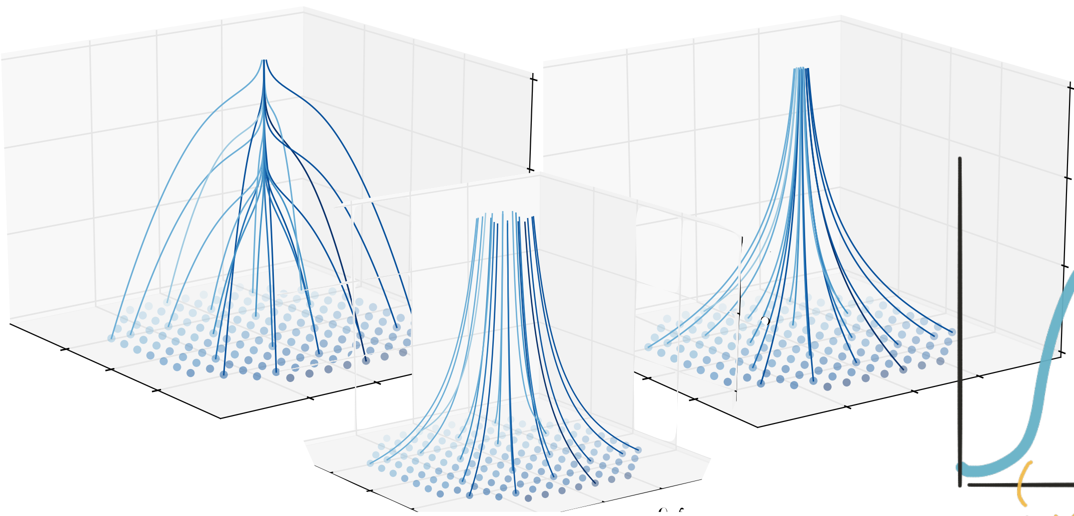
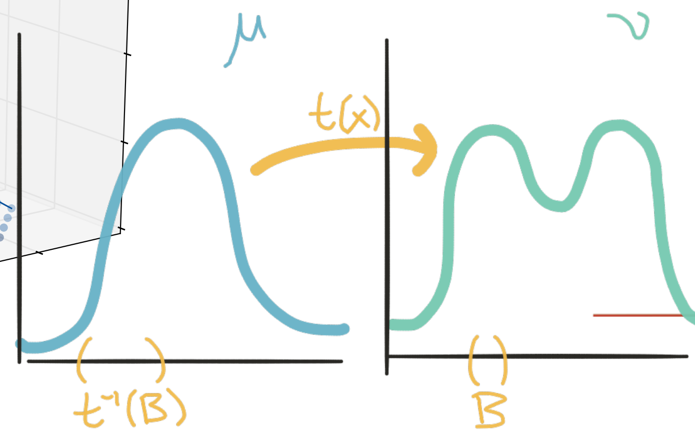
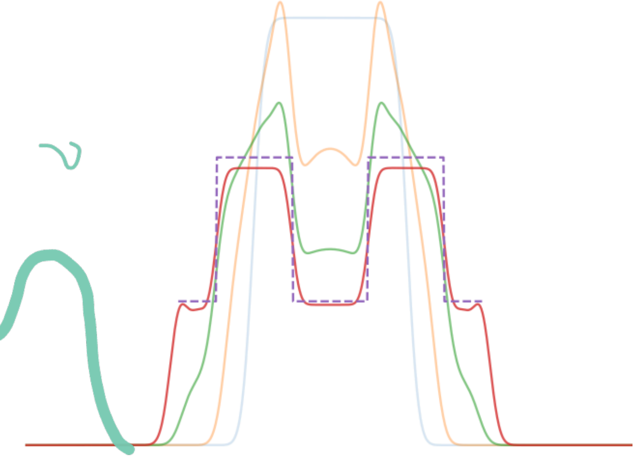




# A blob method for degenerate diffusion and applications to sampling and two layer neural networks.


Katy Craig
University of California, Santa Barbara

joint with José Antonio Carrillo (Oxford), Francesco Patacchini (IFP Energies), Karthik Elamvazhuthi (UCLA), Matt Haberland (Cal Poly), Olga Turanova (Michigan State)



Columbia University, Applied Mathematics Colloquium
November 9th, 2021

# Plan

- Motivation
- Wasserstein gradient flows
- Particle methods (discrete $\leftrightarrow$ continuum)
- Particle method + regularization = blob method for diffusive PDEs
- Numerics

# Sampling/robot coverage algorithms

Consider a target distribution $\bar{\rho} \in \mathscr{P}(\mathbb{R}^d)$.

**Sampling:** How can we choose samples $\{\bar{x}_i\}_{i=1}^N \subseteq \mathbb{R}^d$, so that (with high probability), they accurately represent the desired target distribution?

**Coverage:** How can we program robots to move so that they distribute their locations $\{\bar{x}_i\}_{i=1}^N \subseteq \mathbb{R}^d$ according to $\bar{\rho}$ (deterministically)?

In both cases, we seek to approximate $\bar{\rho}$ by an empirical measure:

$$\bar{\rho}^N := \frac{1}{N} \sum_{i=1}^N \delta_{\bar{x}_i} \xrightarrow{N \to +\infty} \bar{\rho}$$

PDE's can inspire new ways to construct the empirical measure.

# PDEs and sampling/coverage algs

Suppose $\bar{\rho} = e^{-V}$, for $V : \mathbb{R}^d \to \mathbb{R}$ $\lambda$-convex.

**Diffusion:** $\partial_t \rho = \nabla \cdot \left( \rho \nabla \log \left( \rho / \bar{\rho} \right) \right) = \Delta \rho - \nabla \cdot (\rho \nabla \log \bar{\rho})$

$KL(\rho(t), \bar{\rho}) \leq e^{-\lambda t} KL(\rho(0), \bar{\rho})$ [Villani '08,…], $KL(\mu, \nu) = \int \mu \log(\mu/\nu)$

Particle method: $dX_t = \sqrt{2} dB_t - \nabla \log \bar{\rho}(X_t) dt$ [Fo

$$\rho^N(t) := \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i}(t) \xrightarrow{N \to +\infty} \rho(t)$$

**Degenerate diffusion:** $\partial_t \rho = \nabla \cdot \left( \rho \nabla \left( \rho / \bar{\rho} \right) \right)$ =

$KL(\rho(t), \bar{\rho}) \leq e^{-\lambda t} KL(\rho(0), \bar{\rho})$ [Matthes, et al. '09,

Particle method: ?

**Motivation for deg. diff:**

*Sampling*: SVGD, chi-sq.

*PDE:* porous media, swarming, …

*Coverage:* **deterministic** particle method

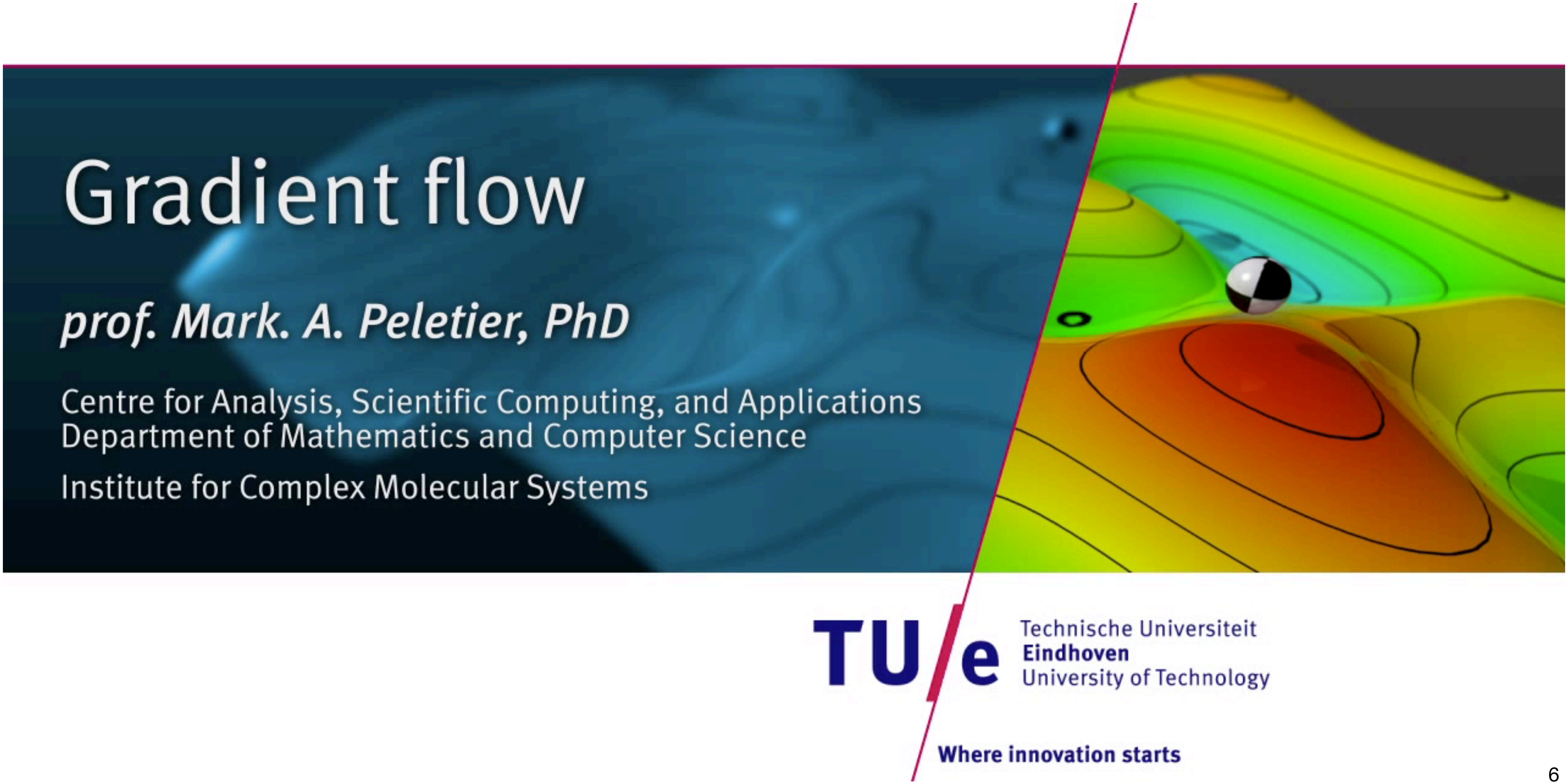
*Optimization*: training neural network with single hidden layer, RBF

