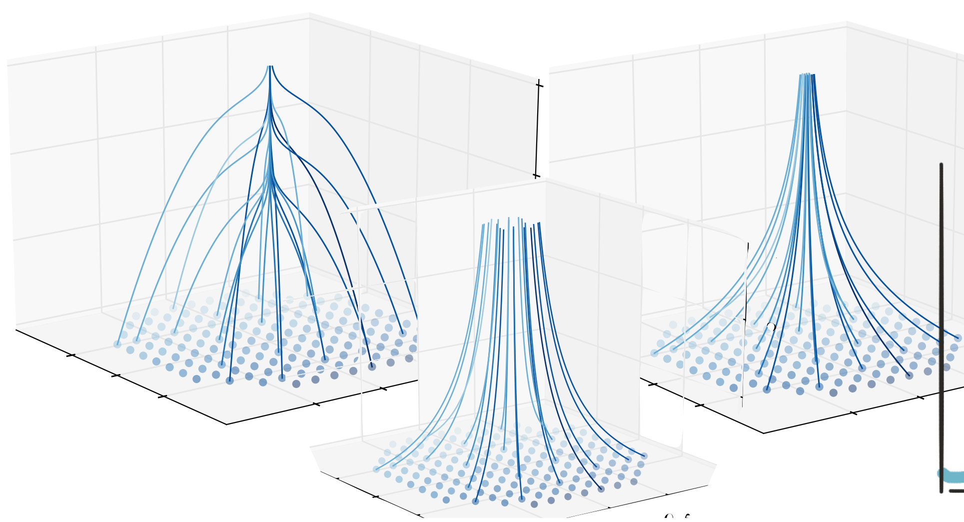
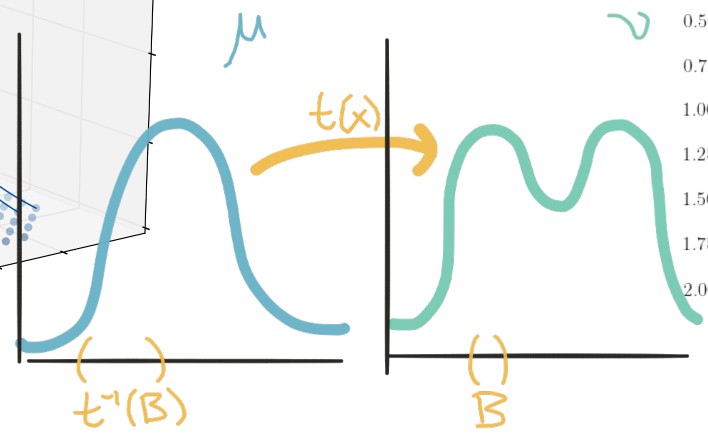
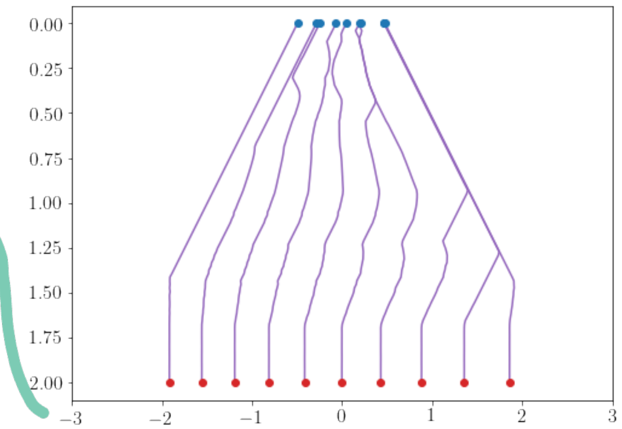






Gradient Flows in the Wasserstein Metric:

From Discrete to Continuum via Regularization

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)

Stanford Operations Research Seminar, November 4th, 2020

Hellman Fellows Fund

Plan

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

PDEs and sampling/coverage algs

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 $V : \mathbb{R}^d \to \mathbb{R}$ 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 2008,…], $KL(\mu, \nu) = \int \mu \log(\mu/\nu)$

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. 200

Particle method: ?

Motivation for deg. diff:

Sampling: SVGD, chi-sq.

