Analysis of a Two-Layer Neural Network via Displacement Convexity

Adel Javanmard
[with Marco Mondelli and Andrea Montanari]

Department of Data Sciences and Operation
University of Southern California

January 2019

Kavli Institute for Theoretical Physics

Non-convex high-dimensional statistics

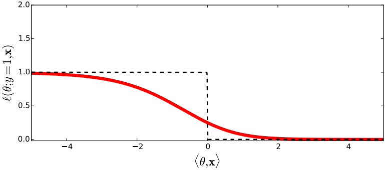
Data

$$\Big\{(y_1, x_1), (y_2, x_2), \cdots, (y_n, x_n)\Big\} \qquad \sim_{iid} \mathbb{P} \in \mathscr{P}(\mathbb{R} \times \mathbb{R}^d)$$

Goal

$$\text{minimize} \quad R(w) = \mathbb{E}\{\ell(w; x, y)\}$$

Example: ‘One-neuron neural network’



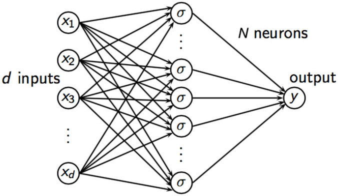
$$z_i = (y_i, x_i),\ y_i \in \{0,1\},\ x_i \in \mathbb{R}^d,\ \mathbb{P}(y_i = 1|x_i) = \sigma(w^\mathsf{T} x_i)$$

$$R(w) = \mathbb{E}\Big[\big(y - \sigma(w^\mathsf{T} x)\big)^2\Big],$$

$$\sigma(u) = \frac{1}{1+e^{-u}}\,.$$

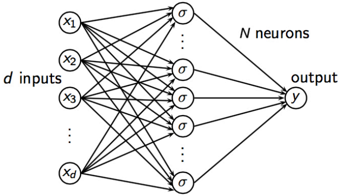
This talk

- More complicated models (two-layers NNs)



This talk

- More complicated models (two-layers NNs)



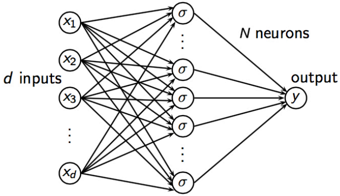
  - Learning a function f on a compact convex domain using simple components:

$$\hat{f}(x; w) = \frac{1}{N} \sum_{i=1}^{N} \sigma(x; w_i)$$

  - $\sigma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a component function ('neuron' or 'unit')

This talk

- More complicated models (two-layers NNs)



  - Learning a function f on a compact convex domain using simple components:

$$\hat{f}(x; w) = \frac{1}{N} \sum_{i=1}^{N} \sigma(x; w_i)$$

  - $\sigma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a component function ('neuron' or 'unit')
  - Learn parameters $\{w_i\}_{i \le N}$ by minimizing

$$R_N(w) = \mathbb{E}[(y - \hat{f}(x; w))^2]$$

Applications

The idea of learning f as linear combination of single components has been also studied in:

- Sparse deconvolution [Donoho 92; Candès, Fernandez-Granda 2014]
- Kernel ridge regression and random feature methods [Cristianini,Shawe-Taylor 2000; Rahimi, Recht 2008]
- Boosting [Schapire 2003; Friedman 2001]

Applications

The idea of learning f as linear combination of single components has been also studied in:

- Sparse deconvolution [Donoho 92; Candès, Fernandez-Granda 2014]
- Kernel ridge regression and random feature methods [Cristianini,Shawe-Taylor 2000; Rahimi, Recht 2008]
- Boosting [Schapire 2003; Friedman 2001]

Challenge: Risk function $R_N(w)$ is highly non-convex!

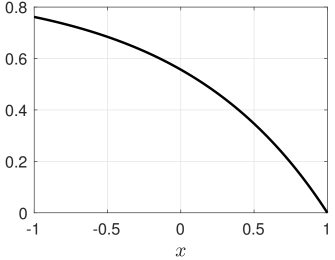
Outline

Model

Data model

Data (x_i, y_i) i.i.d. with $x_i \sim \mathsf{Unif}(\Omega)$ and $y_i = f(x_i) + \varepsilon_i$

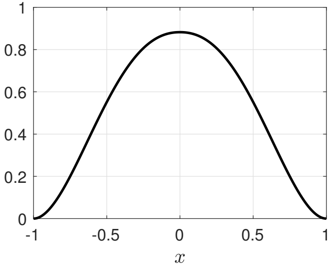
- Ω bounded and convex
- $f : \Omega \to \mathbb{R}_{\geq 0}$
- f α-strongly concave: $\langle z, \nabla^2 f(x) z \rangle \leq -\alpha |z|^2$
- f is a smooth function
- the noise terms ε_i are i.i.d subgaussian with $\mathbb{E}(\varepsilon_i | x_i) = 0$.