# Plan

- Motivation
- Wasserstein gradient flows
- Particle methods (discrete $\leftrightarrow$ continuum)
- Particle method + regularization = blob method for diffusive PDEs
- Numerics

# Gradient flows $\frac{d}{dt}x(t) = -\nabla_d E(x(t))$

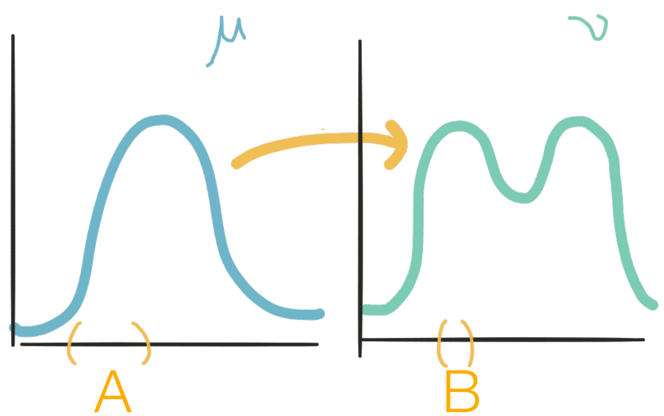
- x(t) evolves in the direction of steepest descent of E, with respect to d
- $x(t+\Delta t) \approx \min_x \frac{1}{2(\Delta t)} d^2(x, x(t)) + E(x(t))$ [De Giorgi '88] [JKO '98]



# Wasserstein metric

- Given $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$, the Wasserstein distance between them is

$$W_2^2(\mu,\nu) = \inf_{\Gamma \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d)} \left\{ \iint_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^2 \, d\Gamma(x,y) : \Gamma(A \times \mathbb{R}^d) = \mu(A), \Gamma(\mathbb{R}^d \times B) = \nu(B) \right\}$$




$\Gamma(A \times B) =$ amount of mass sent from A to B

- W2 **lifts** distance on the underlying space to $\mathscr{P}(\mathbb{R}^d)$: $W_2(\delta_{x_0}, \delta_{y_0}) = |x_0 - y_0|$

- W2 is a **geodesic** metric space

$\mu$ $\nu$

# $W_2$ gradient flows $\partial_t \rho(t) = -\nabla_{W_2} E(\rho(t))$

**Diffusion:** [Jordan, Kinderlehrer, Otto '98,…]

$$\partial_t \rho = \nabla \cdot \left(\rho \nabla \log\left(\rho/\bar{\rho}\right)\right), \quad E(\rho) = \int \rho \log(\bar{\rho}/\rho) = KL(\rho, \bar{\rho})$$

**Degenerate Diffusion:** [Otto '01, Matthes, et al. 2009, Chewi, et. al 2020,…]

$$\partial_t \rho = \nabla \cdot \left(\rho \nabla \left(\rho/\bar{\rho}\right)\right), \quad E(\rho) = \frac{1}{2}\int |\rho - \bar{\rho}|^2/\bar{\rho} = \chi^2(\rho, \bar{\rho}) = \frac{1}{2}\int |\rho|^2/\bar{\rho} + C$$

**Aggregation + Drift:** [McCann '97, Carrillo McCann Villani '05, Carrillo DiFrancesco, Figalli, Laurent, Slepcev '11,…]

$$\partial_t \rho = \nabla \cdot (\rho \nabla (K * \rho)) + \nabla \cdot (\rho \nabla V), \quad E(\rho) = \frac{1}{2}\iint K(x-y) d\rho(x) d\rho(y) + \int V\rho$$

**2-layer neural networks:** [Montanari Mei Nguyen '18, Rotskoff Vanden Eijnden '18, Chizat Bach '18, Sirigano Spiliopoulos '20,…]

$$E(\rho) = \frac{1}{2}\int \left| \int \Phi(x,z) d\rho(x) - f_0(z) \right|^2 d\nu(z)$$

# $W_2$ gradient flows

$\partial_t \rho$

**Choices of K:**

granular media: $K(x)=|x|^3$

swarming: $K(x)=|x|^a/a - |x|^b/b$

chemotaxis: $K(x)=\log(|x|)$

**Choices of Φ:**

$\Phi(x,z) = x_1(\Sigma_i x_i z_i + x_d)_+$

$\Phi(x,z) = \psi(|x-z|)$

**Diffusion:**

$$\partial_t \rho = \nabla \cdot \left(\rho \nabla \log\left(\rho/\bar{\rho}\right)\right), \quad E(\rho) = \int \rho \log(\bar{\rho}/\rho)$$

**Degenerate Diffusion:**

$$\partial_t \rho = \nabla \cdot \left(\rho \nabla \left(\rho/\bar{\rho}\right)\right), \quad E(\rho) = \frac{1}{2}\int |\rho|^2/\bar{\rho}$$

**Aggregation + Drift:**

$$\partial_t \rho = \nabla \cdot (\rho \nabla (K * \rho)) + \nabla \cdot (\rho \nabla V), \quad E(\rho) = \frac{1}{2}\iint K(x-y)d\rho(x)d\rho(y) + \int V\rho$$

**2-layer neural networks:**

$$E(\rho) = \frac{1}{2}\int \left| \int \Phi(x,z)d\rho(x) - f_0(z) \right|^2 d\nu(z)$$

# $W_2$ gradient flows $\partial_t \rho$

**Choices of K:**

granular media: $K(x)=|x|^3$

swarming: $K(x)=|x|^a/a - |x|^b/b$

chemotaxis: $K(x)=\log(|x|)$

**Choices of Φ:**

$$\Phi(x,z) = x_1(\Sigma_i x_i z_i + x_d)_+$$

$$\Phi(x,z) = \psi(|x-z|)$$

**Diffusion:**

$$\partial_t \rho = \nabla \cdot \left(\rho \nabla \log\left(\rho/\bar{\rho}\right)\right), \quad E(\rho) = \int \rho \log(\bar{\rho}/\rho)$$

**Degenerate Diffusion:**

$$\partial_t \rho = \nabla \cdot \left(\rho \nabla \left(\rho/\bar{\rho}\right)\right), \quad E(\rho) = \frac{1}{2}\int |\rho|^2/\bar{\rho}$$

**Aggregation + Drift:**

$$\partial_t \rho = \nabla \cdot (\rho \nabla (K * \rho)) + \nabla \cdot (\rho \nabla V), \quad E(\rho) = \frac{1}{2}\iint K(x-y)d\rho(x)d\rho(y) + \int V\rho$$

**2-layer neural networks:**

$$= \int (\psi * \rho)^2 d\nu$$

$$E(\rho) = \frac{1}{2}\iint \underbrace{\int \Phi(x,z)\Phi(y,z)d\nu(z)}_{K(x,y)} d\rho(x)d\rho(y) - \int \underbrace{\int \Phi(x,z)f_0(z)d\nu(z)}_{V(x)} d\rho(x) + C$$

# Plan

- Motivation
- Wasserstein gradient flows
- Particle methods (discrete $\leftrightarrow$ continuum)
- Particle method + regularization = blob method for diffusive PDEs
- Numerics

# $W_2$ gradient flows

**Diffusion:**

$$\partial_t \rho = \nabla \cdot \left(\rho \nabla \log\left(\rho/\bar{\rho}\right)\right), \quad E(\rho) = \int \rho \log(\bar{\rho}/\rho) = KL(\rho, \bar{\rho})$$

**Degenerate Diffusion:**

$$\partial_t \rho = \nabla \cdot \left(\rho \nabla \left(\rho/\bar{\rho}\right)\right), \quad E(\rho) = \frac{1}{2}\int |\rho|^2/\bar{\rho}$$

**Aggregation + Drift:**

$$\partial_t \rho = \nabla \cdot (\rho \nabla (K * \rho)) + \nabla \cdot (\rho \nabla V), \quad E(\rho) = \frac{1}{2}\iint K(x-y)d\rho(x)d\rho(y) + \int V\rho$$