PDE: porous media, chemotaxis, …

Coverage: **deterministic** particle method

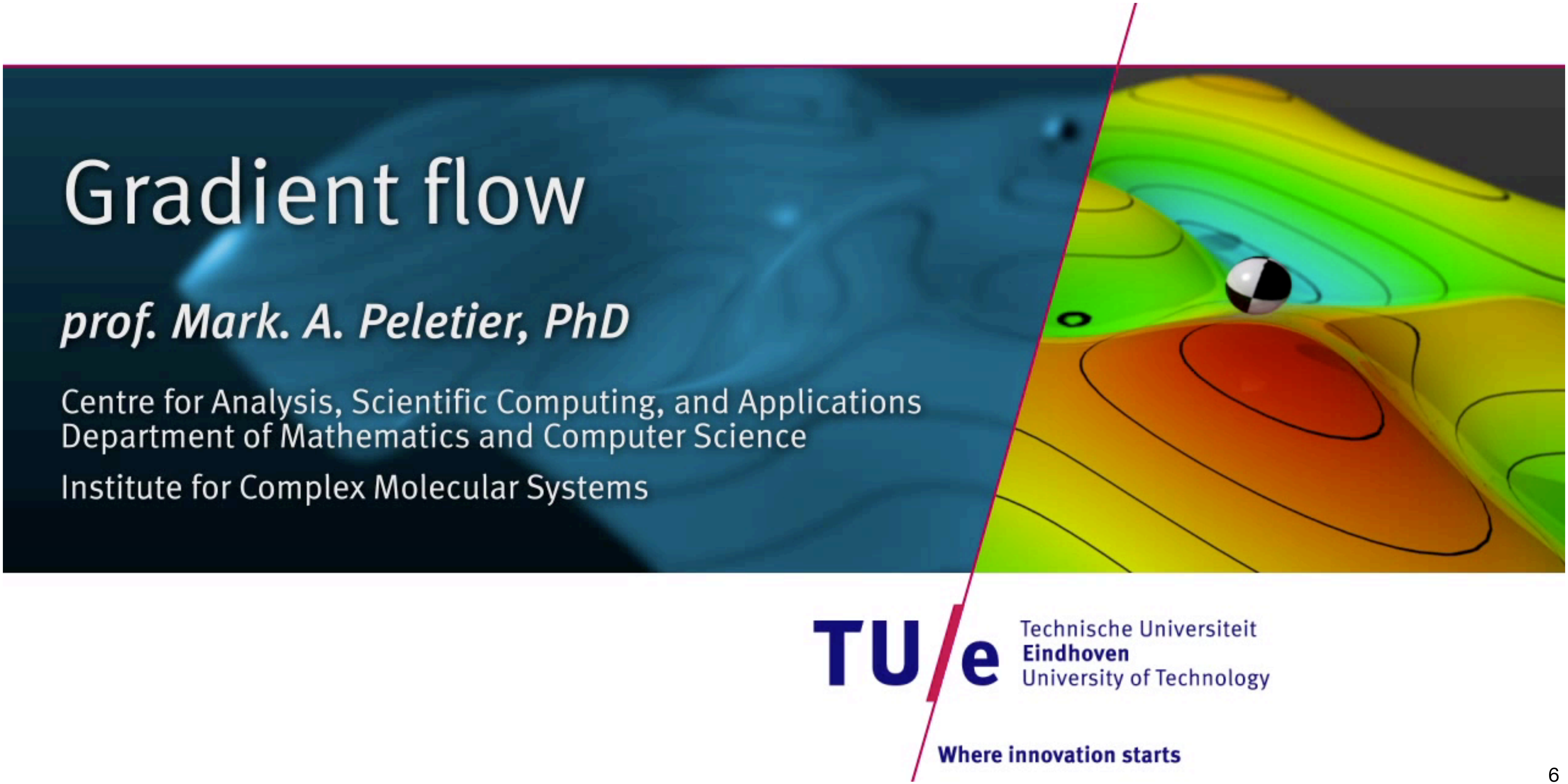
Optimization: training neural network with single hidden layer, RBF

Plan

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

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

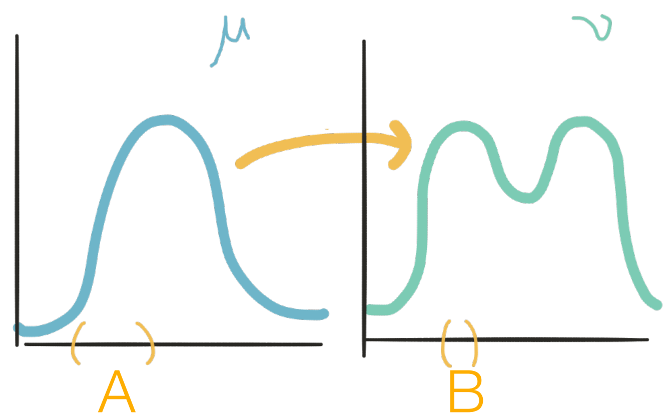
- x(t) evolves in the direction of steepest descent of E, with respect to d
- $x(t+\Delta t) \approx \min_x \frac{1}{2(\Delta t)} d^2(x, x(t)) + E(x(t))$ [De Giorgi '88] [JKO '98]



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 $\mathscr{P}(\mathbb{R}^d)$: $W_2(\delta_{x_0}, \delta_{y_0}) = |x_0 - y_0|$

