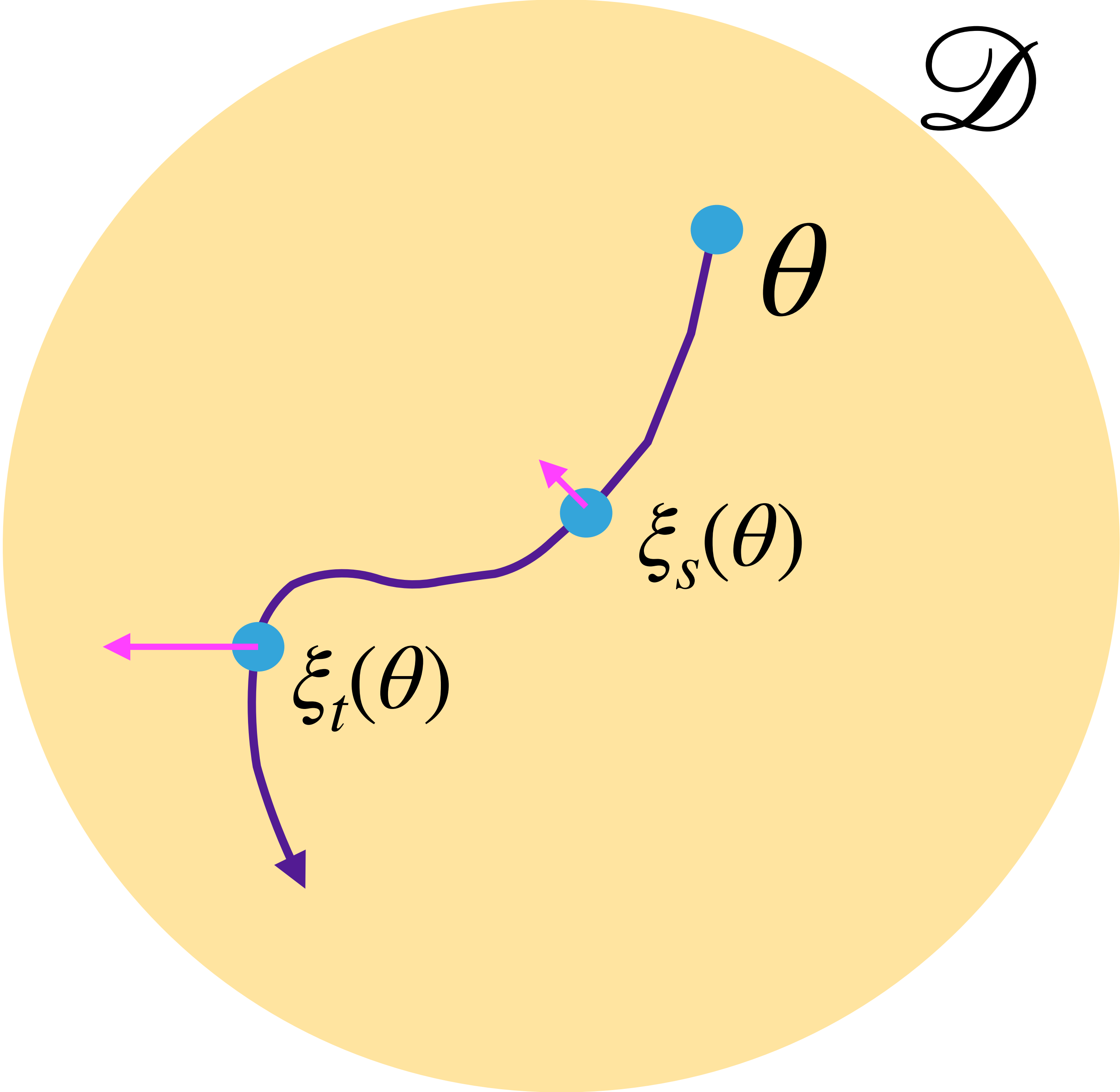
- Let $\xi_t(\theta)$ M-F flow map starting at θ: $\bar{\theta}_i(t) = \xi_t(\theta_i)$.
- Instantaneous potentials $U(\xi_t(\theta); \nu_t)$ are locally strongly convex in a neighborhood of supp(ν^*):
$\exists \tau > 0;\ \nabla^2_\theta U(\xi_t(\theta); \nu_t) \succeq C\sqrt{\mathscr{L}(\nu_t)}\mathsf{P}^{\mathbb{S}}_\theta$ for $\mathrm{dist}(\xi_t(\theta), \mathrm{supp}(\nu^*)) \leq \tau$.
- Implies that ν^* is atomic in current formulation.
- Also exploited in [Chizat'19] [Chen et al.'20] to obtain uniform-in-time, asymptotic (in m), PoC.



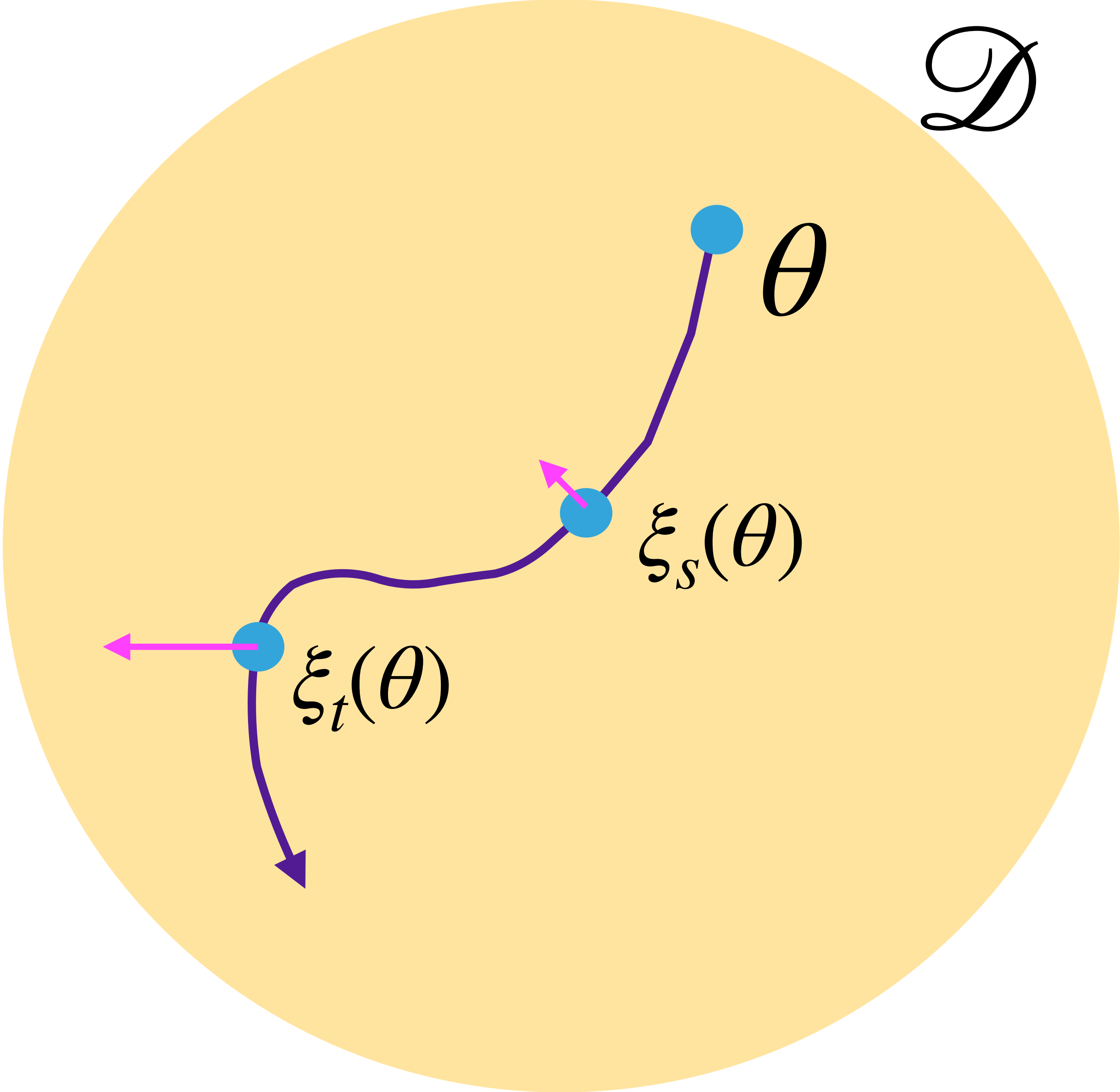
Ingredient 2: Stability

- Local stability matrix now defined for any initial condition:

$$J_\theta(t,s) := \exp\left(\int_s^t \nabla^2 U(\xi_u(\theta);\nu_u)du\right).$$