All $W_2$ gradient flows are solutions of **continuity equations**

$$\partial_t \rho + \nabla \cdot (\rho v[\rho]) = 0, \quad v[\rho] = -\nabla \frac{\partial E}{\partial \rho}$$

# From discrete to continuum

- A key benefit of the **$W_2$ metric** is that it lifts the distance from underlying space to the space of measures.

$$|x - y| = W_2(\delta_x, \delta_y)$$

- A key benefit of **$W_2$ gradient flows** is that they lift the dynamics of systems of ODEs to a PDE. Consider a continuity equation with uniformly Lipschitz continuous **velocity** $v[\rho] : \mathbb{R}^d \to \mathbb{R}^d$:

**PDE** $$\begin{cases} \partial_t \rho + \nabla \cdot (\rho v[\rho]) = 0, \\ \rho(x,0) = \rho_0(x)\,. \end{cases}$$

**ODE** $$\begin{cases} \frac{d}{dt} x_i(t) = v[\rho^N(t)](x_i(t)) \\ x_i(0) = x_{i,0} \end{cases}$$

mean field limit ←

particle method →

$\Updownarrow$ $$\rho^N(t) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i(t)}$$

$$\partial_t \rho^N + \nabla \cdot (\rho^N v[\rho^N]) = 0$$

By Lipschitz continuity, of velocity

$$W_2(\rho^N(t), \rho(t)) \leq e^{\|\nabla v\|_\infty t}\; W_2(\rho_0^N, \rho_0) \xrightarrow{N \to +\infty} 0$$

…what about v not uniformly Lipschitz?

# Wasserstein gradient flows

**Diffusion:**

$$\partial_t \rho = \nabla \cdot \left(\rho \nabla \log\left(\rho/\bar{\rho}\right)\right), \quad E(\rho) = \int \rho \log(\bar{\rho}/\rho) = KL(\rho, \bar{\rho})$$

not Lipschitz

**Degenerate Diffusion:**

$$\partial_t \rho = \nabla \cdot \left(\rho \nabla \left(\rho/\bar{\rho}\right)\right), \quad E(\rho) = \frac{1}{2}\int |\rho|^2/\bar{\rho}$$

not Lipschitz

**Aggregation + Drift:**

$$\partial_t \rho = \nabla \cdot (\rho \nabla (K * \rho)) + \nabla \cdot (\rho \nabla V), \quad E(\rho) = \frac{1}{2}\int (K * \rho)\rho + \int V\rho$$

Lipschitz for $D^2K, D^2V$ bounded

How can we make degenerate diffusion more like aggregation?

Regularize

# Plan

- Motivation
- Wasserstein gradient flows
- Particle methods (discrete ↔ continuum)
- Particle method + regularization = blob method for diffusion
- Numerics

# Blob method for diffusion

**Degenerate Diffusion:**

$$\partial_t \rho = \nabla \cdot \left( \rho \nabla \left( \rho / \bar{\rho} \right) \right), \quad E(\rho) = \int (\psi * \rho)^2 \nu - 2 \int \underbrace{\psi * (f_0 \nu) \rho}_{V}$$

**Approximation of Degenerate Diffusion:**

$$\partial_t \rho = \nabla \cdot \left( \rho \nabla \varphi_\epsilon * \left( \varphi_\epsilon * \rho / \bar{\rho} \right) \right), \quad E_\epsilon(\rho) = \frac{1}{2} \int |\varphi_\epsilon * \rho|^2 / \bar{\rho}$$

**Theorem** (C., Elamvazhuthi, Haberland, Turanova, in preparation): The velocity $v_\epsilon[\rho] = -\nabla \varphi_\epsilon * \left( \varphi_\epsilon * \rho / \bar{\rho} \right)$ is $C_R \epsilon^{-d-2}$ Lipschitz on $\Omega \subseteq B_R(0)$.

Consequently, the particle method is well-posed:

$$\frac{d}{dt} x_i(t) = \sum_{j=1}^{N} f(x_i(t), x_j(t)), \quad f(x_i, x_j) = -\int \nabla \varphi_\epsilon (x_i - x) \varphi_\epsilon (x_j - x) / \bar{\rho}(x) dx$$

**What happens as $N \to +\infty$ *and* $\epsilon \to 0$ ?**

# Convergence of blob method

**Previous work:** $\bar{\rho} = 1$

- [Oelschläger '98]: conv. of particle method to smooth, positive solutions
- [Lions, Mas-Gallic '00]: convergence of bounded entropy solutions as $\epsilon \to 0$ (particles not allowed)
- [Carrillo, C., Patacchini '17]: convergence of bounded entropy solns; allow additional GF terms (aggregation, drift,...), $\partial_t \rho = \Delta \rho^m, m \geq 2$.
- [Javanmard, Mondelli, Montanari '19]: convergence of particle method to smooth, strictly positive solns; allow additional GF terms (2 layer NN)

**Theorem** (C., Elamvazhuthi, Haberland, Turanova, in prep.): Suppose

- $\bar{\rho} = e^{-V}$, for $V : \mathbb{R}^d \to \mathbb{R}$ convex, on a bounded, convex domain $\Omega$.
- $W_2(\rho_0^N, \rho_0) = o(e^{-\frac{1}{\epsilon^{d+2}}})$ for $\rho_0$ with bounded entropy

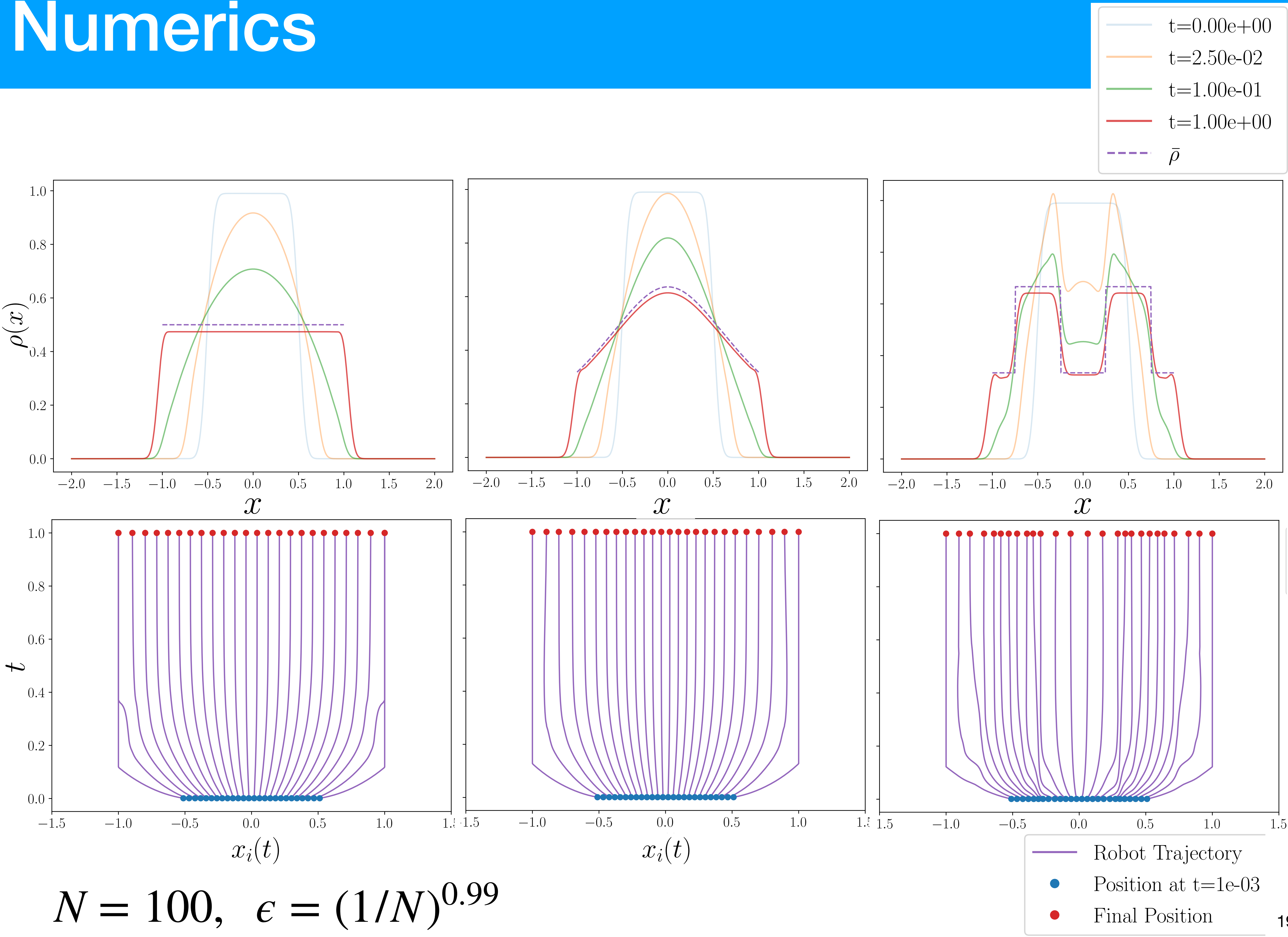
Then $\rho^N(t) \xrightarrow{\epsilon \to 0} \rho(t)$ for all $t \in$

In limiting of 2 layer NN, limiting dynamics are convex GF for $\nu$ log-convex and $f_0 \nu$ concave.