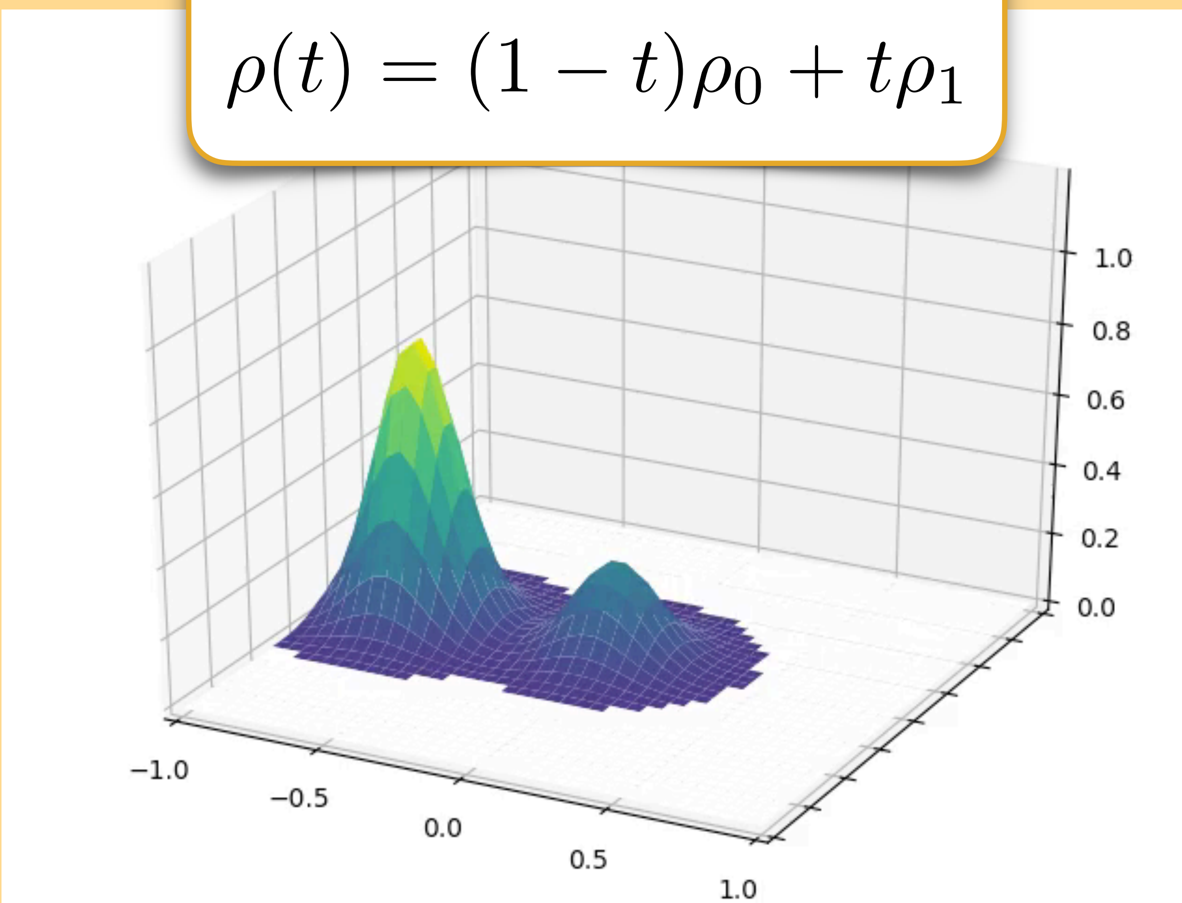
- W2 is a **geodesic** metric space

μ ν

Wasserstein metric

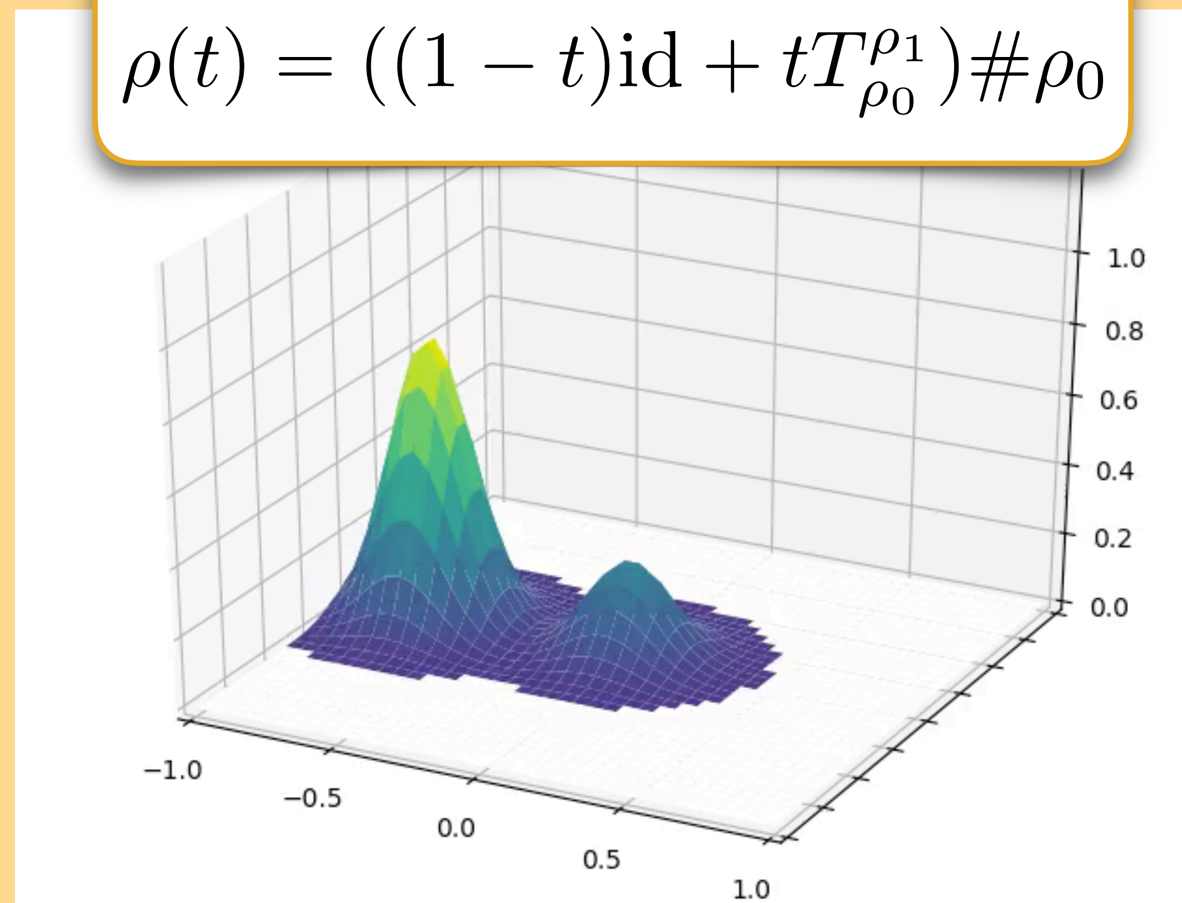
L^2 geodesic

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



W_2 geodesic

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



- W_2 is a **geodesic** metric space

μ ν

Wasserstein gradient flo

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:

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

Degenerate Diffusion:

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

Aggregation + Drift:

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

Training dynamics of 2-layer neural networks: [MMN '18] [RVE '18] [CB '18]…

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

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

Plan

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

Wasserstein gradient flows

Diffusion:

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

Degenerate Diffusion:

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

Aggregation + Drift:

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

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

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

Particle methods

Consider a continuity equation with a uniformly Lipschitz continuous **velocity** $v[\rho] : \mathbb{R}^d \to \mathbb{R}^d$

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

1. Approximate initial data: $\rho_0^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}$
2. Evolve the locations:

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

$$\frac{d}{dt} x_i(t) = v[\rho^N(t)](x_i(t)) \iff \partial_t \rho^N + \nabla \cdot (\rho^N v[\rho^N]) = 0$$

particle method *lifts* solutions of ODEs into PDE framework

3. Since $v[\rho]$ unif Lipschitz,

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

W_2 GF perspective gives tools for proving $v[\rho]$ unif Lipschitz

Benefits of particle methods: deterministic, positivity preserving, adaptive, energy decreasing,… but what about v <u>not unif Lipschitz?</u>

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:

