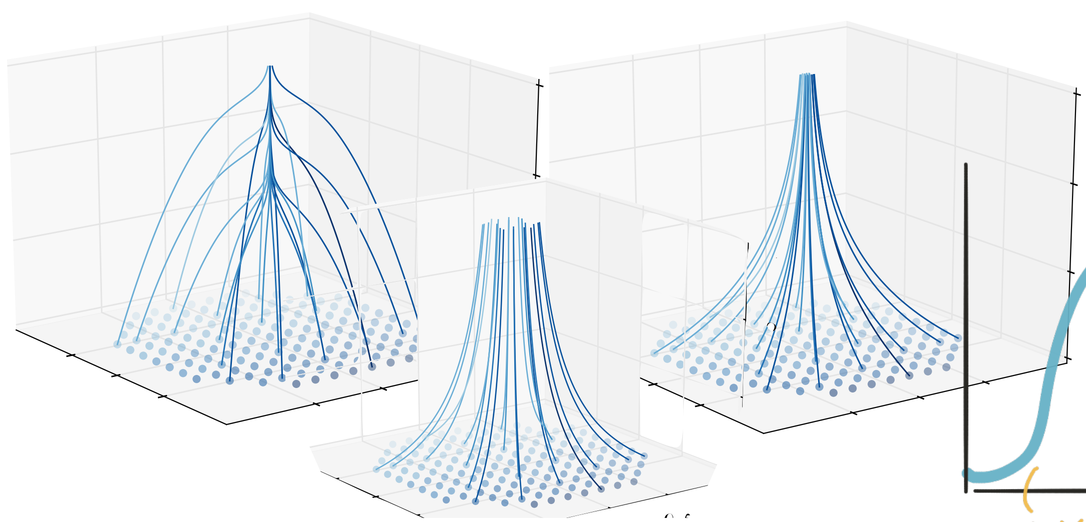
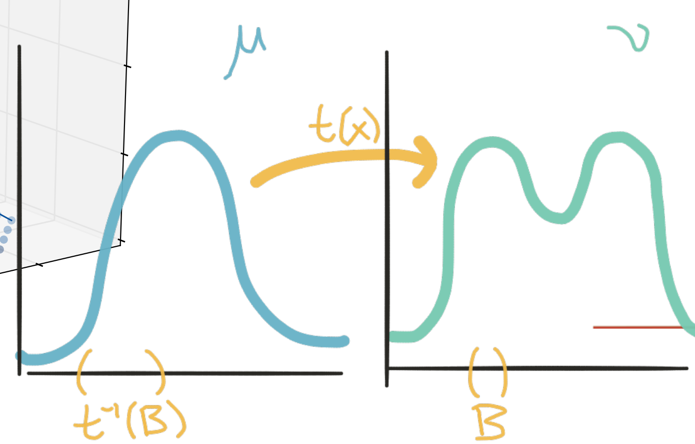
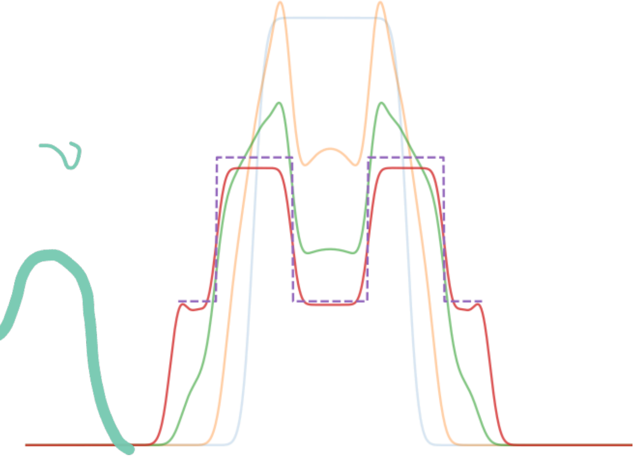
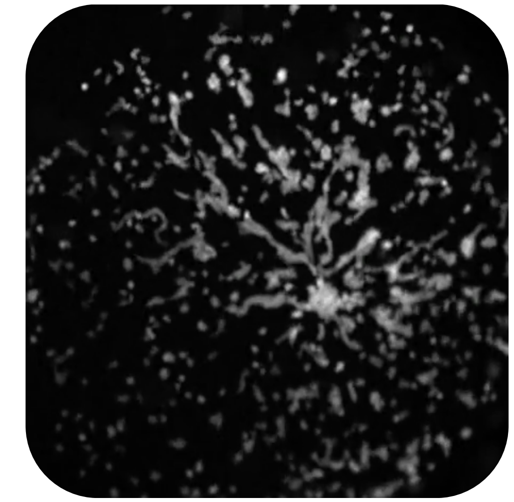



A blob method for degenerate diffusion and applications to sampling and two layer neural networks.