Minimize population risk

$$R_N(w) = \mathbb{E}\left\{\left[y - \frac{1}{N}\sum_{i=1}^{N}\sigma(x, w_i)\right]^2\right\}$$

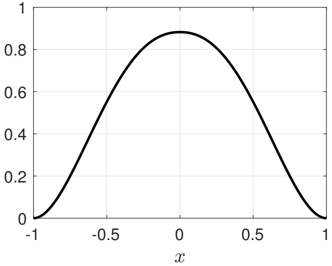
- σ "bump"-like: $\sigma(x; w_i) = K^\delta(x - w_i)$
- $K^\delta(x) = \delta^{-d}K(x/\delta)$



Minimize population risk

$$R_N(w) = \mathbb{E}\Big\{\Big[y - \frac{1}{N}\sum_{i=1}^{N}\sigma(x, w_i)\Big]^2\Big\}$$

- σ "bump"-like: $\sigma(x; w_i) = K^\delta(x - w_i)$
- $K^\delta(x) = \delta^{-d}K(x/\delta)$



Non-convex optimization problem!

Landscape analysis

Landscape analysis?

Often by proving that the landscape is nice or assume that the initialization is close enough to the global minimum.

Partial success...

[Arora, Bhaskara, Ge, Ma, 2014; Janzamin, Sedghi, Anandkumar, 2015; Ge, Lee, Ma, 2017; Soltanolkotabi, Javanmard, Lee, 2017; Zhang, Lee, Jordan, 2017; Zhong, Song, Jain, Bartlett, Dhillon, 2017; ...]

Landscape analysis

Landscape analysis?
Often by proving that the landscape is nice or assume that the initialization is close enough to the global minimum.

Partial success. . .
[Arora, Bhaskara, Ge, Ma, 2014; Janzamin, Sedghi, Anandkumar, 2015; Ge, Lee, Ma, 2017; Soltanolkotabi, Javanmard, Lee, 2017; Zhang, Lee, Jordan, 2017; Zhong, Song, Jain, Bartlett, Dhillon, 2017; . . .]

We do not follow this strategy in our work!

(noisy) Stochastic gradient descent

$$w_i^{k+1} = \mathsf{P}\left\{ w_i^k - \varepsilon \nabla K^\delta(x_k - w_i^k) \left(y_k - \frac{1}{N}\sum_{i=1}^{N} \sigma(x; w_i^k) \right) + \sqrt{2\varepsilon\tau}\, g_i^k \right\}$$

- constant step size ε
- noise term $\sqrt{2\varepsilon\tau}\, g_i^k$ added for smoothness
- $\mathsf{P} =$ orthogonal projection (onto set Ω)

(noisy) Stochastic gradient descent

$$w_i^{k+1} = \mathsf{P}\left\{ w_i^k - \varepsilon \nabla K^\delta(x_k - w_i^k)\left(y_k - \frac{1}{N}\sum_{i=1}^{N} \sigma(x; w_i^k)\right) + \sqrt{2\varepsilon\tau}\, g_i^k \right\}$$

- constant step size ε
- noise term $\sqrt{2\varepsilon\tau}\, g_i^k$ added for smoothness
- $\mathsf{P} =$ orthogonal projection (onto set Ω)

One-Pass: each data point is visited once

Questions

- Does SGD take us to global minimum of $R_N(w)$?
- Number of iterations k to achieve risk $R_N(w^k) \leq R_{\text{target}}$?
- Scaling with N, d, δ?

Main result

Convergence of SGD

Theorem (Javanmard, Mondelli, Montanari, 2018)

Consider the SGD update with initialization $(w_i^0)_{i\le N} \sim_{\text{i.i.d}} \rho^\delta_{\text{init}}$ *and constant step size* ε*. Suppose that the regression function* f *is* α*-strongly concave. Then for any* $k \le T/\varepsilon$*, the following holds with probability at least* $1 - z^{-2}$*,*

$$R_N(w^k) \le R_N(w^0)e^{-2\alpha k\varepsilon} + 8\tau \log|\Omega| + \Delta(N, \varepsilon, d, \delta, z)\,,$$

$$\lim_{\delta\to 0} \lim_{N\to\infty,\varepsilon\to 0} \Delta(N, d, \varepsilon, \delta, z) = 0\,.$$

where $\Delta(N, d, \varepsilon, \delta, z) = \cdots$

Convergence of SGD

Theorem (Javanmard, Mondelli, Montanari, 2018)

Consider the SGD update with initialization $(w_i^0)_{i\le N} \sim_{\text{i.i.d}} \rho^{\delta}_{\text{init}}$ *and constant step size* ε. *Suppose that the regression function* f *is* α*-strongly concave. Then for any* $k \le T/\varepsilon$, *the following holds with probability at least* $1 - z^{-2}$,

$$R_N(w^k) \le R_N(w^0)e^{-2\alpha k\varepsilon} + 8\tau \log|\Omega| + \Delta(N, \varepsilon, d, \delta, z)\,,$$

$$\lim_{\delta\to 0}\lim_{N\to\infty,\varepsilon\to 0} \Delta(N, d, \varepsilon, \delta, z) = 0\,.$$

where $\Delta(N, d, \varepsilon, \delta, z) = \cdots$

- Exponential convregence
- We can get arbitrarily small risk at a **dimension-free** rate!