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

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 K, V smooth

How can we use a particle method for aggregation equations to get a particle method for degenerate diffusion?

Regularize

Plan

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

Blob method for diffusion

Degenerate Diffusion:

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

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

Approximation of Degenerate Diffusion:

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

This particle method is precisely the dynamics of training a neural network with a single hidden layer, with RBF activation function.

Theorem (C., Elamvazhuthi, Haberland, Turanova, i
The velocity $v_\epsilon[\rho] = -\nabla \varphi_\epsilon * \left(\varphi_\epsilon * \rho/\bar{\rho}\right)$ is $C_R \epsilon^{-}$
satisfying $\mathrm{supp}\ \rho \subseteq B_R(0)$.

Consequently, the particle method is well-posed:

$$\frac{d}{dt} x_i(t) = -\nabla \varphi_\epsilon * \left(\varphi_\epsilon * \rho^N(t)/\bar{\rho}\right) = -\nabla \varphi_\epsilon * \left(\frac{1}{N}\sum_{i=1}^{N} \varphi_\epsilon(x_i(t) - x_j(t))/\bar{\rho}(x_i(t))\right)$$

and, for fixed $\epsilon > 0$, as $N \to +\infty$, this converges to the GF of E_ϵ.

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 2000]: convergence of bounded entropy solutions as $\epsilon \to 0$ (particles not allowed) $\int \rho(t) \log \rho(t) < +\infty$
- [Carrillo, C., Patacchini 2017]: convergence of bounded entropy solns; allow additional GF terms (aggregation, drift,…), $\partial_t \rho = \Delta \rho^m, m \geq 1$.
- [Javanmard, Mondelli, Montanari 2019]: convergence of particle method to smooth, strictly positive solns; allow additional GF terms (2 layer NN)

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

- $\bar{\rho} \in C^2(\mathbb{R}^d), \bar{\rho} > 0$
- $W_2(\rho_0^N, \rho_0) = o(e^{-\frac{1}{\epsilon^{d+2}}})$ for ρ_0 with bounded entropy and cpt support

Then $\rho^N(t) \xrightarrow{N \to +\infty} \rho(t)$ for all $t \in [0,T]$.

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 ν log-convex and $f_0\nu$ concave.