# Plan

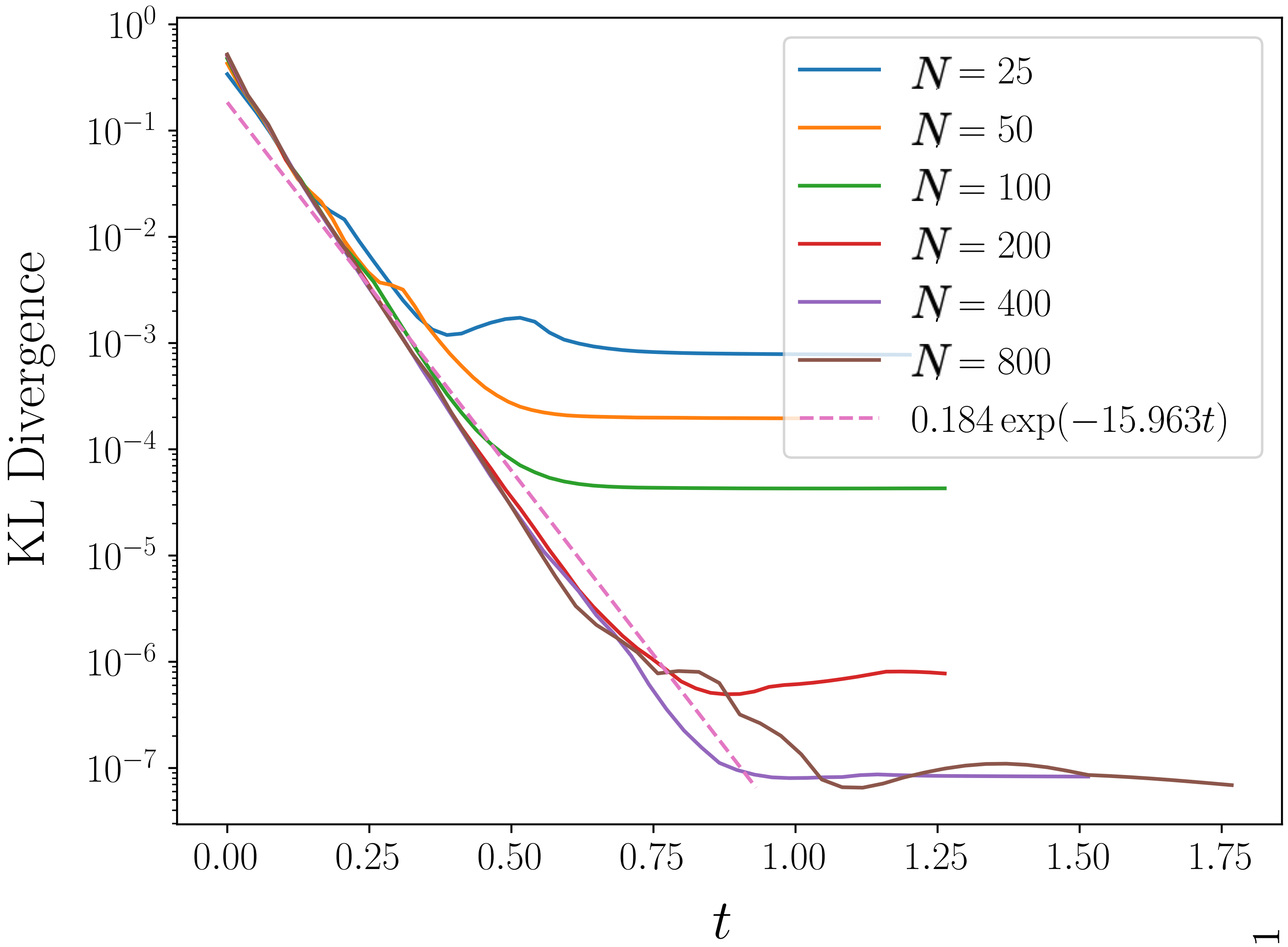
- Motivation:
- Wasserstein gradient flows
- Particle methods (discrete ↔ continuum)
- Particle method + regularization = blob method for diffusive PDEs
- Numerics