Stochastic gradient descent (SGD) and the viscous porous medium PDE

SGD minimizes population risk

$$R_N(w) = \mathbb{E}\Big\{\Big(y - \frac{1}{N}\sum_{j=1}^{N}\sigma(x, w_j)\Big)^2\Big\}$$

SGD minimizes population risk

$$
\begin{aligned}
R_N(w) &= \mathbb{E}\Big\{\Big(y - \frac{1}{N}\sum_{j=1}^{N}\sigma(x, w_j)\Big)^2\Big\} \\
&= R_{\#} + \frac{2}{N}\sum_{i=1}^{N} V(w_i) + \frac{1}{N^2}\sum_{i,j=1}^{N} U(w_i, w_j)\,,
\end{aligned}
$$

SGD minimizes population risk

$$
\begin{aligned}
R_N(w) &= \mathbb{E}\Big\{\Big(y - \frac{1}{N}\sum_{j=1}^{N}\sigma(x, w_j)\Big)^2\Big\} \\
&= R_{\#} + \frac{2}{N}\sum_{i=1}^{N} V(w_i) + \frac{1}{N^2}\sum_{i,j=1}^{N} U(w_i, w_j)\,,
\end{aligned}
$$

$$
\begin{aligned}
V(w) &\equiv -\mathbb{E}\{y\,\sigma(x, w)\}\,, \\
U(w_1, w_2) &\equiv \mathbb{E}\{\sigma(x, w_1)\sigma(x, w_2)\}\,.
\end{aligned}
$$

SGD minimizes population risk

$$
\begin{aligned}
R_N(w) &= \mathbb{E}\Big\{\Big(y - \frac{1}{N}\sum_{j=1}^{N}\sigma(x, w_j)\Big)^2\Big\} \\
&= R_{\#} + \frac{2}{N}\sum_{i=1}^{N} V(w_i) + \frac{1}{N^2}\sum_{i,j=1}^{N} U(w_i, w_j)\,,
\end{aligned}
$$

$$
\begin{aligned}
V(w) &\equiv -\mathbb{E}\{y\,\sigma(x, w)\}\,, \\
U(w_1, w_2) &\equiv \mathbb{E}\{\sigma(x, w_1)\sigma(x, w_2)\}\,.
\end{aligned}
$$

- **Exchangeable!**
- $U(\,\cdot\,,\,\cdot\,) \succeq 0$

Exchangeability $\Rightarrow$

$R_N(w)$ depends on $w_1, \dots, w_N$ only through $\widehat{\rho}^{(N)} = \sum_{i=1}^N \delta_{w_i}/N$:

$$R : \mathscr{P}(\Omega) \to \mathbb{R}$$

$$R(\rho) \equiv R_{\#} + 2 \int V(w)\, \rho(\mathrm{d}w) + \int U(w_1, w_2)\, \rho(\mathrm{d}w_1)\, \rho(\mathrm{d}w_2)$$

Exchangeability $\Rightarrow$

$R_N(w)$ depends on $w_1, \dots, w_N$ only through $\widehat{\rho}^{(N)} = \sum_{i=1}^N \delta_{w_i}/N$:

$$R : \mathscr{P}(\Omega) \to \mathbb{R}$$

$$R(\rho) \equiv R_{\#} + 2\int V(w)\,\rho(\mathrm{d}w) + \int U(w_1, w_2)\,\rho(\mathrm{d}w_1)\,\rho(\mathrm{d}w_2)$$

- Algorithms for optimizig $R(\rho)$ [Bengio et al. 2006]
- Use ∞-dimensional formulation to analyze SGD [Mei et al., 2018]

Exchangeability $\Rightarrow$

$R_N(w)$ depends on $w_1, \dots, w_N$ only through $\widehat{\rho}^{(N)} = \sum_{i=1}^{N} \delta_{w_i}/N$:

$$R : \mathscr{P}(\Omega) \to \mathbb{R}$$

$$R(\rho) \equiv R_{\#} + 2\int V(w)\, \rho(\mathrm{d}w) + \int U(w_1, w_2)\, \rho(\mathrm{d}w_1)\, \rho(\mathrm{d}w_2)$$