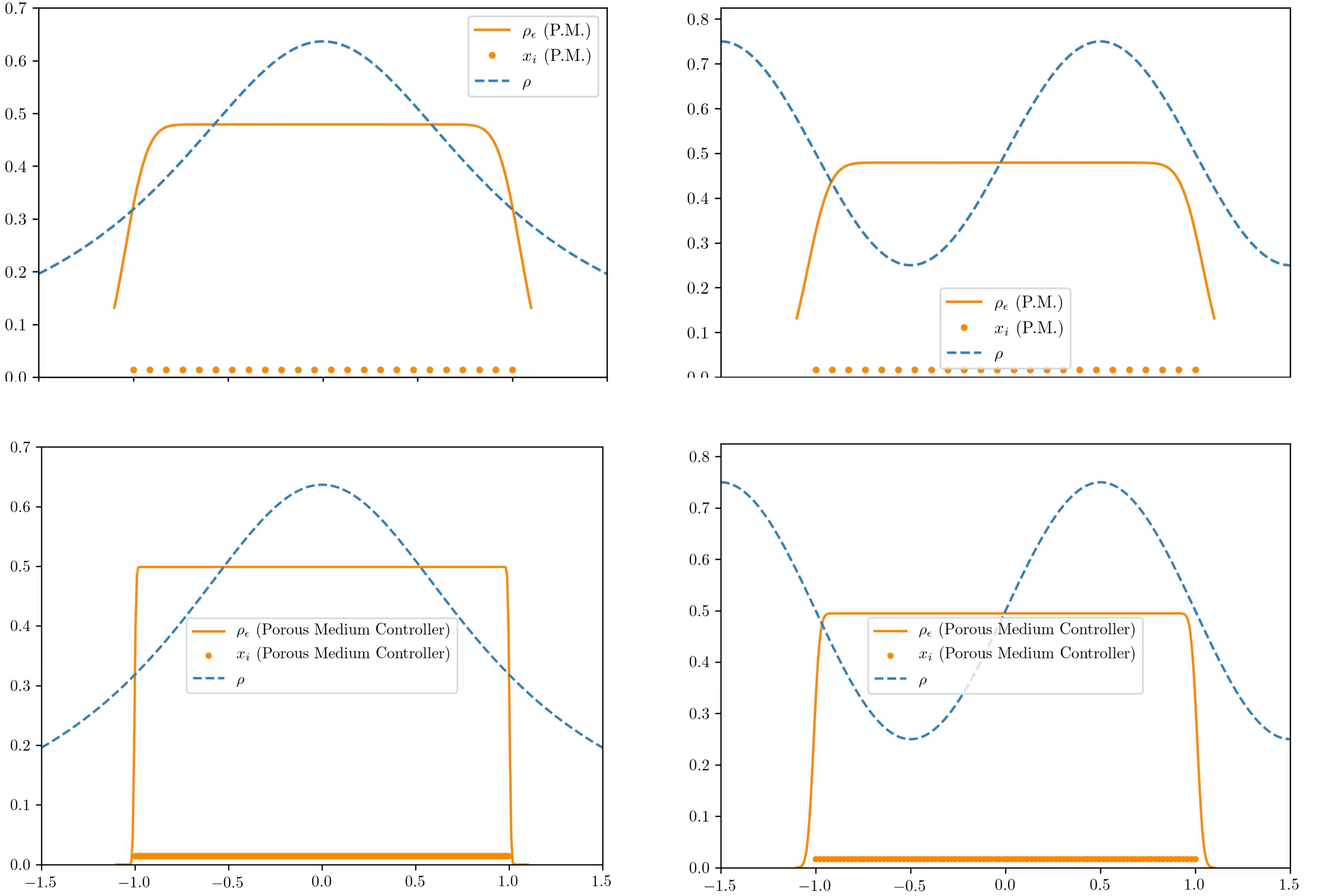
$$E(\rho) = \int (\psi * \rho)^2 \nu - 2 \int \underbrace{\psi * (f_0 \nu)}_{V} \rho$$

Plan

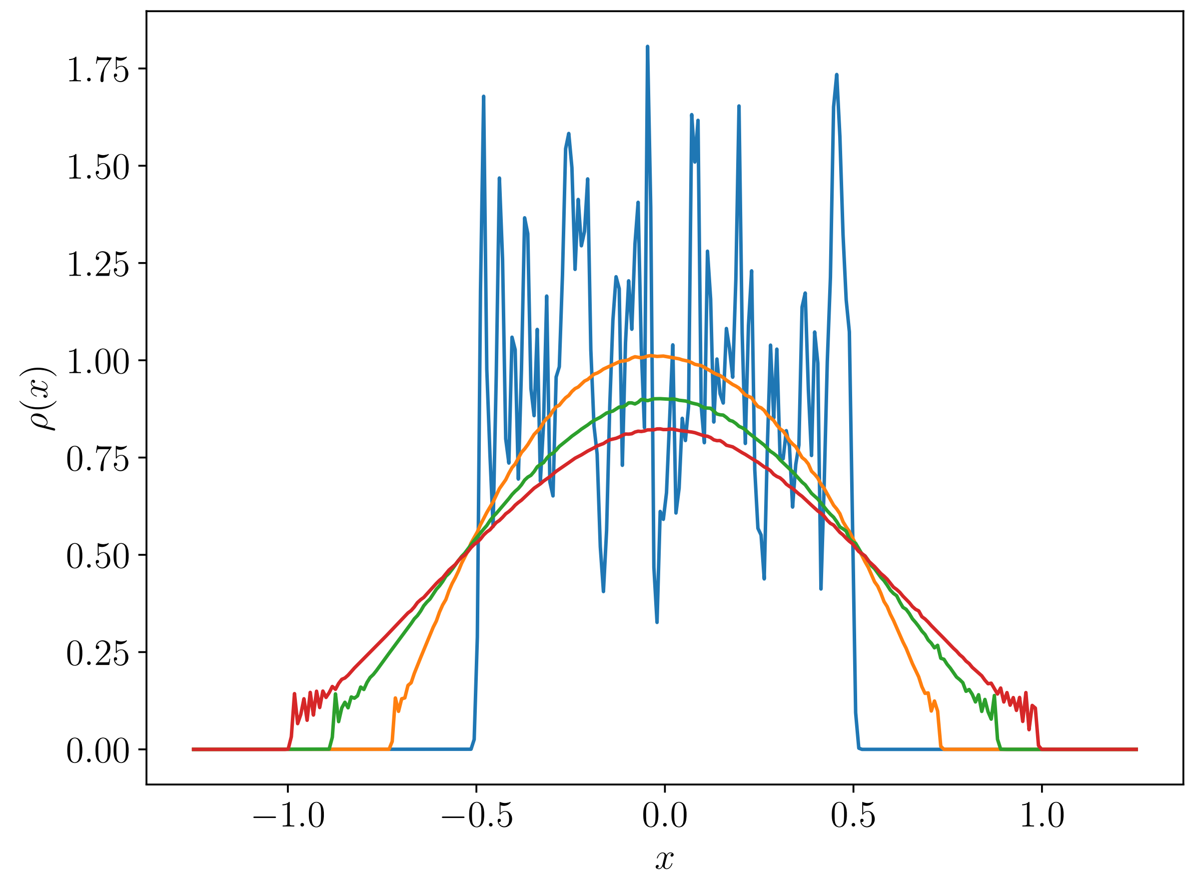
- Motivation:
  - Diffusive PDEs and sampling/coverage algorithms
  - Training dynamics for neural networks with a single hidden layer
- Wasserstein gradient flows
- Particle methods (discrete $\leftrightarrow$ continuum)
- Particle method + regularization = blob method for diffusive PDEs
- Numerics