# Numerics

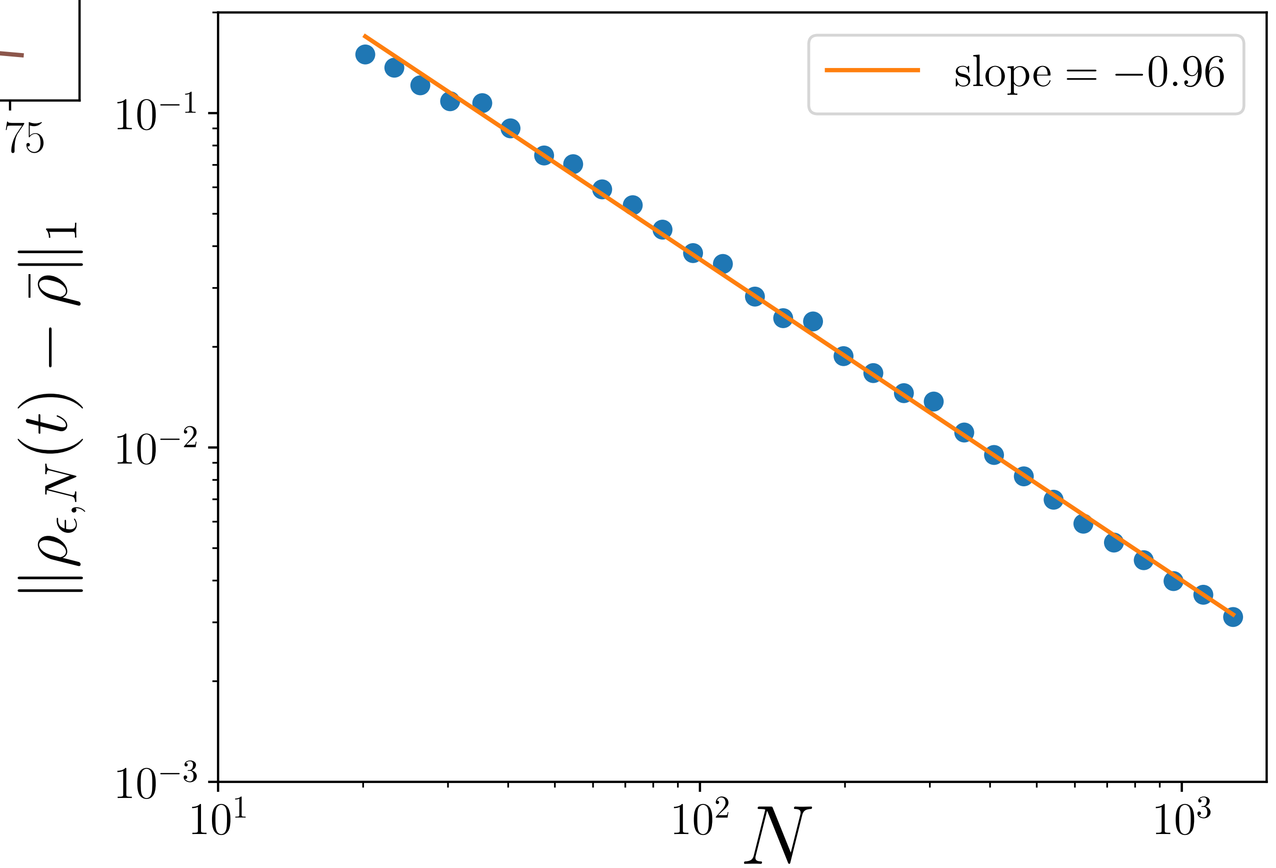



$$N = 100, \quad \epsilon = (1/N)^{0.99}$$

# Numerics




$\bar{\rho}$ log concave



# Numerics

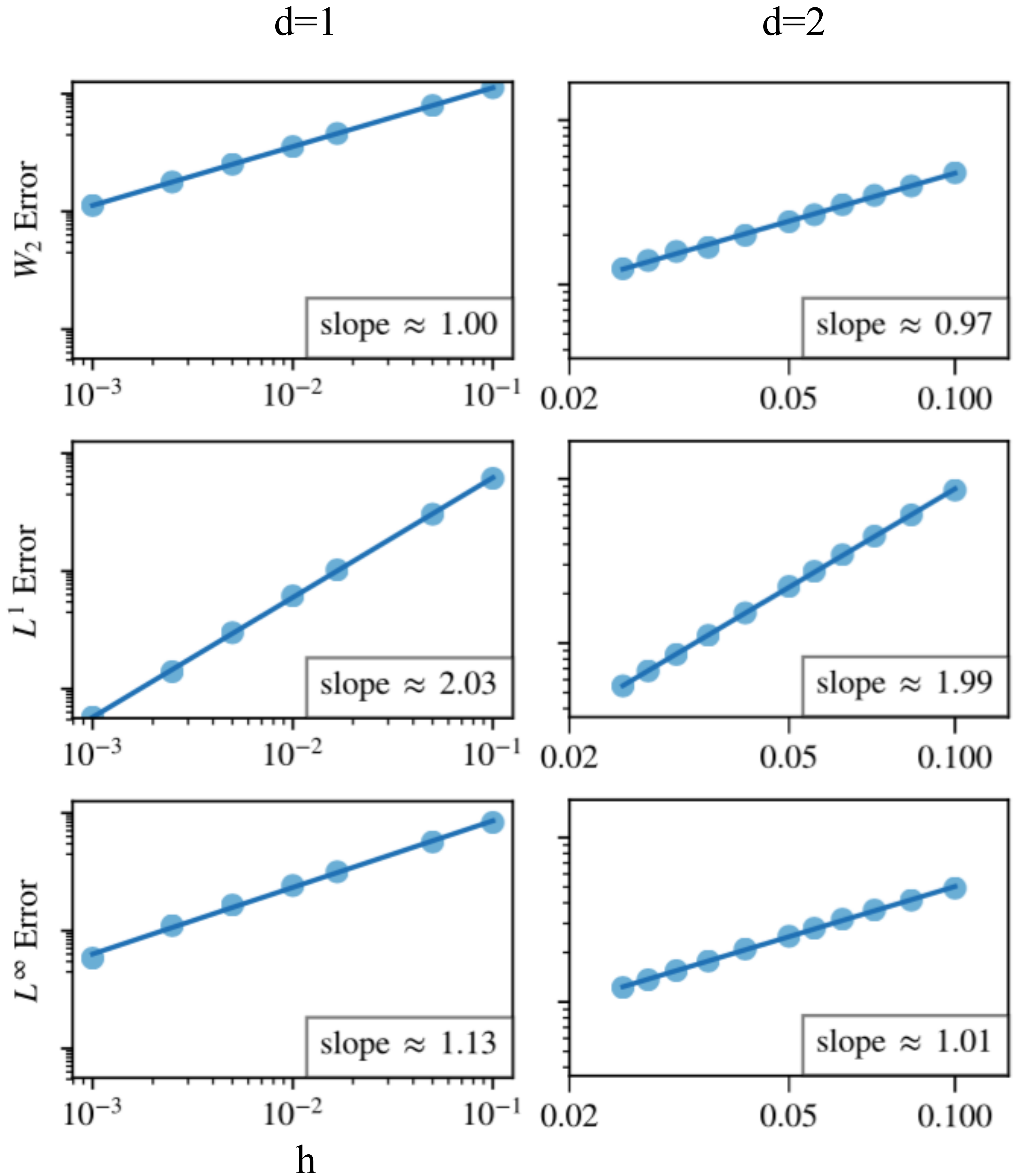
$\bar{\rho} = 1$

Rate of convergence of $\rho_\epsilon(x,t)$ to $\rho(x,t)$, where $\partial_t \rho = \Delta \rho^2$.

$\rho_0^N$ samples $\rho_0$ on a uniform grid

$h = (1/N)^{1/d}$

$\epsilon = h^{.95}$



# Open questions

- general $\bar{\rho}$
- less information on $\bar{\rho}$

$$\frac{d}{dt}x_i(t) = \sum_{j=1}^{N} f(x_i(t), x_j(t)), \quad f(x_i, x_j) = -\int \nabla \varphi_\epsilon(x_i - x)\varphi_\epsilon(x_j - x)/\bar{\rho}(x)dx$$

- Quantitative rate of convergence depending on N and $\epsilon$?
- Can better choice of RBF lead to faster rates of convergence? Help fight against curse of dimensionality? $\mathcal{O}(N^{-m/d})$
- Can random batch method [Jin, Li, Liu '20] lower computational cost from $O(N^2)$ while preserving long-time behavior?

# Thank you!

# Implications

**Sampling:** Spatially discrete, deterministic particle method for sampling according to chi-squared divergence (c.f. [Chewi, et. al. '20])

**PDE:** Provably convergent numerical method for diffusive gradient flows with low regularity (merely bounded entropy)

**Coverage:** *Deterministic* particle method well-suited to robotics

**Optimization:**

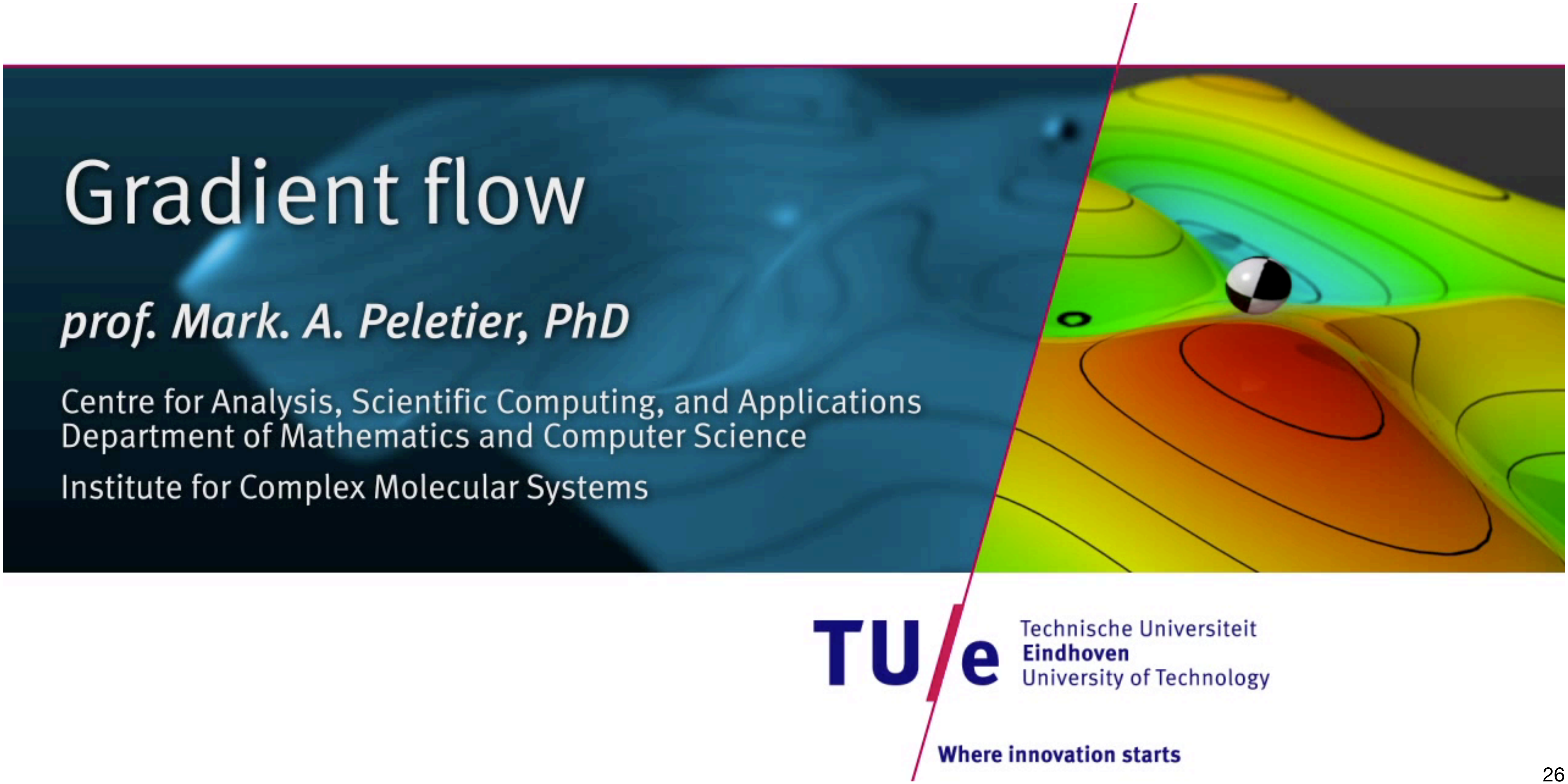
- Particle method equivalent to training dynamics for neural networks with a singular hidden layer, RBF activation.
- Our result identifies limiting dynamics in the over parametrized regime ($N \to +\infty$) as variance of the RBF decreases to zero ($\epsilon \to 0$), $\nu \neq 1$.
- Limiting dynamics are *convex* GF for $\nu$ log-convex and $f_0\nu$ concave.

$$E(\rho) = \int (\psi * \rho)^2 \nu - 2 \int \underbrace{\psi * (f_0 \nu)}_{V} \rho$$

# Backup

# Gradient flows $\frac{d}{dt}x(t) = -\nabla_d E(x(t))$

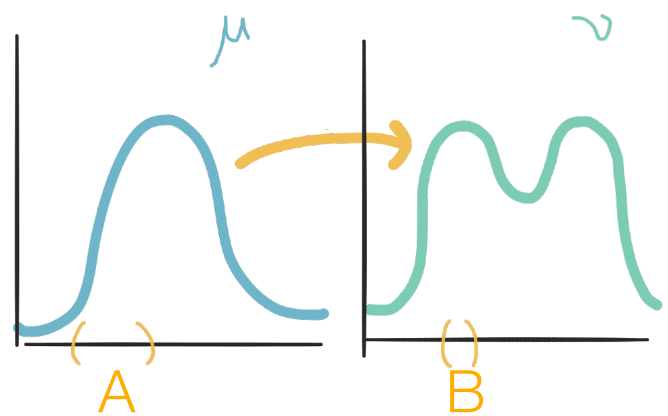
- x(t) evolves in the direction of steepest descent of E, with respect to d
- $x(t+\Delta t) \approx \min_x \frac{1}{2(\Delta t)} d^2(x, x(t)) + E(x(t))$ [De Giorgi '88] [JKO '98]