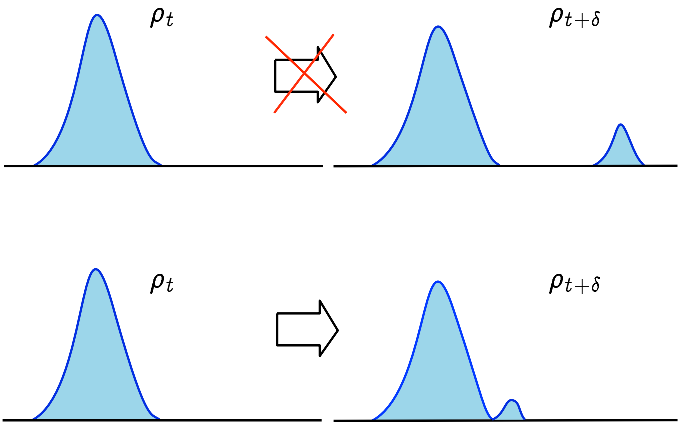
- Algorithms for optimizig $R(\rho)$ [Bengio et al. 2006]
- Use ∞-dimensional formulation to analyze SGD [Mei et al., 2018]

Is SGD a descent algorithm for $R(\rho)$?

$$R(\rho) \equiv R_{\#} + 2\int V(w)\,\rho(\mathrm{d}w) + \int U(w_1, w_2)\,\rho(\mathrm{d}w_1)\,\rho(\mathrm{d}w_2)$$

- $R(\rho)$ convex
- Did we trivialize the problem?

Not at all!



- Not all 'small changes' in ρ can be realized by SGD dynamics
- Mass must be conserved locally

Does SGD have a scaling limit?

Evolution in the space of distributions ρ?

Scaling limit: A flow in $\mathscr{P}(\Omega)$

Claim $k = t/\varepsilon$, $N \to \infty$, $\varepsilon \to 0$:

$$\hat{\rho}_k^{(N)} \equiv \frac{1}{N}\sum_{i=1}^{N} \delta_{w_i^k} \quad \Rightarrow \quad \rho_t$$

$$\partial_t \rho_t = \nabla_w \cdot \Big(\rho_t \nabla_w \Psi^\delta(w;\rho_t)\Big) + \tau \Delta \rho_t(w)\,,$$

$$\Psi^\delta(w;\rho) \equiv \frac{\delta R(\rho)}{\delta \rho(w)} = V(w) + \int U(w,w')\,\rho(\mathrm{d}w)$$

$$= -K^\delta * f(w) + K^\delta * K^\delta * \rho(w)\,.$$

Scaling limit: A flow in $\mathscr{P}(\Omega)$

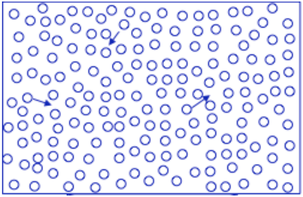
Claim $k = t/\varepsilon$, $N \to \infty$, $\varepsilon \to 0$:

$$\hat{\rho}_k^{(N)} \equiv \frac{1}{N} \sum_{i=1}^{N} \delta_{w_i^k} \quad \Rightarrow \quad \rho_t$$

$$\begin{aligned}
\partial_t \rho_t &= \nabla_w \cdot \Big(\rho_t \nabla_w \Psi^\delta(w; \rho_t) \Big) + \tau \Delta \rho_t(w) \,, \\
\Psi^\delta(w; \rho) &\equiv \frac{\delta R(\rho)}{\delta \rho(w)} = V(w) + \int U(w, w')\, \rho(\mathrm{d}w) \\
&= -K^\delta * f(w) + K^\delta * K^\delta * \rho(w) \,.
\end{aligned}$$

High-level description

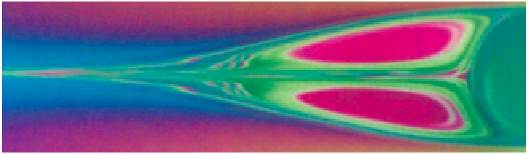
Stochastic gradient descent (SGD)



Microscopic dynamics of a gas with N particles

$$\downarrow$$

Viscous porous medium equation (PDE)



Macroscopic dynamics of a gas through a porous medium

Related work (connection of SGD to PDE)

- Non-quantitative (no convergence rates)
- No (Neumann) boundary conditions
  - Mei, Montanari, Nguyen [PNAS 2018]
  - Rotskoff, Vanden-Eijnden arXiv:1805.00915
  - Sirignano, Spiliopoulos arXiv:1805.01053
  - Chizac, Bach arXiv:1805.09545