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)

Simons Center for Theory of Computing
“Sampling Algorithms and Geometries on Probability Distributions”
September 7th, 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 $V : \mathbb{R}^d \to \mathbb{R}$ λ-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, 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

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

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:

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

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: [MMN '18] [RVE '18] [CB '18]

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

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

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

Choices of Φ:

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

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

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:

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

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 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:

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

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

3. Since $v[\rho]$ unif Lipschitz, $W_2(\rho^N(t), \rho(t)) \leq e^{\|\nabla v\|_\infty t} \, W_2(\rho_0^N, \rho_0) \xrightarrow{N \to +\infty} 0$

…what about v <u>not unif Lipschitz?</u>

Wasserstein gradient flows

Diffusion:

not Lipschitz

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

Degenerate Diffusion:

not Lipschitz

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

Aggregation + Drift:

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

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 $\leftrightarrow$ continuum)
- Particle method + regularization = blob method for diffusion
- Numerics

Blob method for diffusion

Degenerate Diffusion:

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

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:

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

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)
- [Carrillo, C., Patacchini 2017]: convergence of bounded entropy solns; allow additional GF terms (aggregation, drift,...), $\partial_t \rho = \Delta \rho^m, m \geq 2$.
- [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

- $\bar{\rho} = e^{-V}$, for $V : \mathbb{R}^d \to \mathbb{R}$ convex, on a bounded, convex domain Ω.
- $W_2(\rho_0^N, \rho_0) = o(e^{-\frac{1}{\epsilon^{d+2}}})$ for ρ_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 ν log-convex and $f_0 \nu$ concave.

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:

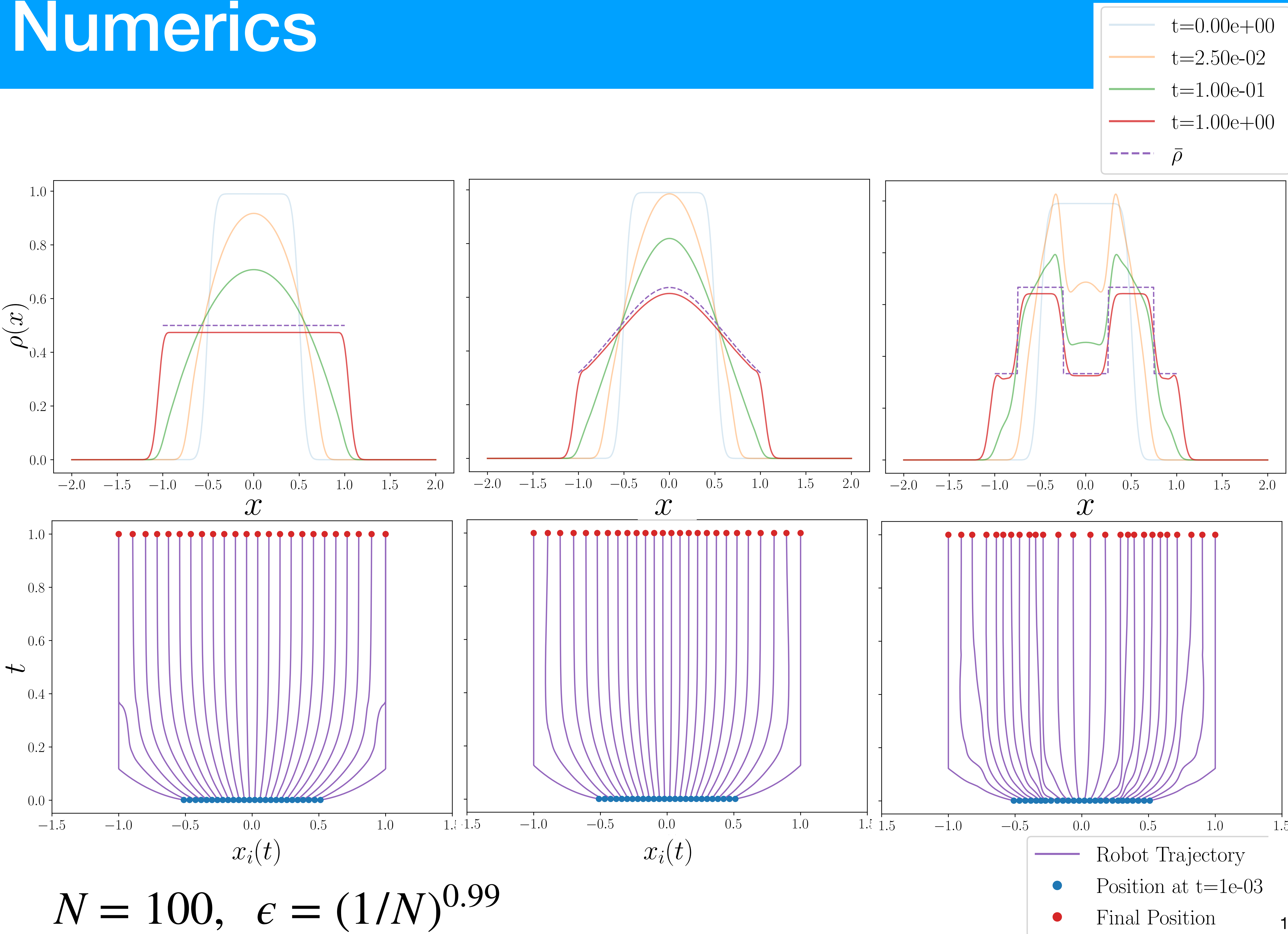
- Particle method equivalent to training dynamics for neural networks with a singular hidden layer, RBF activation.
- Our result identifies limiting dynamics in the over parametrized regime ($N \to +\infty$) as variance of the RBF decreases to zero ($\epsilon \to 0$), $\nu \neq 1$.
- Limiting dynamics 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 ↔ continuum)
- Particle method + regularization = blob method for diffusive PDEs
- Numerics