Ingredient 2: Stability



- Local stability matrix now defined for any initial condition:

$$J_\theta(t,s) := \exp\left(\int_s^t \nabla^2 U(\xi_u(\theta);\nu_u)du\right).$$

- For a desired convergence time T, we assume:

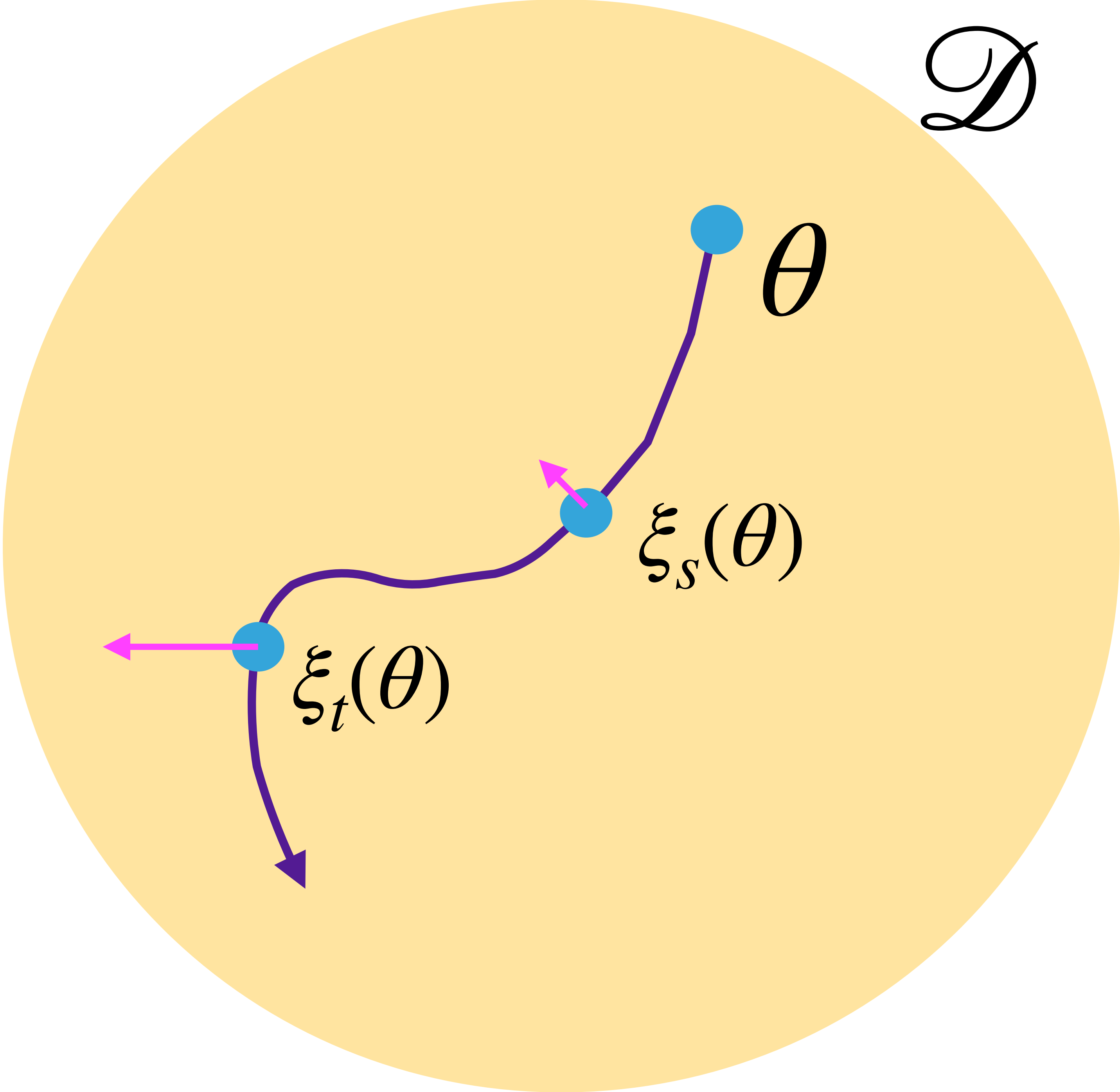
1. *Uniform* Stability: $\sup_{s\le t\le T,\theta} \|J_\theta(t,s)\| = \mathrm{poly}(d,T)$,

2. *Average* Stability far from convergence:

$$\sup_{s\le t\le T,\theta'} \mathbb{E}_\theta[\|J_\theta(t,s)H_{\theta,\theta'}(s)\| \cdot \mathbf{1}(\mathrm{dist}(\xi_t(\theta),\mathrm{supp}(\nu^*)) > \tau)] \lesssim \mathrm{poly}(\tau^{-1})/T$$

.

Ingredient 2: Stability



- Local stability matrix now defined for any initial condition:

$$J_\theta(t, s) := \exp\left(\int_s^t \nabla^2 U(\xi_u(\theta); \nu_u) du\right).$$

- For a desired convergence time T, we assume:

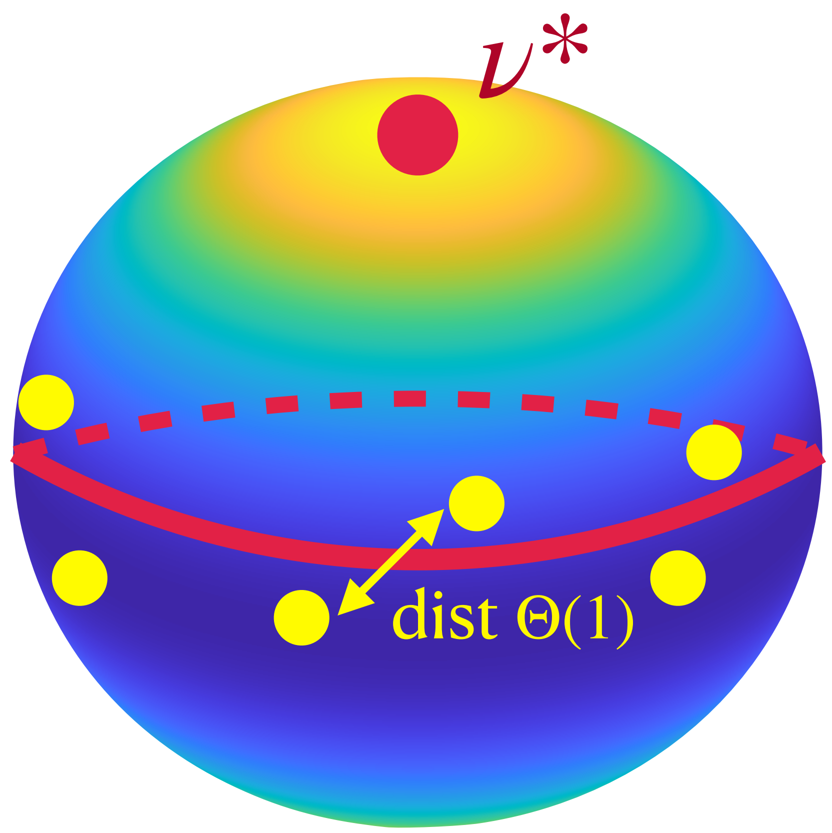
1. *Uniform* Stability: $\sup_{s \le t \le T, \theta} \|J_\theta(t, s)\| = \mathrm{poly}(d, T)$,