- Convergence in Poly(d) time, but for a different continuous flow.
  - Wei, Lee, Liu, Ma arXiv:1810.05369

Proof sketch

Step 1: SGD is close to the PDE ($N \to \infty$, $\varepsilon \to 0$)

- Propagation-of-chaos argument. [Sznitman 91, Mei et al. 2018]

Step 1: SGD is close to the PDE ($N \to \infty$, $\varepsilon \to 0$)

- Propagation-of-chaos argument. [Sznitman 91, Mei et al. 2018]
  SGD update:

$$w_i^{k+1} = \mathsf{P}\{w_i^k + F_i(x_{k+1}, y_{k+1}; w^k)\}$$

$$F_i(x_{k+1}, y_{k+1}, w^k) = -\varepsilon \nabla \sigma(x_{k+1}, w_i^k)\Big(y_{k+1} - \frac{1}{N}\sum_{i=1}^{N} \sigma(x_{k+1}, w_i^k)\Big) + \sqrt{2\tau\varepsilon}\, g_i^{k+1}\,.$$

Step 1: SGD is close to the PDE ($N \to \infty$, $\varepsilon \to 0$)

- Propagation-of-chaos argument. [Sznitman 91, Mei et al. 2018]
  SGD update:

$$w_i^{k+1} = \mathsf{P}\{w_i^k + F_i(x_{k+1}, y_{k+1}; w^k)\}$$

$$F_i(x_{k+1}, y_{k+1}, w^k) = -\varepsilon \nabla \sigma(x_{k+1}, w_i^k)\Big(y_{k+1} - \frac{1}{N}\sum_{i=1}^{N} \sigma(x_{k+1}, w_i^k)\Big) + \sqrt{2\tau\varepsilon}\, g_i^{k+1}\,.$$

- Intuition $\mathcal{F}_k = \sigma(\{y_i, x_i\}_{i<k})$:

$$\begin{aligned}\mathbb{E}\{F_i(x_{k+1}, y_{k+1}, w^k)|\mathcal{F}_k\} &= -\varepsilon\, \nabla V(w_i^k) - \varepsilon\, \frac{1}{N}\sum_{j=1}^{N} \nabla_1 U(w_i^k, w_j^k) \\ &= -\varepsilon \int [\nabla V(w_i^k) + \nabla_1 U(w_i^k, w)]\, \hat{\rho}_k^{(N)}(\mathrm{d}w) \\ &= -\varepsilon\, \nabla \Psi(w_i^k, \hat{\rho}_k^{(N)})\,.\end{aligned}$$

Step 1: SGD is close to the PDE ($N \to \infty$, $\varepsilon \to 0$)

Nonlinear dynamics

$$\mathrm{d}X_t = -\nabla\Psi^\delta(X_t,\rho_t)\mathrm{d}t + \sqrt{2\tau}\mathrm{d}B_t + \mathrm{d}\Phi_t\,, \qquad X_0 \sim \rho_0$$

- $(\Phi_t)_{t\geq 0}$ enforces the reflecting boundary condition
- $(B_t)_{t\geq 0}$ is standard d-dim Brownian motion

Step 1: SGD is close to the PDE ($N \to \infty$, $\varepsilon \to 0$)

Nonlinear dynamics

$$\mathrm{d}X_t = -\nabla \Psi^{\delta}(X_t, \rho_t)\mathrm{d}t + \sqrt{2\tau}\mathrm{d}B_t + \mathrm{d}\Phi_t\,, \qquad X_0 \sim \rho_0$$

- $(\Phi_t)_{t \geq 0}$ enforces the reflecting boundary condition
- $(B_t)_{t \geq 0}$ is standard d-dim Brownian motion

SGD update is close to the fixed point of nonlinear dynamics

$$\rho_t^{\delta} = \mathrm{Law}(X_t) \quad \forall t \in [0, T]\,.$$

Step 1: SGD is close to the PDE ($N \to \infty$, $\varepsilon \to 0$)

Nonlinear dynamics

$$\mathrm{d}X_t = -\nabla\Psi^\delta(X_t, \rho_t)\mathrm{d}t + \sqrt{2\tau}\mathrm{d}B_t + \mathrm{d}\Phi_t\,, \qquad X_0 \sim \rho_0$$

- $(\Phi_t)_{t\geq 0}$ enforces the reflecting boundary condition
- $(B_t)_{t\geq 0}$ is standard d-dim Brownian motion

SGD update is close to the fixed point of nonlinear dynamics

$$\rho_t^\delta = \mathrm{Law}(X_t) \quad \forall t \in [0, T]\,.$$

- By an application of Ito's integral, ρ_t^δ is a weak solution of

$$\partial_t \rho_t(x) = \nabla(\rho_t(x)\nabla^\delta \Psi^\delta(x, \rho_t)) + \tau\Delta\rho_t(x)$$

$$\langle n(x), \rho_t(x)\nabla\Psi^\delta(x, \rho_t) + \tau\nabla\rho_t(x)\rangle = 0, \quad \forall x \in \partial\Omega.$$