# Wasserstein metric

- Given $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$, the Wasserstein distance between them is

$$W_2^2(\mu,\nu) = \inf_{\Gamma \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d)} \left\{ \iint_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^2 \, d\Gamma(x,y) : \Gamma(A \times \mathbb{R}^d) = \mu(A), \Gamma(\mathbb{R}^d \times B) = \nu(B) \right\}$$




$\Gamma(A \times B) =$ amount of mass sent from A to B

- W2 **lifts** distance on the underlying space to $\mathscr{P}(\mathbb{R}^d)$: $W_2(\delta_{x_0}, \delta_{y_0}) = |x_0 - y_0|$

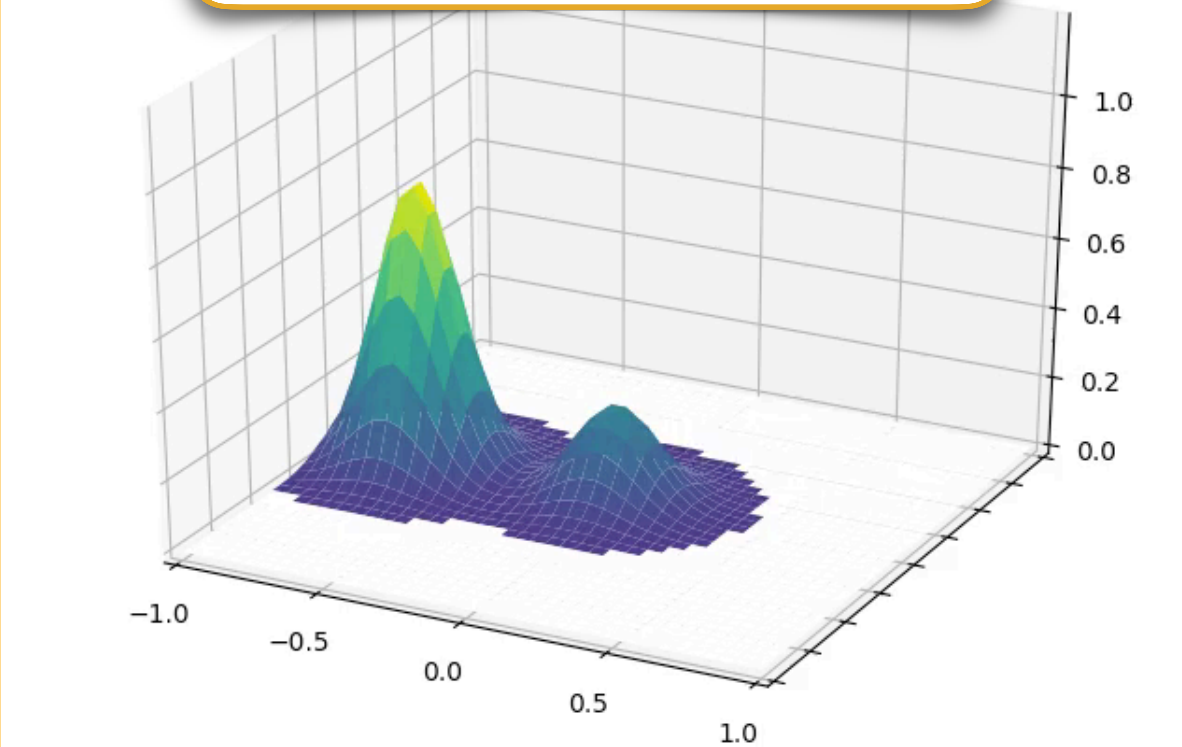
- W2 is a **geodesic** metric space

$\mu$ $\nu$

# Wasserstein metric

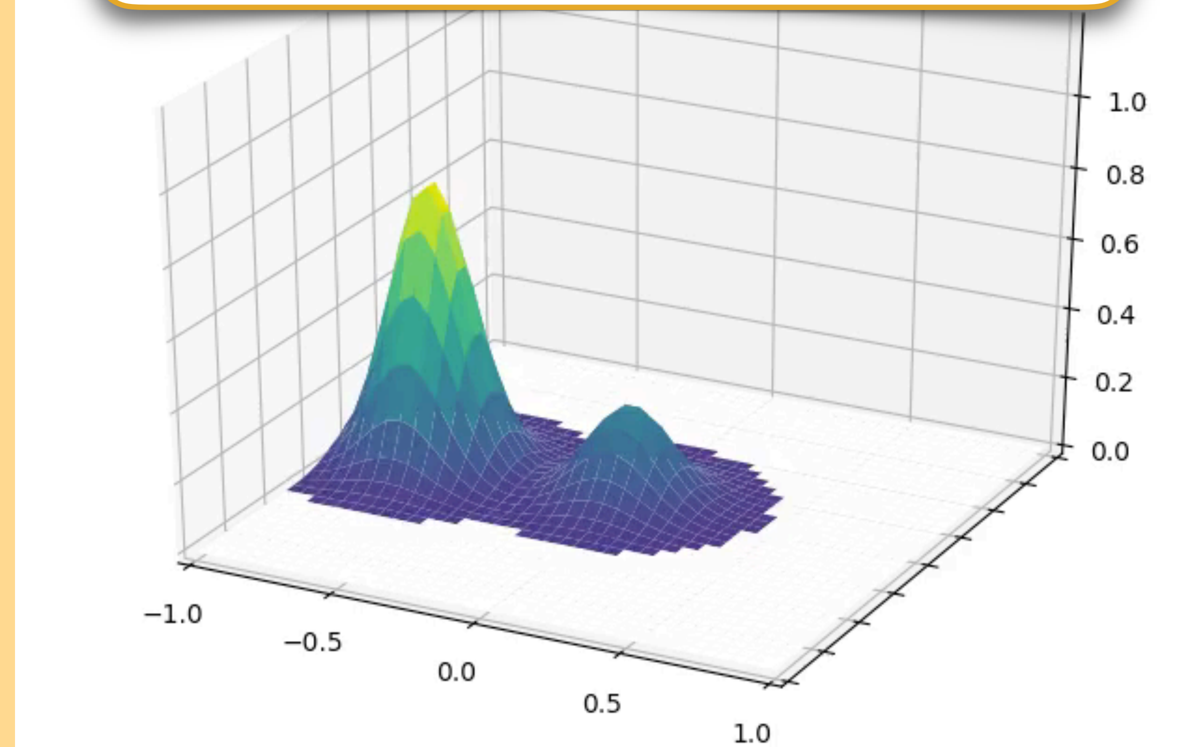
L² geodesic

$$\rho(t) = (1-t)\rho_0 + t\rho_1$$




W₂ geodesic

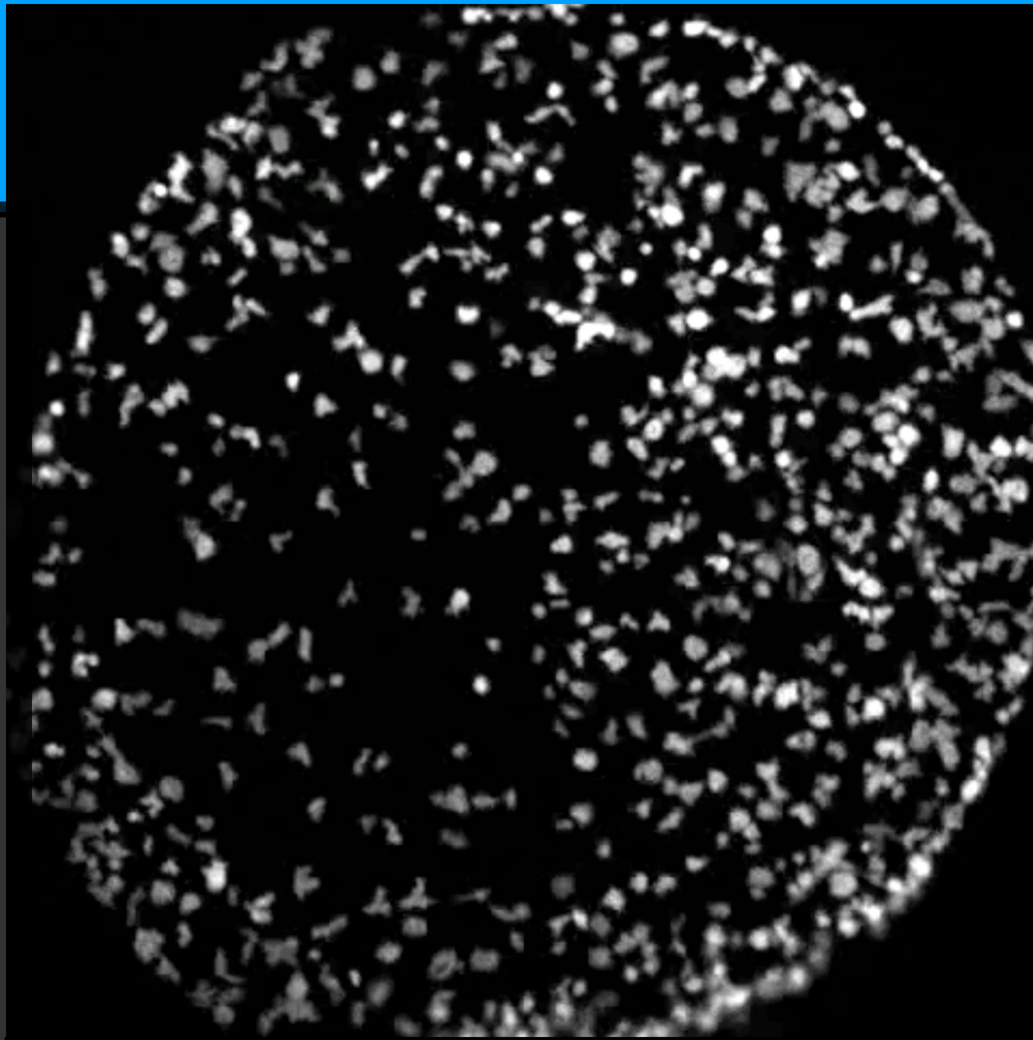
$$\rho(t) = ((1-t)\mathrm{id} + tT_{\rho_0}^{\rho_1})\#\rho_0$$