"Self-concordance" property:
sharpness $\|D_\theta(t)\| \lesssim \|\nabla U(\theta_t, \nu_t)\|$

2. *Average* Stability far from convergence:

$$\sup_{s \le t \le T, \theta'} \mathbb{E}_\theta[\|J_\theta(t, s) H_{\theta,\theta'}(s)\| \cdot \mathbf{1}(\mathrm{dist}(\xi_t(\theta), \mathrm{supp}(\nu^*)) > \tau)] \lesssim \mathrm{poly}(\tau^{-1})/T$$
.

Neurons 'dispersed' before converging



Main Result

- Under local strong convexity and stability, we have quantitative PoC:
- **Theorem** [GWB'25], informal: Assume ***LSC*** and ***Stability*** over horizon T, plus technical regularity assumptions. Then whp $\mathscr{E}(\nu_T, \nu_T^{(m)}) \lesssim \dfrac{\text{poly}(d, T)}{\sqrt{m}}$.

Main Result

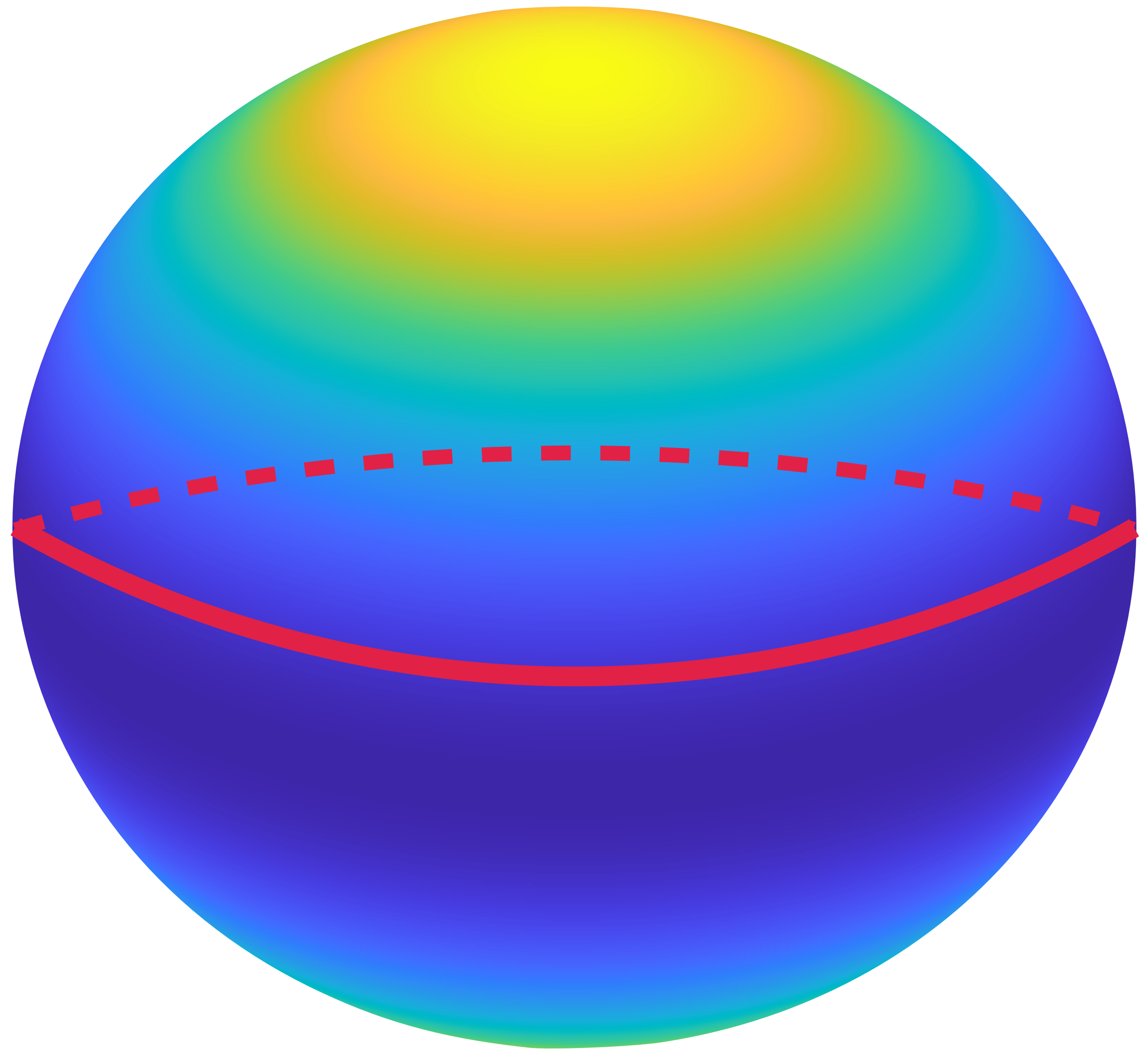
- Under local strong convexity and stability, we have quantitative PoC:
- **Theorem** [GWB'25], informal: Assume ***LSC*** and ***Stability*** over horizon T, plus technical regularity assumptions. Then whp $\mathscr{E}(\nu_T, \nu_T^{(m)}) \lesssim \dfrac{\mathrm{poly}(d, T)}{\sqrt{m}}$.
  - If MF converges at horizon $T = \mathrm{poly}(d)$, then poly-sized finite net does too.
  - Result extends to empirical risk with additional $O(\sqrt{d/n})$ term.

Main Result

- Under local strong convexity and stability, we have quantitative PoC:
- **Theorem** [GWB'25], informal: Assume ***LSC*** and ***Stability*** over horizon T, plus technical regularity assumptions. Then whp $\mathscr{E}(\nu_T, \nu_T^{(m)}) \lesssim \dfrac{\mathrm{poly}(d, T)}{\sqrt{m}}$.
  - If MF converges at horizon $T = \mathrm{poly}(d)$, then poly-sized finite net does too.
  - Result extends to empirical risk with additional $O(\sqrt{d/n})$ term.