Neumann boundary condition

Step 2: the PDE is close to Viscous Porous Medium Equation ($\delta \to 0$)

PDE ($\delta > 0$)

$$\partial_t \rho_t(w) = \nabla(\rho_t(w)\nabla^\delta \Psi^\delta(w, \rho_t)) + \tau \Delta \rho_t(w)$$

$$\langle n(w), \rho_t(w)\nabla\Psi^\delta(w, \rho_t) + \tau \nabla \rho_t(w)\rangle = 0, \quad \forall w \in \partial\Omega.$$

Recall that

$$\Psi^\delta(w, \rho_t) = -K^\delta * f(w) + K^\delta * K^\delta * \rho(w)$$

Step 2: the PDE is close to Viscous Porous Medium Equation ($\delta \to 0$)

PDE ($\delta > 0$)

$$\partial_t \rho_t(w) = \nabla(\rho_t(w)\nabla^\delta \Psi^\delta(w,\rho_t)) + \tau\Delta\rho_t(w)$$
$$\langle n(w), \rho_t(w)\nabla\Psi^\delta(w,\rho_t) + \tau\nabla\rho_t(w)\rangle = 0, \quad \forall w \in \partial\Omega.$$

Recall that

$$\Psi^\delta(w,\rho_t) = -K^\delta * f(w) + K^\delta * K^\delta * \rho(w)$$

As $\delta \to 0$, the weak solution of the above converges to a weak solution of the (PME):

Viscous Porous Medium Equation

$$\partial_t \rho_t(x) = -\nabla(\rho_t(w)\nabla f(w)) + \frac{1}{2}\Delta(\rho_t^2(w)) + \tau\Delta\rho_t(w)$$
$$\langle n(w), \rho_t(w)\nabla(f(w) - \rho_t(w) - \tau\nabla\rho_t(w))\rangle = 0, \quad \forall w \in \partial\Omega.$$

Indeed, there is a unique weak solution to PME.

Viscous Porous Medium Equation

$$\partial_t \rho_t(x) = -\nabla(\rho_t(w)\nabla f(w)) + \frac{1}{2}\Delta(\rho_t^2(w)) + \tau\Delta\rho_t(w)$$
$$\langle n(w), \rho_t(w)\nabla(f(w) - \rho_t(w) - \tau\nabla\rho_t(w))\rangle = 0, \quad \forall w \in \partial\Omega.$$

What does this minimize?

Viscous Porous Medium Equation

$$\partial_t \rho_t(x) = -\nabla(\rho_t(w)\nabla f(w)) + \frac{1}{2}\Delta(\rho_t^2(w)) + \tau\Delta\rho_t(w)$$
$$\langle n(w), \rho_t(w)\nabla(f(w) - \rho_t(w) - \tau\nabla\rho_t(w))\rangle = 0, \quad \forall w \in \partial\Omega.$$

What does this minimize?

Free energy function:

$$F_\tau(\rho) = \frac{1}{2}R(\rho) - \tau \mathrm{Ent}(\rho)\,,$$
$$R(\rho) \equiv \int_\Omega \|f(w) - \rho(w)\|_2^2 \, \mathrm{d}w \qquad \text{(risk)}$$
$$\mathrm{Ent}(\rho) \equiv -\int \rho(w)\log\rho(w)\, \mathrm{d}w \qquad \text{(entropy)}$$

PME solution minimizes the free energy $F_\tau(\rho)$

Proposition

If ρ_t is the solution of viscous PME, then $F_\tau(\rho_t)$ is non-increasing:

$$\frac{d}{dt}F_\tau(\rho_t) = -\int \left\|\nabla\Big(\rho(w) - f(w) + \tau \log \rho_t(w)\Big)\right\|_2^2 \rho_t(w)dw < 0\,.$$

Step 3: A result on viscous PME

Theorem (Carrillo et al. [CJMTU] 2001)

For the free energy function $F_\tau(\rho)$ and the viscous PME solution ρ_t we have

1. *There exists a unique minimizer ρ^* of the free energy $F_\tau(\rho)$.*
2. *For any $t \geq 0$, we have*

$$F_\tau(\rho_t) - F_\tau(\rho^*) \leq (F_\tau(\rho_0) - F_\tau(\rho^*))e^{-2\alpha t}.$$

(Recall) f is α-strongly concave.

[Carrillo, Jüngel, Markowich, Toscani, Unterreiter (2001), Carrillo, macCann, Villani (2003),(2006)]

- Dimension free convergence rate (no dependent on d)!
- Displacement convexity plays a key role!