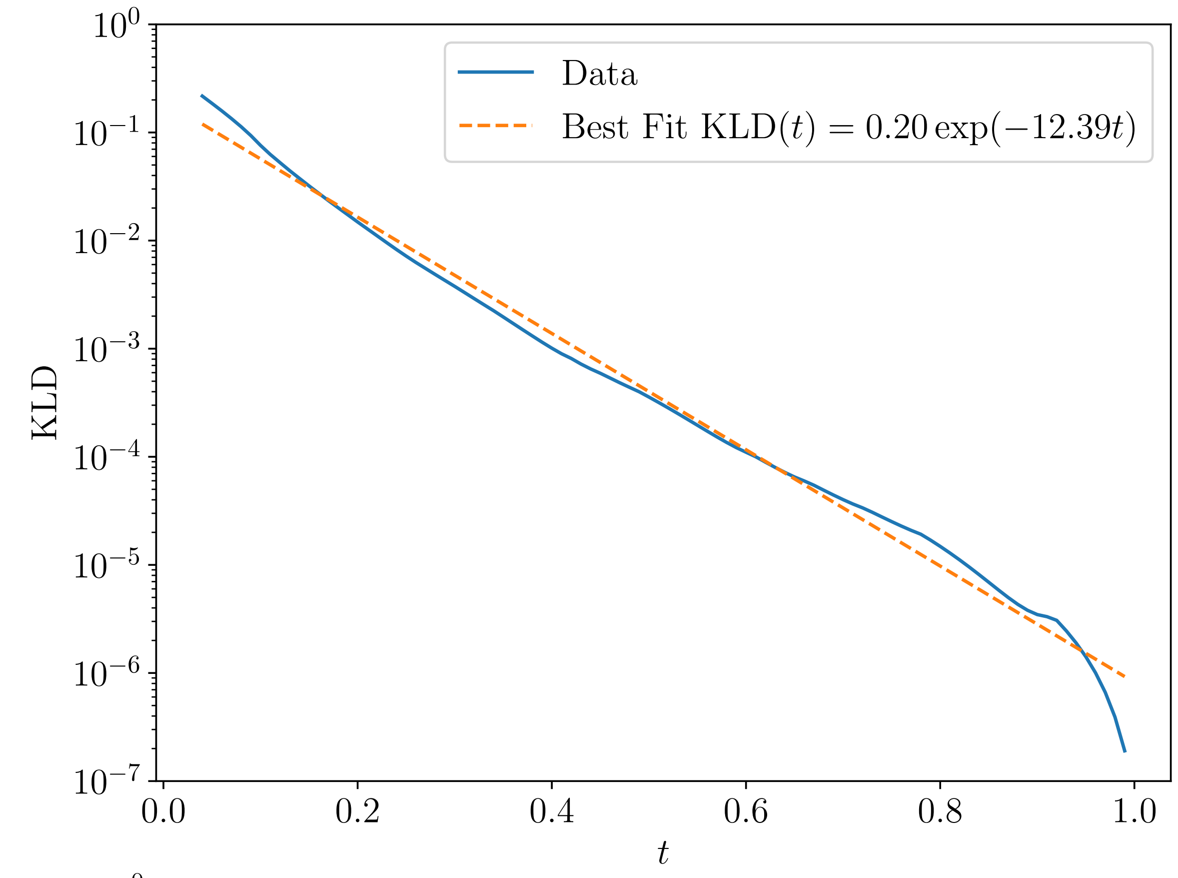
Numerics

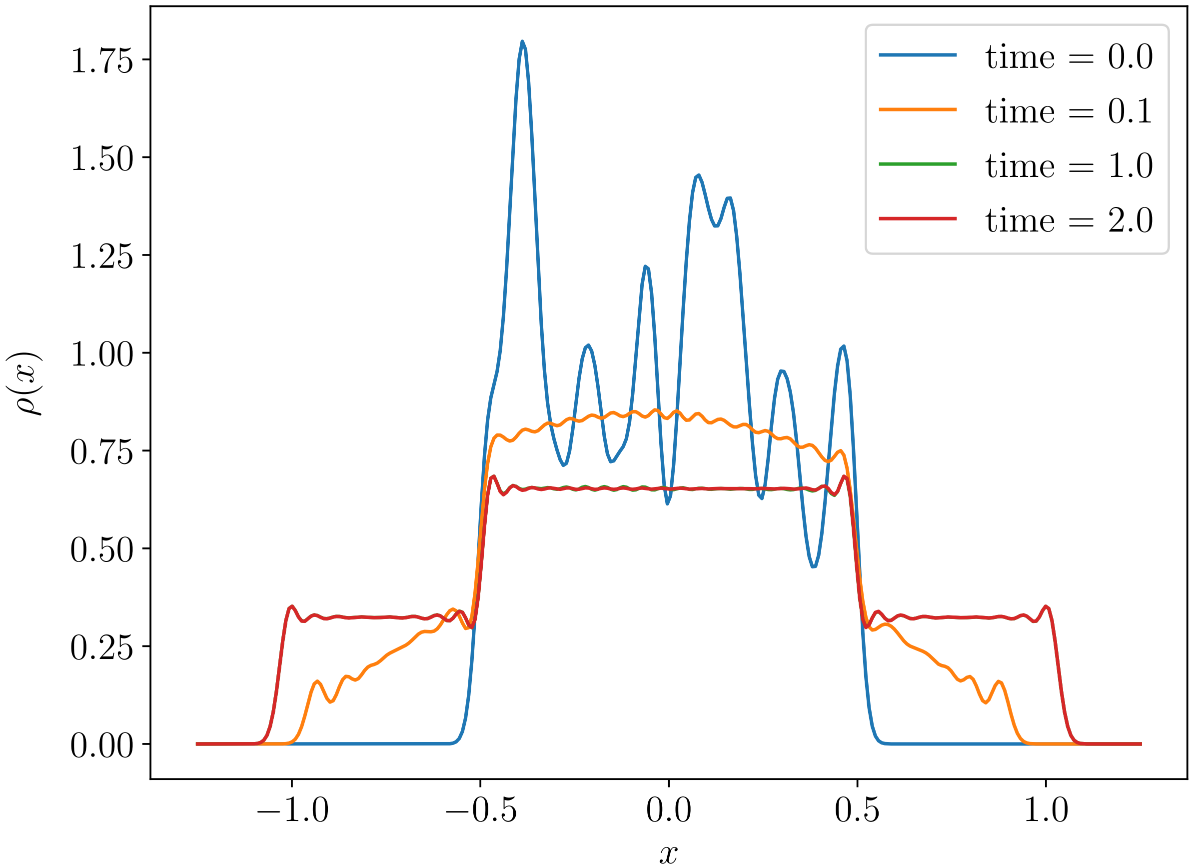
$$\rho_\epsilon(x,t) = \frac{1}{N}\sum_{i=1}^{N} \varphi_\epsilon(x - x_i(t)) = \varphi_\epsilon * \rho^N(t)$$