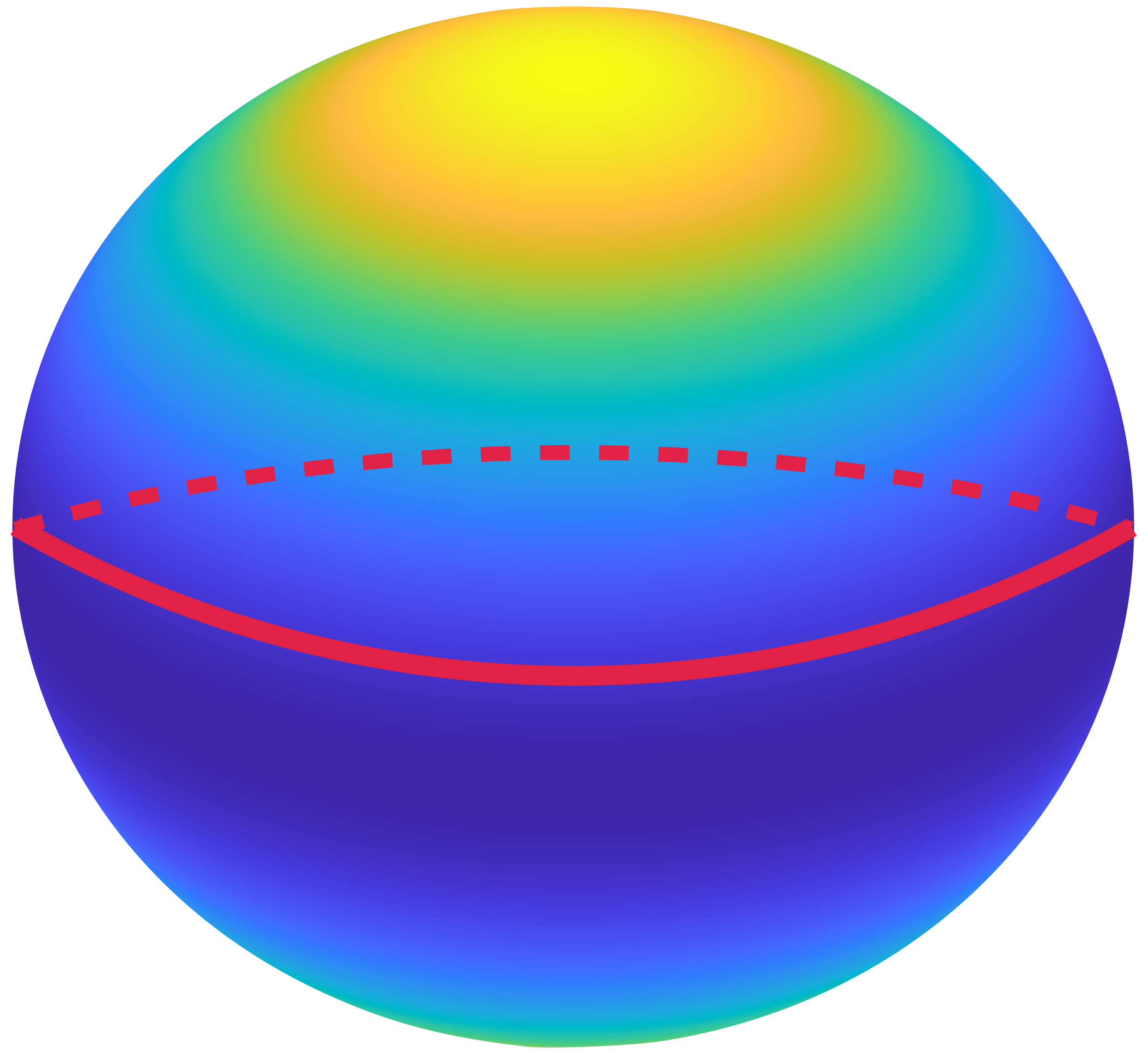
When can we verify these assumptions?

Application: Single-Index Models

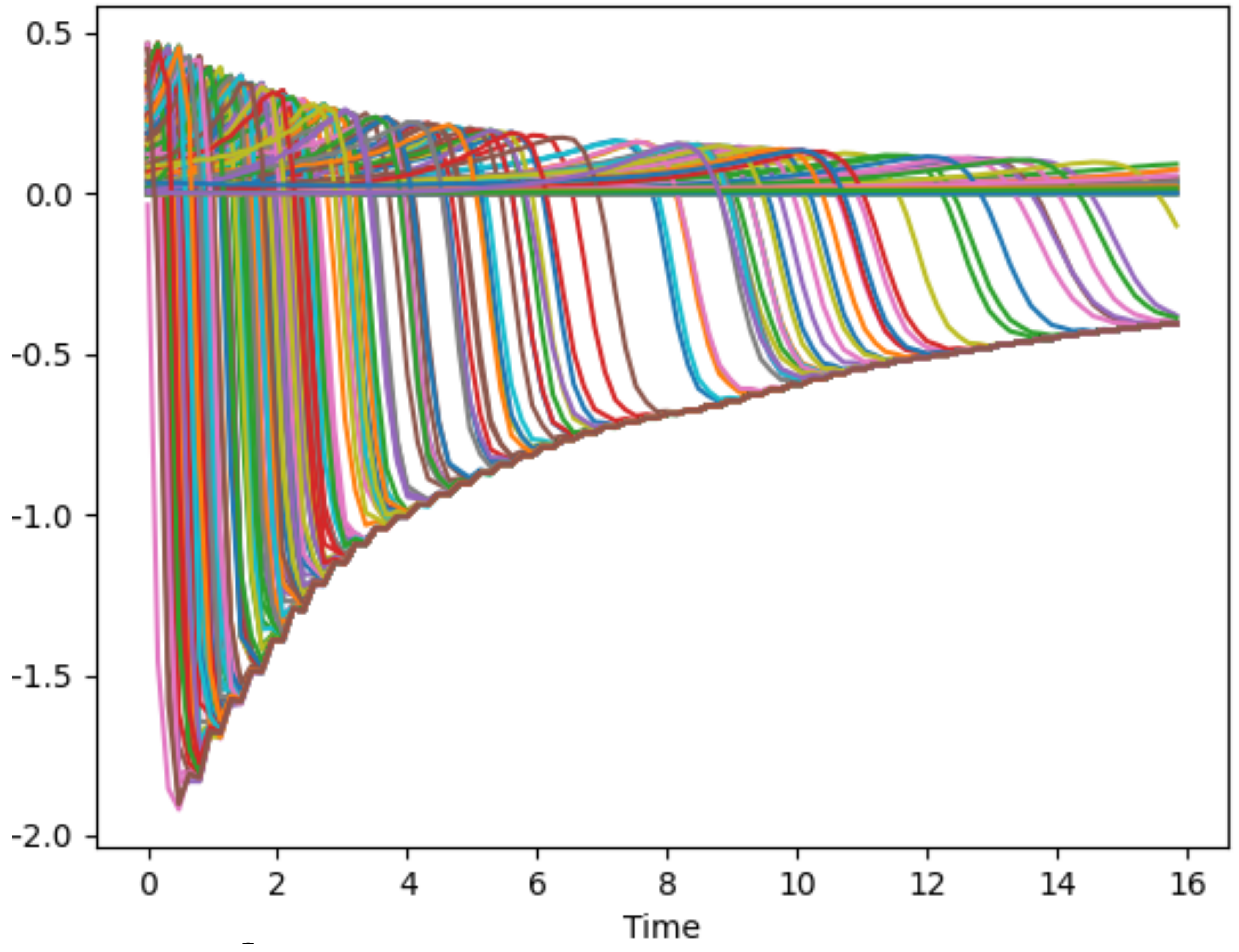


- Well-specified, Gaussian setting: $x \sim \mathcal{N}(0, I_d)$, $y = \rho(\theta^* \cdot x) + w$,
- ρ :even function with Information-Exponent $k^\star \geq 4$.
- **Theorem** [GWB'25]: Let $f_{\nu_t^{(m)}}(x) = \frac{1}{m} \sum_{j \leq m} \rho(\theta_j(t) \cdot x)$ trained with L2-loss on n iid samples for $T = O(\delta^{-k^\star+1} d^{k^\star/2-1})$. Then if $m \gtrsim d^{13k^\star}$, $n \gtrsim d^{11k^\star}$, we have whp $\|f_{\nu_T^{(m)}} - f^*\|^2 = O(\delta^2)$.