Step 3: A result on viscous PME

Theorem (Carrillo et al. [CJMTU] 2001)

For the free energy function $F_\tau(\rho)$ and the viscous PME solution ρ_t we have

1. *There exists a unique minimizer ρ^* of the free energy $F_\tau(\rho)$.*
2. *For any $t \geq 0$, we have*

$$F_\tau(\rho_t) - F_\tau(\rho^*) \leq (F_\tau(\rho_0) - F_\tau(\rho^*))e^{-2\alpha t}\,.$$

(Recall) f is α-strongly concave.

[Carrillo, Jüngel, Markowich, Toscani, Unterreiter (2001), Carrillo, macCann, Villani (2003),(2006)]

- Dimension free convergence rate (no dependent on d)!
- Displacement convexity plays a key role!

Displacement convexity

Wasserstein geodesics

W_2 (Wasserstein) distance between two probability measures ρ_0, ρ_1

$$W_2(\rho_0, \rho_1)^2 = \inf_{\gamma \in \Gamma(\rho_0, \rho_1)} \int \|x - y\|_2^2 \, \gamma(\mathrm{d}x, \mathrm{d}y) \, .$$

Wasserstein geodesics

W_2 (Wasserstein) distance between two probability measures ρ_0, ρ_1

$$W_2(\rho_0, \rho_1)^2 = \inf_{\gamma \in \Gamma(\rho_0, \rho_1)} \int \|x - y\|_2^2 \, \gamma(\mathrm{d}x, \mathrm{d}y)\,.$$

- **W_2 geodesic between ρ_0 and ρ_1:**
  Let $(X_0, X_1) \sim \gamma_*$ and define ρ_t to be the distribution of

  $$X_t = (1-t)X_0 + tX_1$$

  (γ^* the optimal coupling)
- The curve $t \mapsto \rho_t$ is the geodesic between ρ_0 and ρ_1.

Displacement convexity

Displacement Convexity

- Convexity along geodesics
- A function $F(\rho)$ is λ-strongly displacement convex if

$$(1-t)F(\rho_0) + tF(\rho_1) - F(\rho_t) \geq \frac{1}{2}\lambda t(1-t)\,.$$

Recall that ρ_t is the W_2 geodesic between ρ_0 and ρ_1.

Displacement convexity

Displacement Convexity

- Convexity along geodesics
- A function $F(\rho)$ is λ-strongly displacement convex if

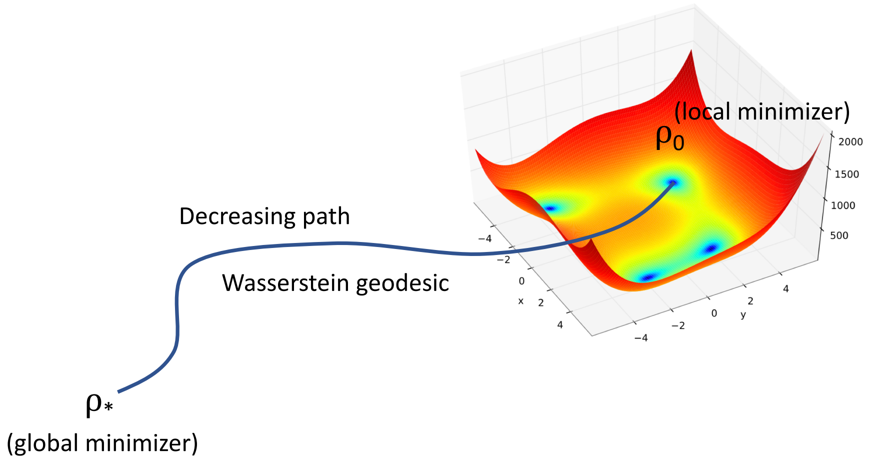
$$(1-t)F(\rho_0) + tF(\rho_1) - F(\rho_t) \geq \frac{1}{2}\lambda t(1-t)\,.$$

Recall that ρ_t is the W_2 geodesic between ρ_0 and ρ_1.

A nice property:
Our free energy $F_\tau(\rho) = \frac{1}{2}R(\rho) - \tau\,\mathrm{Ent}(\rho)$ is strongly displacement convex.

How does displacement convexity help?

How does displacement convexity help?