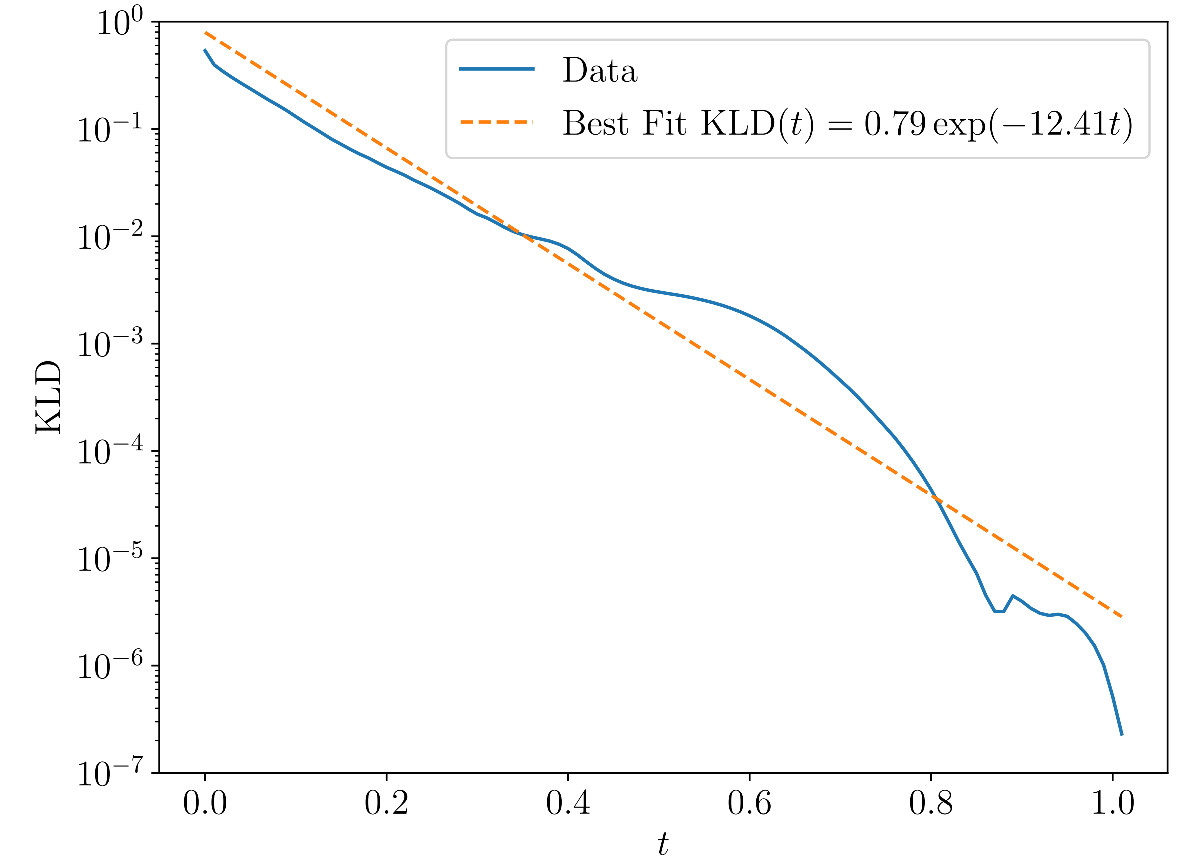


Numerical results: sampling









Numerics

$$\rho_\epsilon(x,t) = \frac{1}{N}\sum_{i=1}^{N} \varphi_\epsilon(x - x_i(t)) = \varphi_\epsilon * \rho^N(t)$$

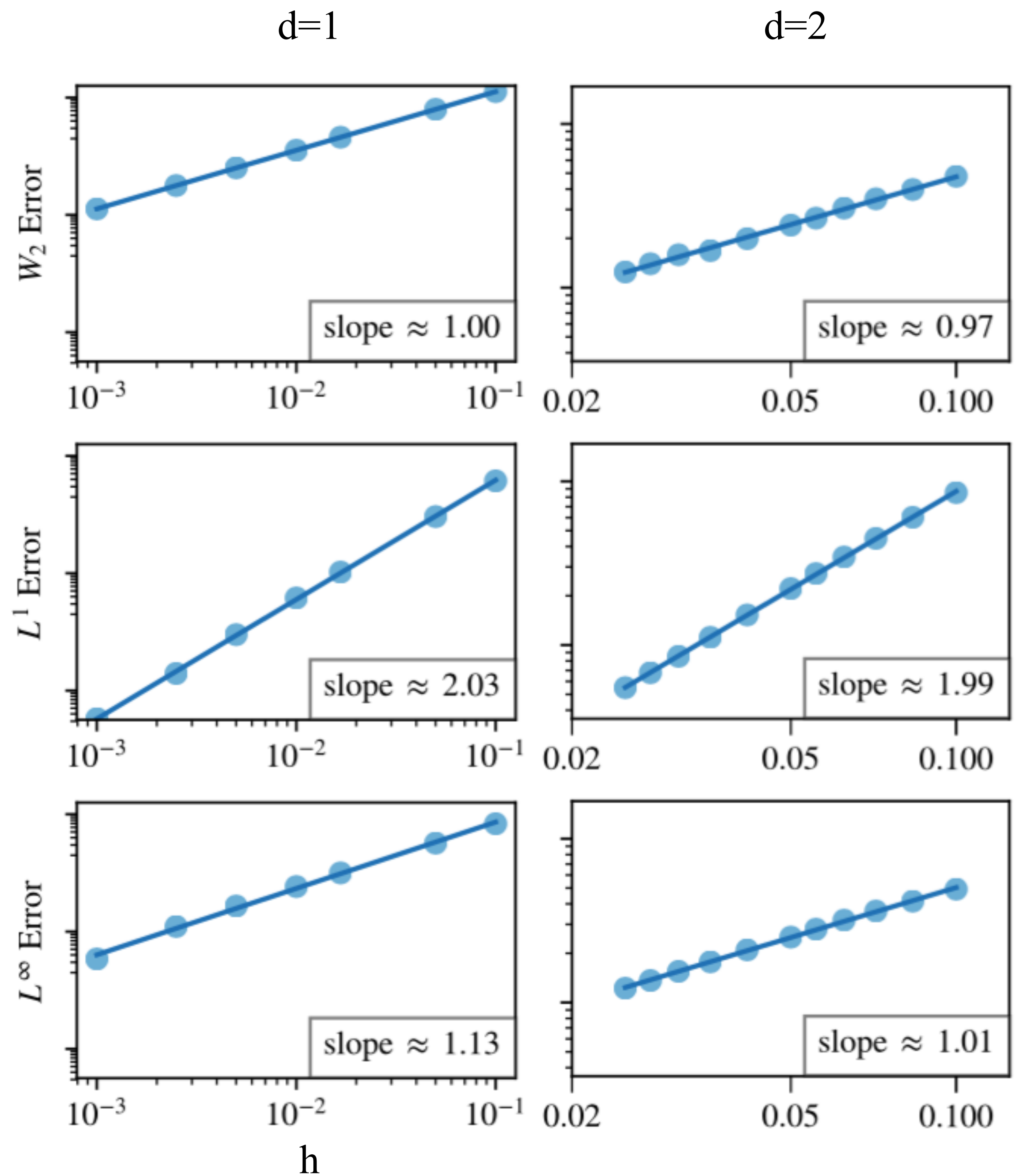
$\bar{\rho} = 1$

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

ρ_0^N samples ρ_0 on a uniform grid

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

$\epsilon = h^{.95}$



Open questions

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

Thank you!