Application: Single-Index Models



- Well-specified, Gaussian setting: $x \sim \mathcal{N}(0, I_d)$, $y = \rho(\theta^* \cdot x) + w$,
- ρ :even function with Information-Exponent $k^\star \geq 4$.
- **Theorem** [GWB'25]: Let $f_{\nu_t^{(m)}}(x) = \frac{1}{m}\sum_{j \leq m} \rho(\theta_j(t) \cdot x)$ trained with L2-loss on n iid samples for $T = O(\delta^{-k^\star+1} d^{k^\star/2-1})$. Then if $m \gtrsim d^{13k^\star}, n \gtrsim d^{11k^\star}$, we have whp $\|f_{\nu_T^{(m)}} - f^*\|^2 = O(\delta^2)$.
- $k^\star = 2$ violates current stability assumptions; covered in [Damian et al,'22], [Mahankali et al].
- Exploits *self-concordance* of SIM landscapes: $\|\nabla^2 U(\theta)\| \simeq (\theta \cdot \theta^*)^{-1} \|\nabla U(\theta)\|$.



Proof Overview

$$\frac{d}{dt}\Delta_i(t) = D_i(t)\Delta_i(t) - \mathbb{E}_j[H_{i,j}\Delta_j(t)] + \epsilon_i(t)\,.$$

Self-interaction: driven by local Hessian $\nabla^2 U(\theta; \nu_t)$

Interactions: driven by neuron repulsion kernel $\nabla_\theta \nabla_{\theta'} K(\theta, \theta')$

Source term: at Monte-Carlo scale $O(1/\sqrt{m})$

Proof Overview

$$\frac{d}{dt}\Delta_i(t) = D_i(t)\Delta_i(t) - \mathbb{E}_j[H_{i,j}\Delta_j(t)] + \epsilon_i(t)\,.$$

Self-interaction: driven by local Hessian $\nabla^2 U(\theta; \nu_t)$
Interactions: driven by neuron repulsion kernel $\nabla_\theta \nabla_{\theta'} K(\theta, \theta')$
Source term: at Monte-Carlo scale $O(1/\sqrt{m})$

- Ignoring neuron interactions: exploit uniform stability

$$\sup_{s \le t \le T, \theta} \|J_\theta(t, s)\| = \mathrm{poly}(d, T)\,.$$

- Ignoring self-interactions: PSD kernel contracts $\mathbb{E}_i \|\Delta_i(t)\|^2$.

Proof Overview

$$\frac{d}{dt}\Delta_i(t) = D_i(t)\Delta_i(t) - \mathbb{E}_j[H_{i,j}\Delta_j(t)] + \epsilon_i(t)\,.$$

Self-interaction: driven by local Hessian $\nabla^2 U(\theta; \nu_t)$
Interactions: driven by neuron repulsion kernel $\nabla_\theta \nabla_{\theta'} K(\theta, \theta')$
Source term: at Monte-Carlo scale $O(1/\sqrt{m})$

- Ignoring neuron interactions: exploit uniform stability

$$\sup_{s \le t \le T, \theta} \|J_\theta(t, s)\| = \mathrm{poly}(d, T)\,.$$

- Ignoring self-interactions: PSD kernel contracts $\mathbb{E}_i \|\Delta_i(t)\|^2$.
- **Main challenge**: interplay between these terms.
- Coupling dynamics driven by sparse fluctuations → 'natural' metric is $W_1(\nu_t^{(m)}, (\nu_t)^{(m)}) \le \mathbb{E}_i \|\Delta_i(t)\|$.

Proof Overview

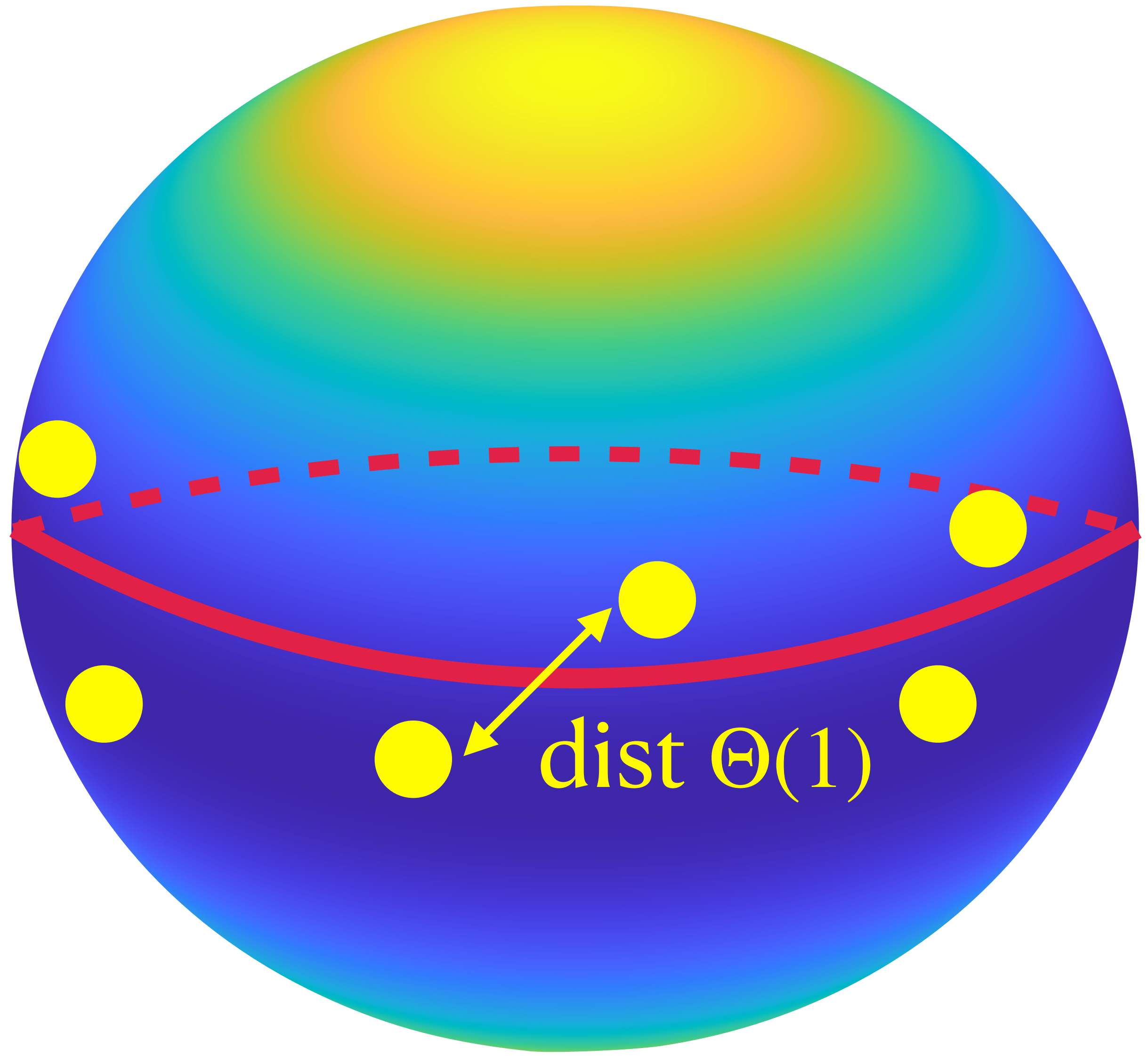
$$\frac{d}{dt}\Delta_i(t) = D_i(t)\Delta_i(t) - \mathbb{E}_j[H_{i,j}\Delta_j(t)] + \epsilon_i(t)\,.$$

Self-interaction: driven by local Hessian $\nabla^2 U(\theta;\nu_t)$
Interactions: driven by neuron repulsion kernel $\nabla_\theta \nabla_{\theta'} K(\theta,\theta')$
Source term: at Monte-Carlo scale $O(1/\sqrt{m})$

- Ignoring neuron interactions: exploit uniform stability

$$\sup_{s\le t\le T,\theta} \|J_\theta(t,s)\| = \mathrm{poly}(d,T)\,.$$

- Ignoring self-interactions: PSD kernel contracts $\mathbb{E}_i\|\Delta_i(t)\|^2$.
- **Main challenge**: interplay between these terms.
- Coupling dynamics driven by sparse fluctuations → 'natural' metric is $W_1(\nu_t^{(m)}, (\nu_t)^{(m)}) \le \mathbb{E}_i\|\Delta_i(t)\|$.
- Near initialisation, dynamics are driven by local term $D_i(t)$, thanks to the average stability assumption (neurons are dispersed before converging).