- $W_2$ is a **geodesic** metric space

$\mu$ $\nu$

# $W_2$ gradient flows



Video Credit: Thomas Gregor,
Laboratory for the Physics of Life, Princeton University

## Neural network with single hidden layer

$$E(\mu) = \frac{1}{2}\int \left| \int \Phi(x,z)d\mu(x) - f_0(z)\right|^2 d\nu(z),$$

$$= \frac{1}{2}\iint \underbrace{\int \Phi(x,z)\Phi(y,z)d\nu(z)}_{K(x,y)} d\mu(y)d\mu(x)$$

$$- \int \underbrace{\int \Phi(x,z)f_0(z)d\nu(z)}_{V(x)} d\mu(x) + C$$

## Swarming, granular media, porous media…

$$E(\mu) = \frac{1}{2}\iint K(x-y)d\mu(x)d\mu(y) + \int V(x)d\mu(x)$$

$$+ \frac{1}{m-1}\int \mu^m(x)dx$$

$$\partial_t \mu = \nabla \cdot ((\nabla K * \mu)\mu) \qquad + \Delta \mu^m$$

## Choices of K:

granular media: $K(x)=|x|^3$

swarming: $K(x)=|x|^a/a - |x|^b/b$

chemotaxis: $K(x)=\log(|x|)$

# $W_2$ gradient flows $\frac{d}{dt}\mu(t) = -\nabla_{W_2} E(\mu(t))$

## Neural network with single hidden layer

$$E(\mu) = \frac{1}{2}\int \left| \int \Phi(x,z)d\mu(x) - f_0(z) \right|^2 d\nu(z),$$

$$= \frac{1}{2}\iint \underbrace{\int \Phi(x,z)\Phi(y,z)d\nu(z)}_{K(x,y)} d\mu(y)d\mu(x)$$

$$- \int \underbrace{\int \Phi(x,z) f_0(z) d\nu(z)}_{V(x)} d\mu(x) + C$$

## Choices of Φ:

$$\Phi(x,z) = x_1(\Sigma_i x_i z_i + x_d)_+$$

$$\Phi(x,z) = \psi(|x-z|)$$

$$\iint K d\mu d\mu = \int (\psi * \mu)^2 d\nu$$

## Swarming, granular media, porous media…

$$E(\mu) = \frac{1}{2}\iint K(x-y)d\mu(x)d\mu(y) + \int V(x)d\mu(x)$$

$$+ \frac{1}{m-1}\int \mu^m(x)dx$$

$$\partial_t \mu = \nabla \cdot ((\nabla K * \mu)\mu) + \nabla \cdot (\nabla V \mu) + \Delta \mu^m$$

## Choices of K:

granular media: $K(x)=|x|^3$

swarming: $K(x)=|x|^a/a - |x|^b/b$

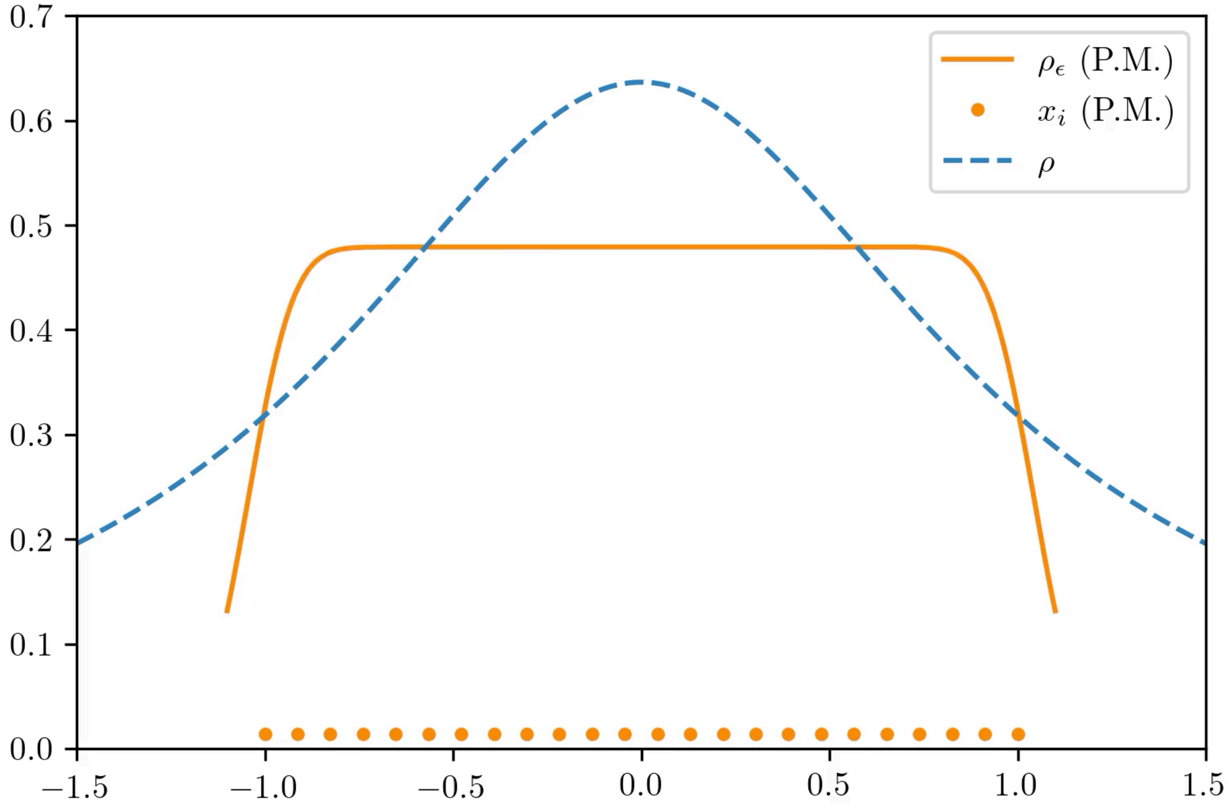
chemotaxis: $K(x)=\log(|x|)$

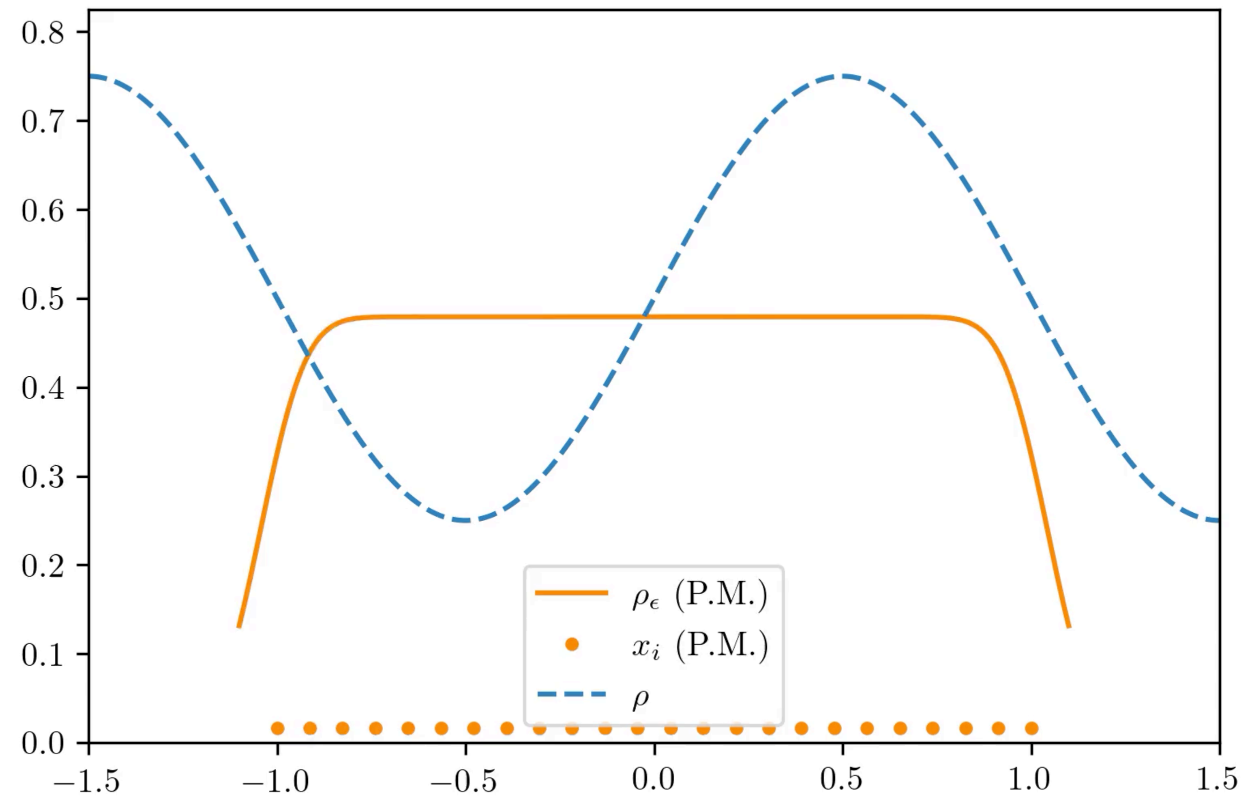
# Gradient flows $\frac{d}{dt}x(t) = -\nabla_d E(x(t))$