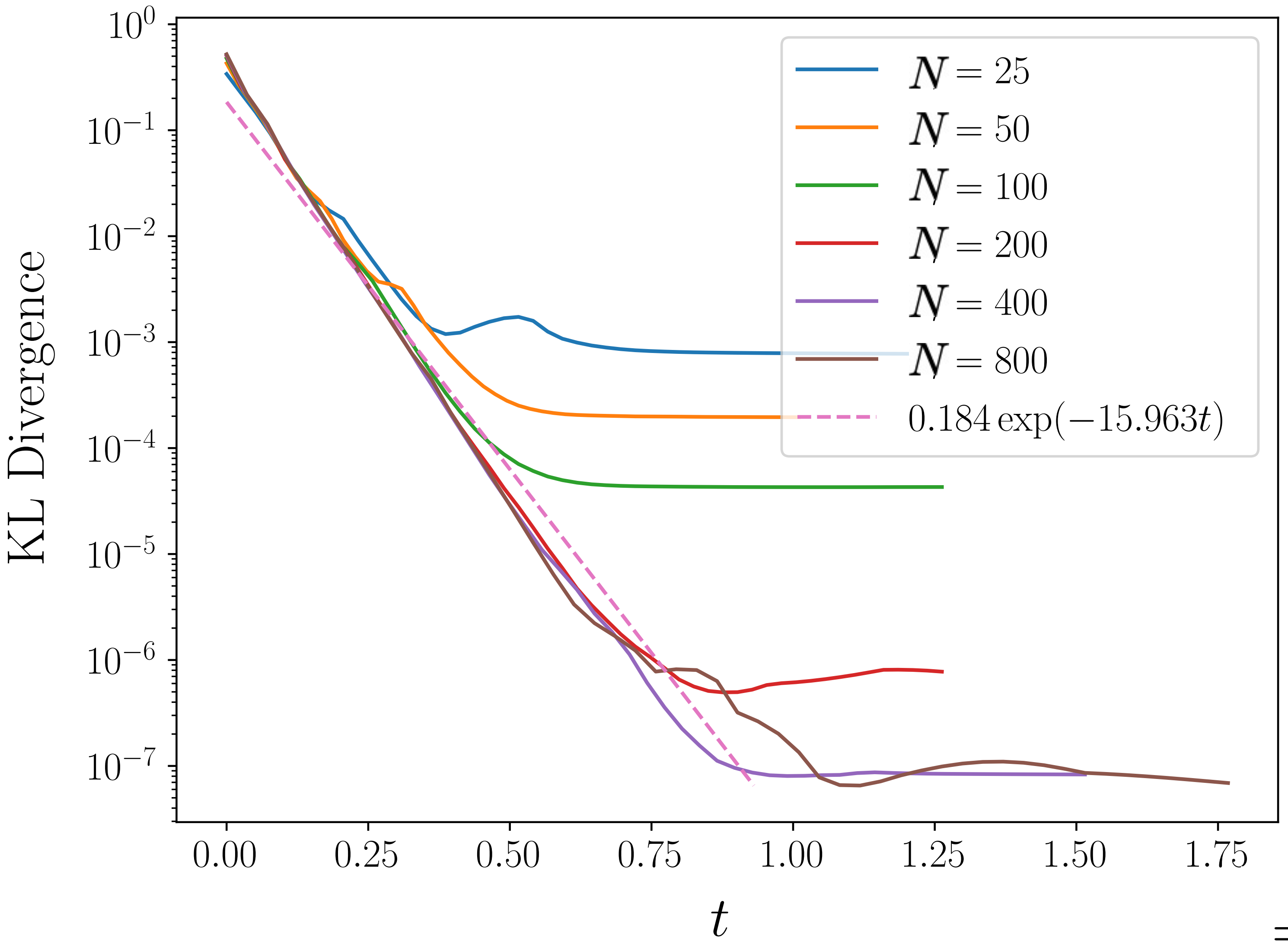
Numerics

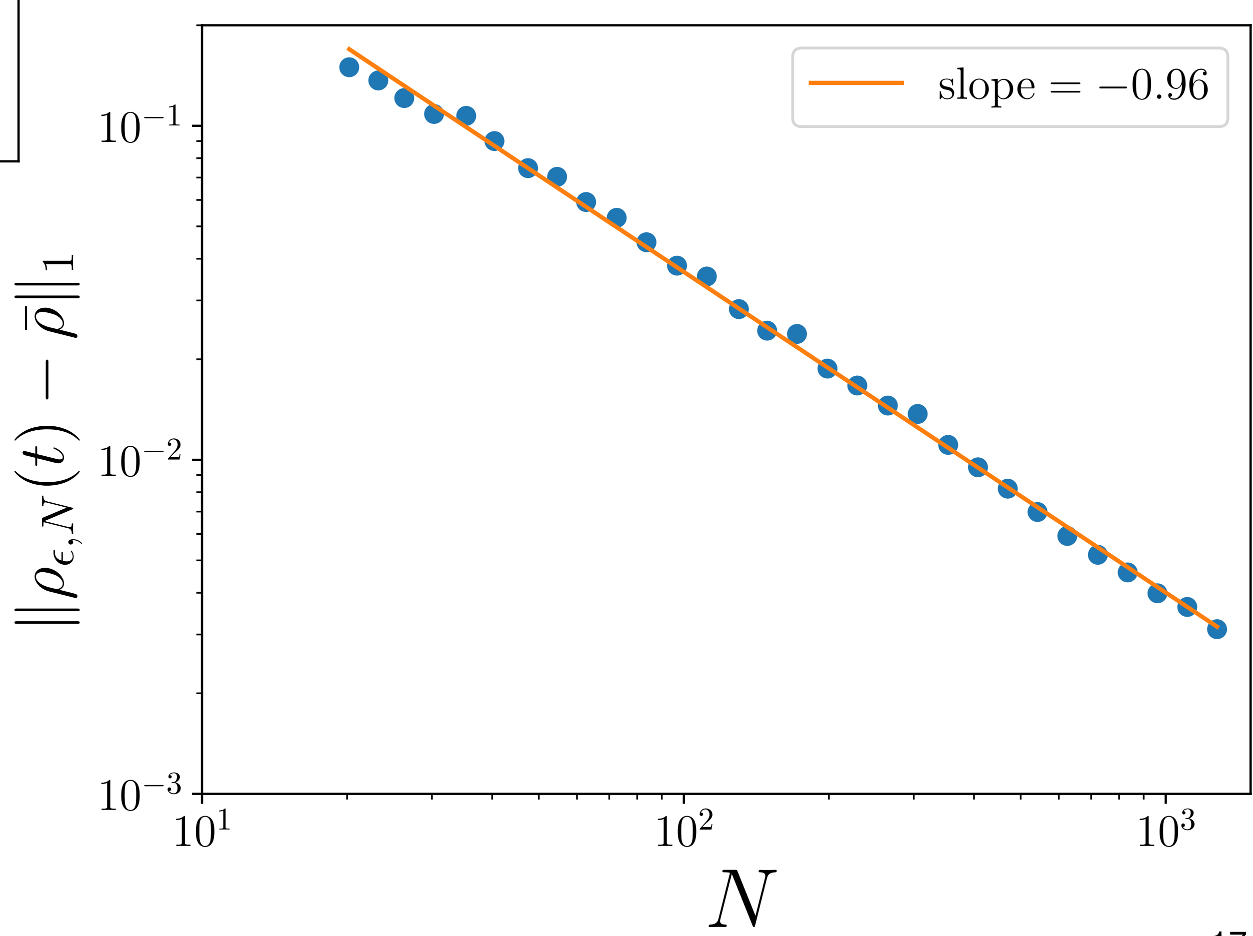


$N = 100, \quad \epsilon = (1/N)^{0.99}$

Numerics

$\bar{\rho}$ log concave





Open questions

- general $\bar{\rho}$

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

$$f_{w,z}(x) = -\int \varphi_\epsilon(x-w)\varphi_\epsilon(x-z)/\bar{\rho}(x)dx$$

- 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? $\mathcal{O}(N^{-m/d})$

- Can random batch method [Jin, Li, Liu '20] lower computational cost from $O(N^2)$ while preserving long-time behavior?

Thank you!