Proof Overview

$$\frac{d}{dt}\Delta_i(t) = D_i(t)\Delta_i(t) - \mathbb{E}_j[H_{i,j}\Delta_j(t)] + \epsilon_i(t)\,.$$

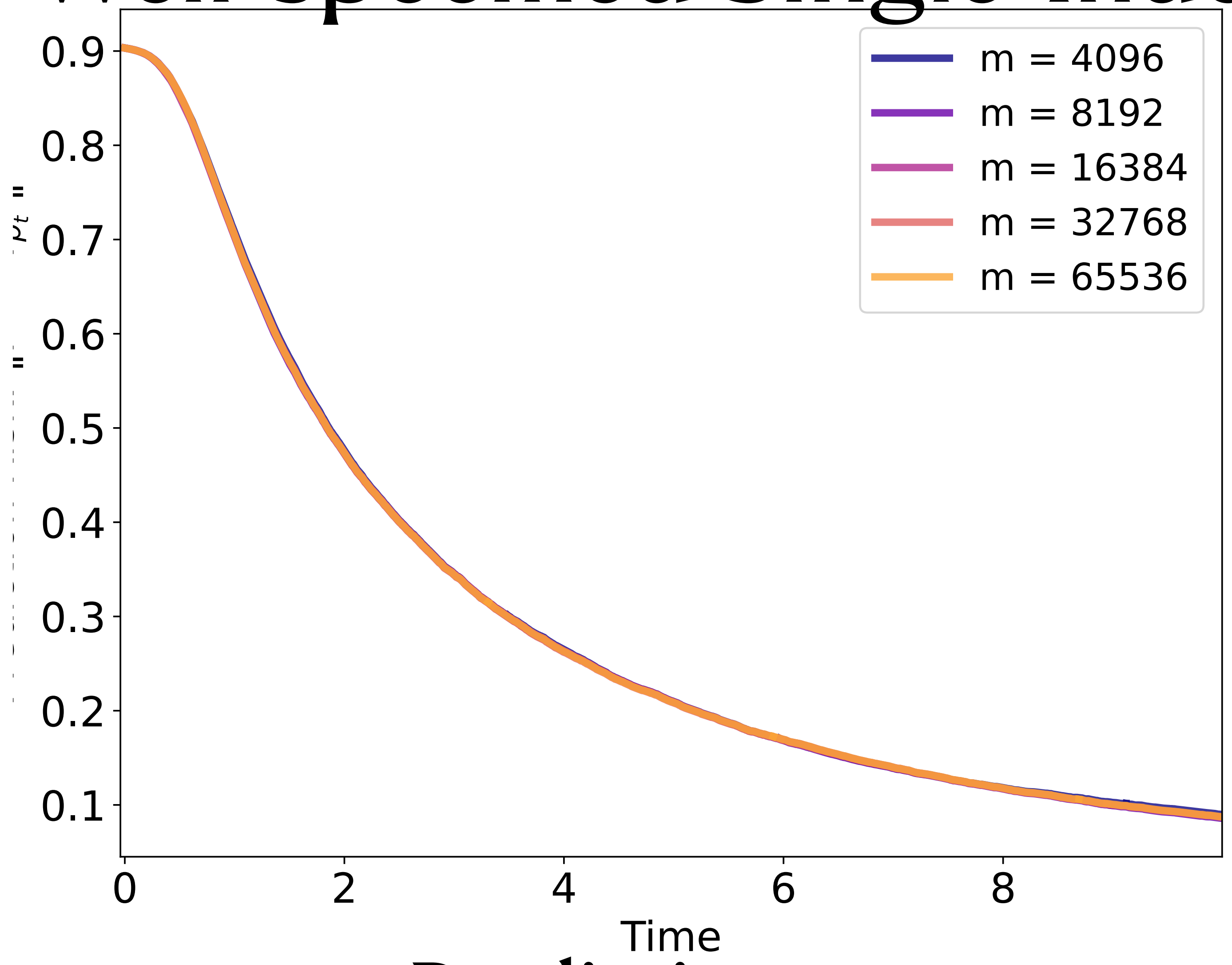
Self-interaction: driven by local Hessian $\nabla^2 U(\theta; \nu_t)$
Interactions: driven by neuron repulsion kernel $\nabla_\theta \nabla_{\theta'} K(\theta, \theta')$
Source term: at Monte-Carlo scale $O(1/\sqrt{m})$

- Near convergence, dynamics are driven by interaction terms $H_{i,j}(t)$:
- **Balanced Interaction Lemma**: If $\mathbb{E}_i\|\Delta_i(s)\|_1$ is small, then interaction dynamics cannot increase it too much:
  Let $\frac{d}{dt}\Delta = -H\Delta$, and consider eigendecomposition $H(\infty) = \sum_{\lambda \in \Lambda} \lambda P_\lambda$.
  For $t \geq s$, we have $\|\Delta(t)\|_1 \leq \|\Delta(s)\|_1 \sum_\lambda \|P_\lambda\|_\infty = \Theta(|\Lambda|)\|\Delta(s)\|_1$.
- Exploited by designing appropriate potential function $\Phi(t)$ that combines interaction at convergence H_∞ and surrogate quantity of interest $\mathbb{E}_i\|\Delta_i(t)\|$.