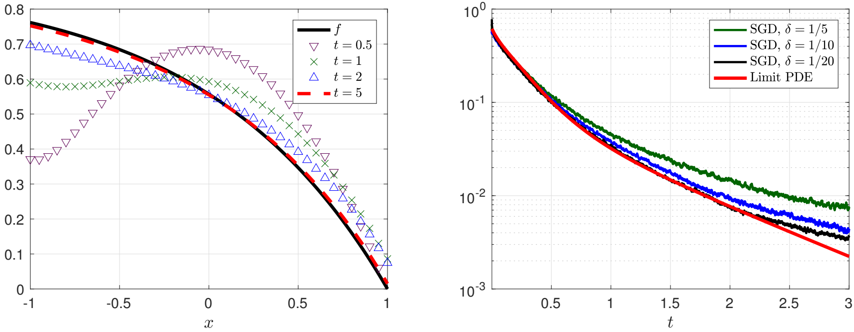
Numerical experiments

Simulation setup 1

- $d = 1$, $\Omega = [-1, 1]$, $f(x) = c(e - e^x)$
- The kernel is chosen as

$$K(x) = C\kappa(\|x\|), \quad \kappa(t) = \begin{cases} 1 - t^2 - 2t^3 + 2t^4 & \text{for } t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- ρ_0 is truncated gaussian with $\sigma = 1/3$ and we choose $N = 200$.

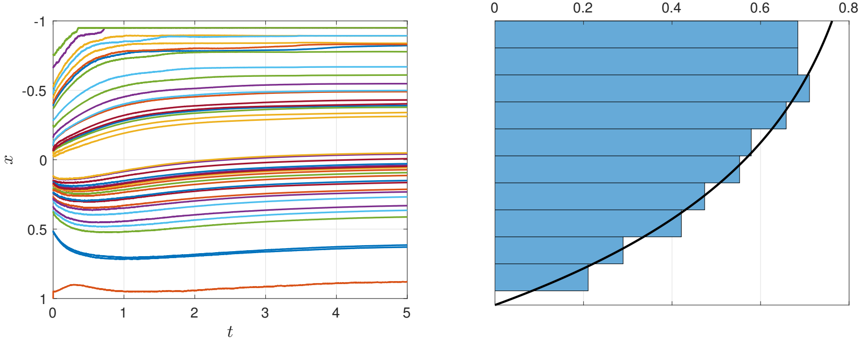


Simulation setup 1

- $d = 1$, $\Omega = [-1, 1]$, $f(x) = c(e - e^x)$
- The kernel is chosen as

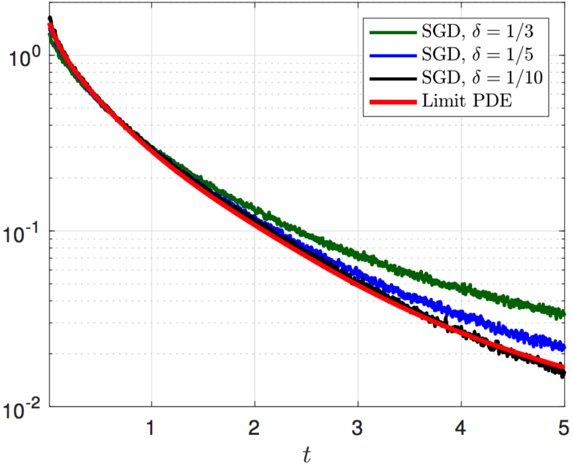
$$K(x) = C\kappa(\|x\|), \quad \kappa(t) = \begin{cases} 1 - t^2 - 2t^3 + 2t^4 & \text{for } t \le 1 \\ 0 & \text{otherwise} \end{cases}$$

- ρ_0 is truncated gaussian with $\sigma = 1/3$ and we choose $N = 200$.



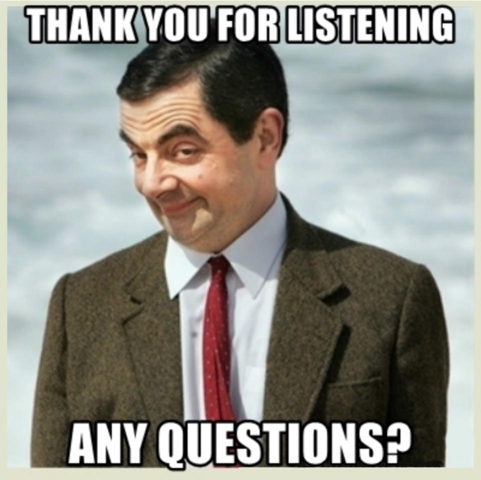
Simulation setup 2

- $d = 2$, $\Omega = [-1, 1]^d$, $f(x) = c_2(c_1 - \log(e^{\langle q_1, x\rangle}) + e^{\langle q_2, x\rangle})$
- same kernel as berfore



Conclusion

- Learning functions on compact domain using simple components ('bump-like')
- Formulate the problem as learning a two-layer neural net
- Stochastic gradient descent (SGD) is close to viscous porous medium equation
- dimension-free convergence rate to global optimum


THANK YOU FOR LISTENING
ANY QUESTIONS?