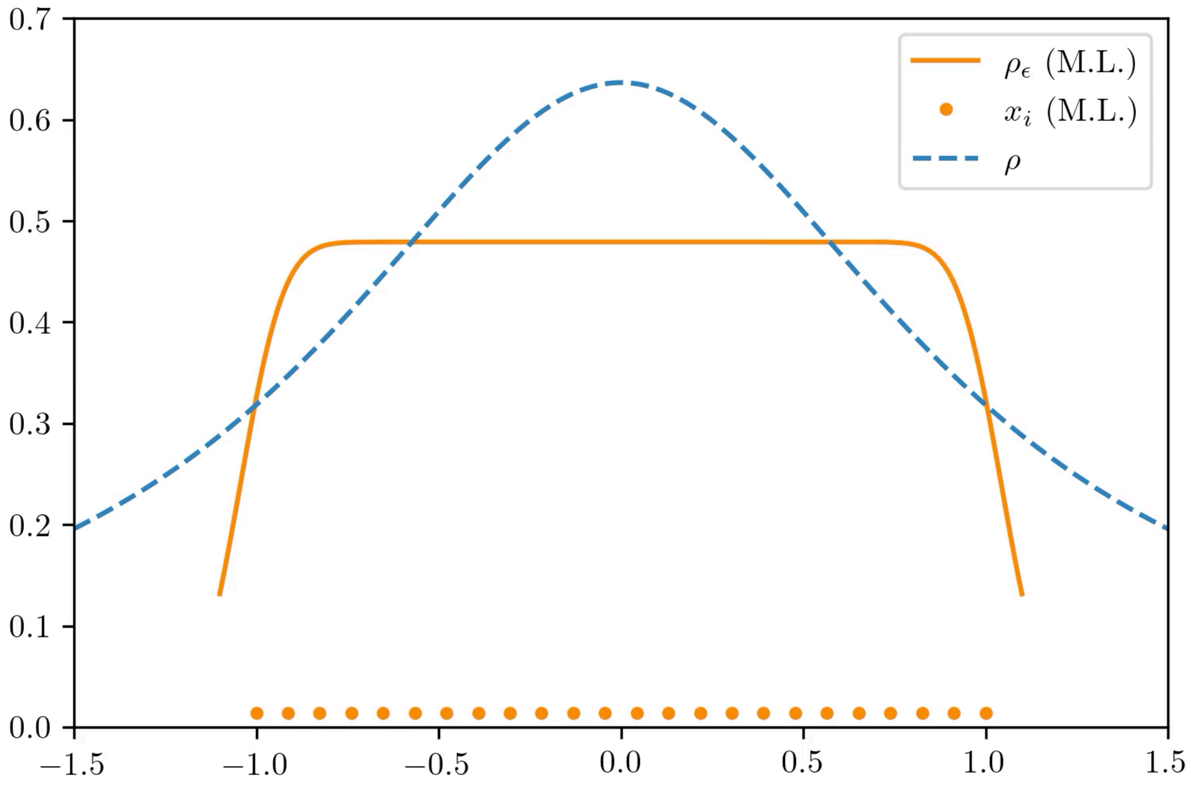
| metric | definition of gradient | formula for gradient |
|---|---|---|
| $\mathbb{R}^d$ | $\langle \nabla E(x), v \rangle = \lim_{h\to 0} \frac{E(x+hv)-E(x)}{h}$ | $\nabla E(x) = \left[\frac{\partial E}{\partial x_i}\right]$ |
| $L^2$ | | |
| $W_2$ | $\langle \nabla_{W_2} E(\mu), -\nabla\cdot(\xi\mu)\rangle_\mu = \lim_{h\to 0} \frac{E((\mathrm{id}+h\xi)\#\mu) - E(\mu)}{h}$ | $\nabla_{W_2} E(\mu) = -\nabla\cdot\left(\mu\nabla\frac{\partial E}{\partial\mu}\right)$ |

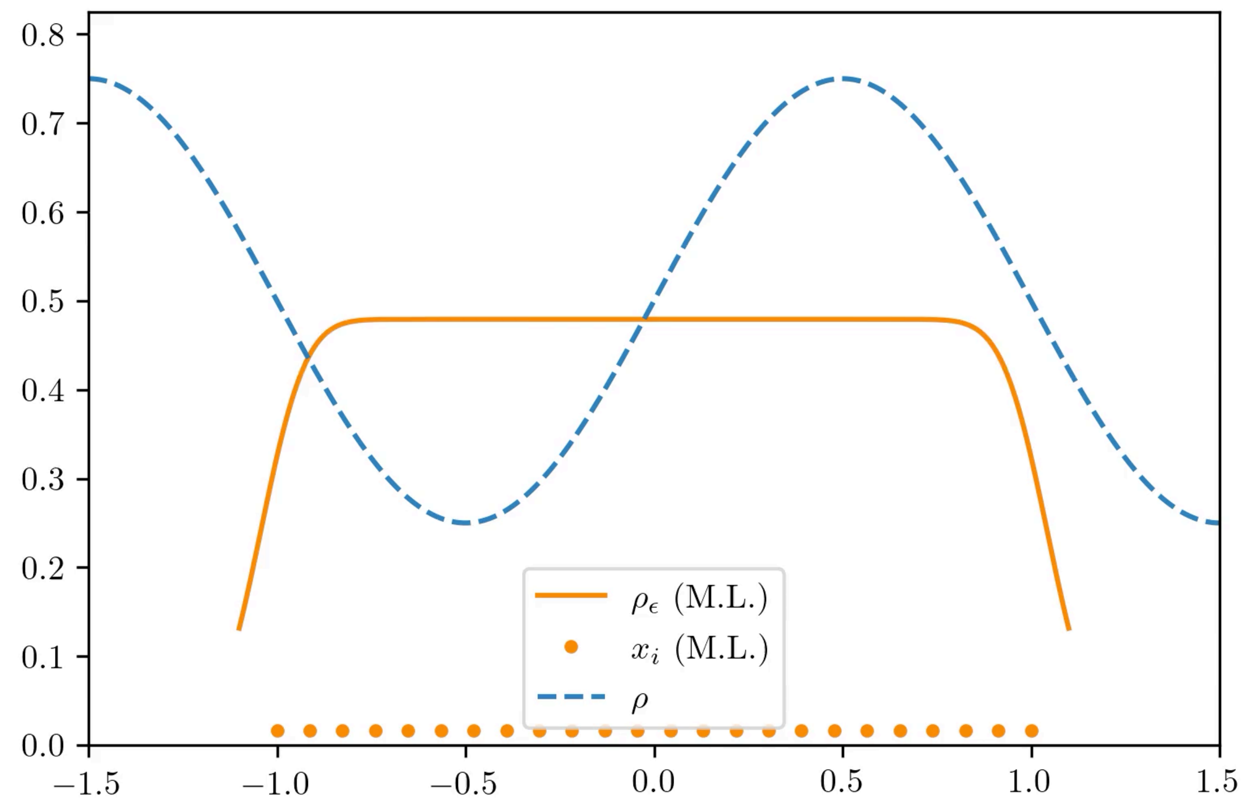
$$\partial_t \mu(t) = -\nabla_{W_2} E(\mu(t)) \iff \partial_t\mu - \nabla\cdot\left(\mu\nabla\frac{\partial E}{\partial\mu}\right) = 0$$

# Application: coverage algorithm



# Outline

- W2 lifts discrete to continuum
- W2 GFs also provide novel tools from passing between discrete and continuum
- Diffusive robot coverage algorithms / Sampling / Training dynamics for neural networks with a single hidden layer
    - These are W2 GFs
    - Particle method well-posed for epsilon >0, converges as N\to +\infty
    - Gamma convergence as epsilon to 0
    - Convergence of particle method as epsilon to 0 and N \to +\infty
    - Emergence of convexity in the limit
- Open problems: right now, N has to grow much faster than epsilon; nothing quantitative on rate of convergence to diffusive equation