Experiments

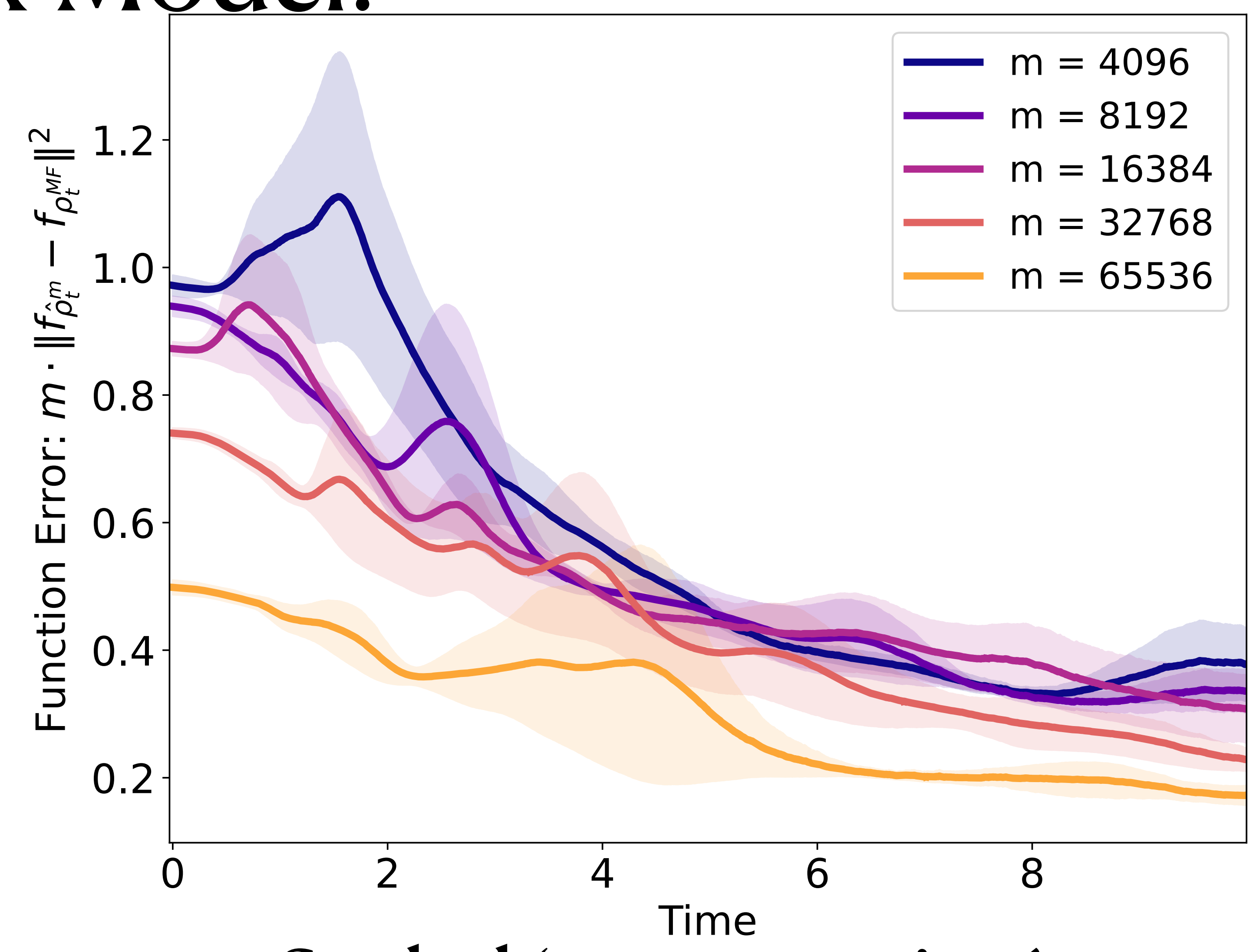
Name	Target Function	Activation/Network Design	LSC?	Symmetric?	J_{avg} assm?
He_4	$\text{He}_4(x^\top e_1)$	$\sigma = \text{He}_4$	Yes	Yes	Yes
Circle	$\mathbb{E}_{w\sim\mathbb{S}^1}\text{He}_4(x^\top w)$	$\sigma = \text{He}_4$	No	Yes	Yes
Misspecified	$0.8\text{He}_4(x^\top e_1) + 0.6\text{He}_6(x^\top e_1)$	$\sigma = \text{He}_4 + \text{He}_6$	No	No	Yes
$\text{Random}_{6,6}$	He_4 link, 6 random teachers in $\mathbb{R}^6$	$\sigma = \text{He}_4$	Yes	No	Yes?
Staircase	$0.25x_1 + 0.75\text{XOR}_4(x_{[4]})$	$\sigma =$ SoftPlus, 2nd layer ± 8	Yes	No	No
XOR_4	$\text{XOR}_4(x_{[4]})$	$\sigma =$ SoftPlus, 2nd layer ± 8	Yes	No	?

- Well-specified Single-Index Model:



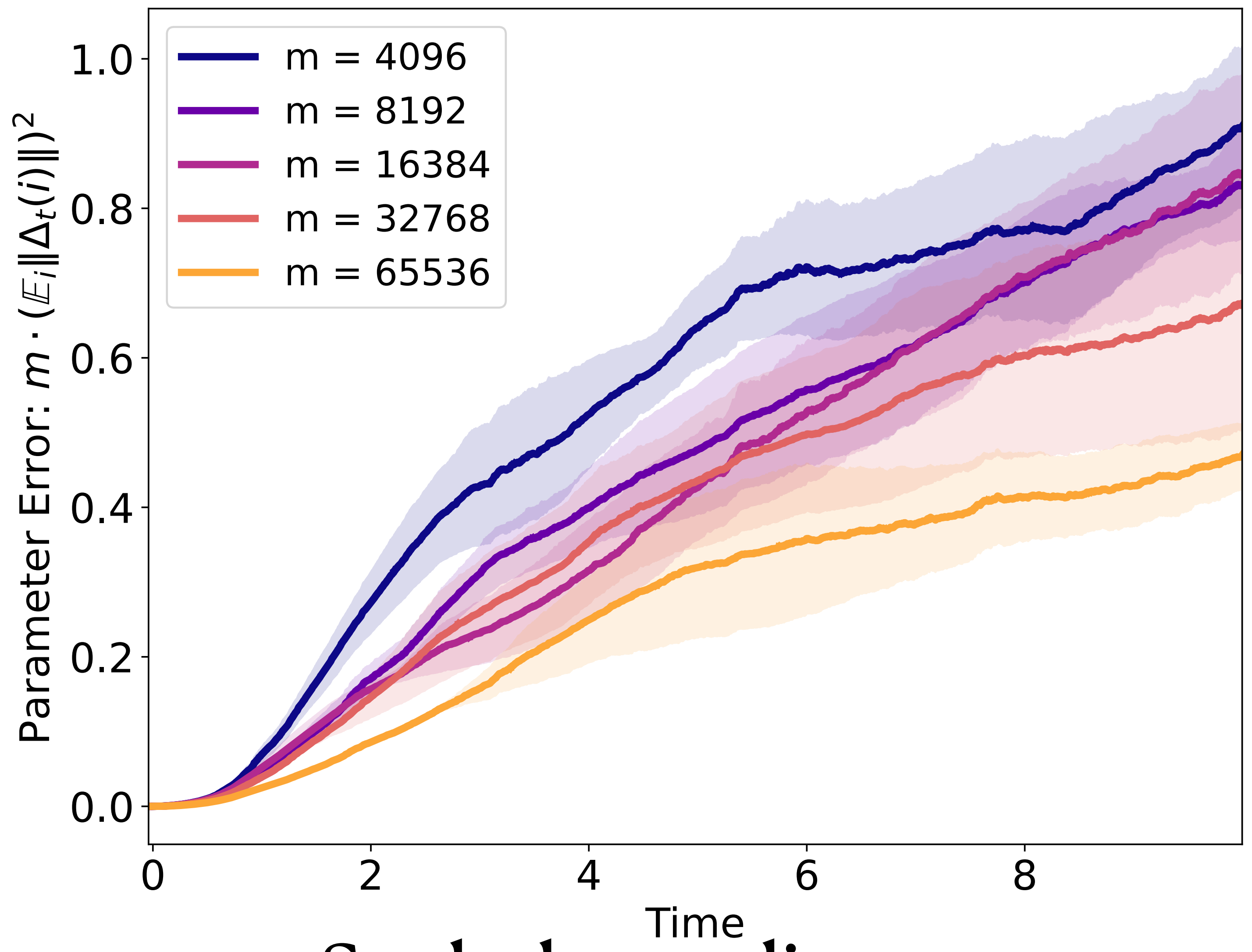
Prediction error

$\mathscr{E}(\nu_t^{(m)}, \nu^*)$



Scaled 'commutation' error

$m\mathscr{E}(\nu_t^{(m)}, (\nu_t)^{(m)})$

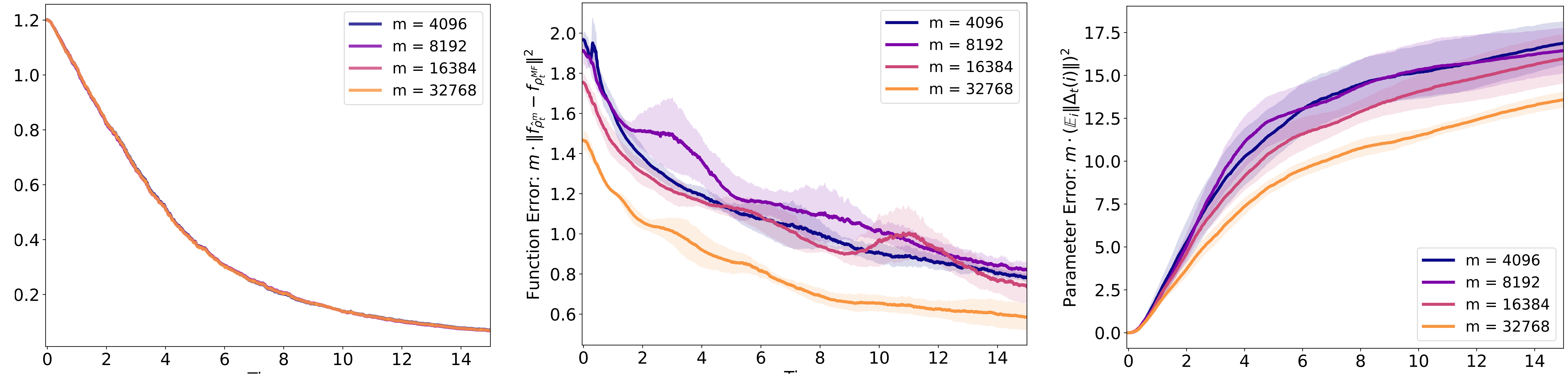


Scaled coupling error

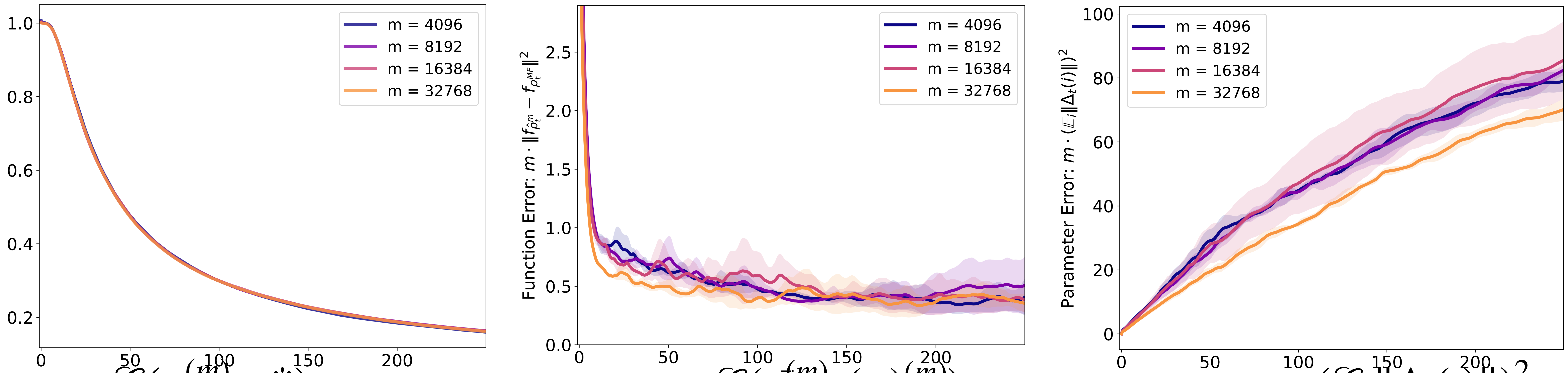
$m(\mathscr{E}_j\|\Delta_j(t)\|)^2$

Experiments

- Misspecified Single-Index Model:

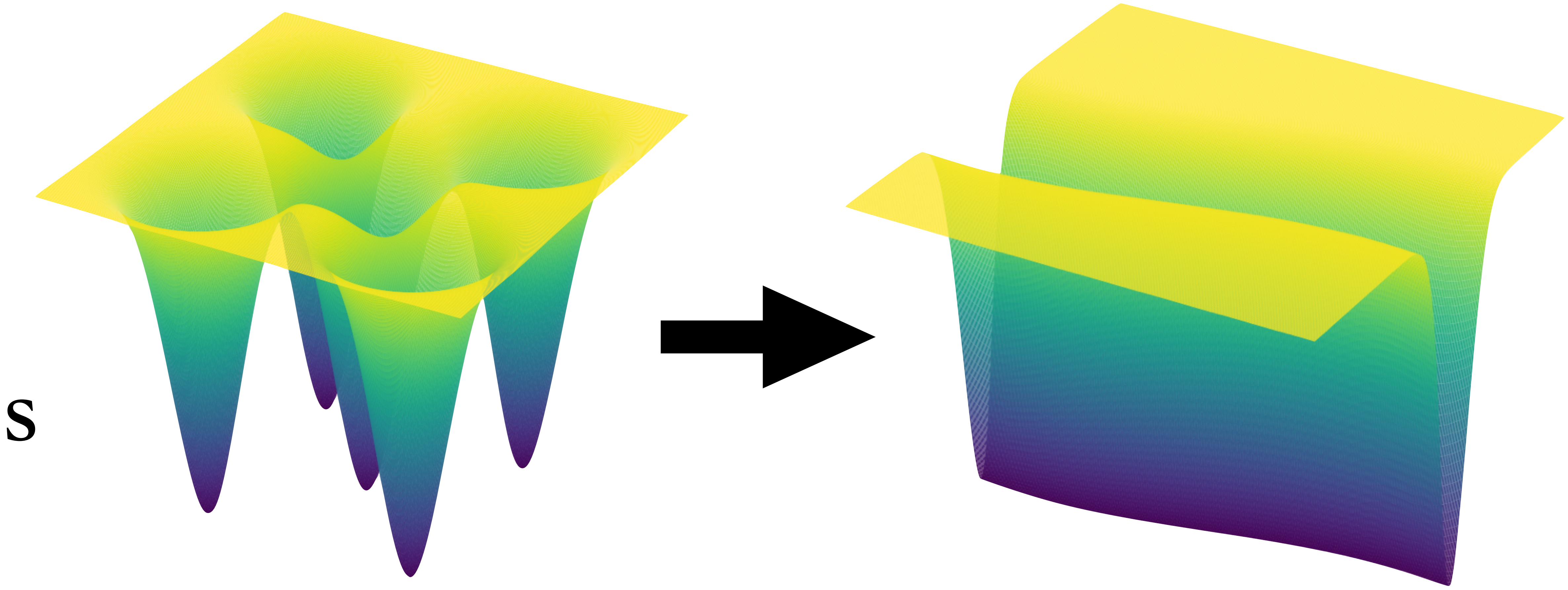


- 4-parity (misspecified Multi-index model):



$\mathscr{E}(\nu_t^{(m)}, \nu^*)$ $\quad m\mathscr{E}(\nu_t^{(m)}, (\nu_t)^{(m)})$ $\quad m(\mathscr{E}_j\|\Delta_j(t)\|)^2$

Next Steps / Questions



- Relaxing LSC to allow mis-specified problems
- Establishing stability properties beyond 'self-concordant' SIM-MIM-type problems? BBP-like?
- Effect of step-size: Links between sharpness and velocity related to central flow [Cohen & Damian et al]?
- Relationship with DMFT analysis of fluctuations [Bordelon et al]?
- Links between PoC and scaling laws, beyond linear models [Paquette et al.]?

Thanks!

References:

- Propagation-of-Chaos in Single-Hidden Layer Neural Networks beyond Logarithmic time, with Denny Wu and Margalit Glasgow, COLT 25.