Scaling limits of SGD over large networks

Zaid Harchaoui[1,3], Sewoong Oh[1,4], Soumik Pal[2], Raghav Somani[1] and Raghav Tripathi[2]

[1]UW CSE, [2]UW Math, [3]UW Statistics & [4]Google

February 2, 2023





Plan

- Introduction: Interacting particle system
- 2 layer Neural Networks
- Optimization on graphons
- Future directions and Deep Neural Networks

Prologue: Interacting particle systems

Problem

For $n \in \mathbb{N}$, consider $R_n(x) \coloneqq \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{1}{2}(x_i - x_j)^2, \quad \text{for } x \in \mathbb{R}^n$. Minimize R_n.

Prologue: Interacting particle systems

Problem

For $n \in \mathbb{N}$, consider $R_n(x) \coloneqq \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{1}{2}(x_i - x_j)^2$, for $x \in \mathbb{R}^n$. Minimize R_n.

- Can perform GD to solve - particle gradient flow:

$$\begin{aligned} \mathrm{d} X_i(t) &= -n\, \partial_i R_n(X(t))\, \mathrm{d}t \\ &= -\frac{1}{n} \sum_{j=1}^{n} (X_i(t) - X_j(t))\, \mathrm{d}t \end{aligned} \qquad \forall\, i \in [n].$$

Prologue: Interacting particle systems

Problem

For $n \in \mathbb{N}$, consider $R_n(x) \coloneqq \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{1}{2}(x_i - x_j)^2$, for $x \in \mathbb{R}^n$. Minimize R_n.

- Can perform GD to solve - particle gradient flow:

$$\begin{aligned} \mathrm{d} X_i(t) &= -n\, \partial_i R_n(X(t))\, \mathrm{d}t \\ &= -\frac{1}{n} \sum_{j=1}^{n} (X_i(t) - X_j(t))\, \mathrm{d}t \end{aligned} \qquad \forall\, i \in [n].$$

- R_n is *permutation invariant* and hence a function R of empirical measure

Prologue: Interacting particle systems

Problem

For $n \in \mathbb{N}$, consider $R_n(x) \coloneqq \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{1}{2}(x_i - x_j)^2$, for $x \in \mathbb{R}^n$. Minimize R_n.

- Can perform GD to solve - particle gradient flow:

$$\begin{aligned} \mathrm{d} X_i(t) &= -n\, \partial_i R_n(X(t))\, \mathrm{d}t \\ &= -\frac{1}{n} \sum_{j=1}^{n} (X_i(t) - X_j(t))\, \mathrm{d}t \end{aligned} \qquad \forall\, i \in [n].$$

- R_n is *permutation invariant* and hence a function R of empirical measure defined by

$$R(\rho) \coloneqq \iint_{\mathbb{R}\times\mathbb{R}} \frac{1}{2}(x-y)^2\, \mathrm{d}\rho(x)\, \mathrm{d}\rho(y) = \mathrm{Var}[\rho] .$$

Prologue: Interacting particle systems

Problem

For $n \in \mathbb{N}$, consider $R_n(x) \coloneqq \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{1}{2}(x_i - x_j)^2$, for $x \in \mathbb{R}^n$. Minimize R_n.

- Can perform GD to solve - particle gradient flow:

$$\begin{aligned} \mathrm{d}X_i(t) &= -n\,\partial_i R_n(X(t))\,\mathrm{d}t \\ &= -\frac{1}{n}\sum_{j=1}^{n}(X_i(t) - X_j(t))\,\mathrm{d}t \qquad \forall\, i \in [n]. \end{aligned}$$

- R_n is *permutation invariant* and hence a function R of empirical measure defined by

$$R(\rho) \coloneqq \iint_{\mathbb{R}\times\mathbb{R}} \frac{1}{2}(x-y)^2\,\mathrm{d}\rho(x)\,\mathrm{d}\rho(y) = \mathrm{Var}[\rho]\,.$$

- It is known that $\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i(t)} =: \hat{\rho}_t^{(n)} \xrightarrow{n\to\infty} \rho_t$.
- $t \mapsto \rho_t$ is the gradient flow of $R\colon \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}), \mathbb{W}_2)$

$$\partial_t \rho_t = -\nabla_{\mathbb{W}_2} R(\rho_t)$$

Prologue: Interacting particle systems

Problem

For $n \in \mathbb{N}$, consider $R_n(x) \coloneqq \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{1}{2}(x_i - x_j)^2$, for $x \in \mathbb{R}^n$. Minimize R_n.

- Can perform GD to solve - particle diffusion:

$$\begin{aligned} \mathrm{d}X_i(t) &= -n\,\partial_i R_n(X(t))\,\mathrm{d}t + \mathrm{d}B_i(t) \\ &= -\frac{1}{n}\sum_{j=1}^{n}(X_i(t) - X_j(t))\,\mathrm{d}t + \mathrm{d}B_i(t) \end{aligned} \qquad \forall\, i \in [n].$$

- R_n is *permutation invariant* and hence a function R of empirical measure defined by

$$R(\rho) \coloneqq \iint_{\mathbb{R}\times\mathbb{R}} \frac{1}{2}(x-y)^2\,\mathrm{d}\rho(x)\,\mathrm{d}\rho(y) = \mathrm{Var}[\rho]\ .$$

- It is known that $\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(t)} \eqqcolon \hat{\rho}_t^{(n)} \xrightarrow{n\to\infty} \rho_t$.
- $t \mapsto \rho_t$ is the gradient flow of $R\colon \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}), \mathbb{W}_2)$

$$\partial_t \rho_t = -\nabla_{\mathbb{W}_2}(R + \mathrm{Ent})(\rho_t)$$

Summary

Particle gradient flow/diffusion

Objective: $R_n\colon \mathbb{R}^n \to \mathbb{R}$

$$\mathrm{d}X_i(t) = -n\,\partial_i R_n(X(t))\,\mathrm{d}t + \mathrm{d}B_i(t)$$

Wasserstein gradient flow

Objective: $R\colon \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$

$$\partial_t \rho_t = -\nabla_{\mathbb{W}_2}(R + \mathrm{Ent})(\rho_t)$$

Summary

Particle gradient flow/diffusion

Objective: $R_n\colon \mathbb{R}^n \to \mathbb{R}$

$$\mathrm{d}X_i(t) = -n\,\partial_i R_n(X(t))\,\mathrm{d}t + \mathrm{d}B_i(t)$$

Wasserstein gradient flow

Objective: $R\colon \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$

$$\partial_t \rho_t = -\nabla_{\mathbb{W}_2}(R + \mathrm{Ent})(\rho_t)$$

Meta Theorem(s)

- Particle system gradient descent approximates the Wasserstein gradient flow of measures

Summary

Particle gradient flow/diffusion

Objective: $R_n \colon \mathbb{R}^n \to \mathbb{R}$

$$\mathrm{d}X_i(t) = -n\,\partial_i R_n(X(t))\,\mathrm{d}t + \mathrm{d}B_i(t)$$

Wasserstein gradient flow

Objective: $R \colon \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$

$$\partial_t \rho_t = -\nabla_{\mathbb{W}_2}(R + \mathrm{Ent})(\rho_t)$$

Meta Theorem(s)

- Particle system gradient descent approximates the Wasserstein gradient flow of measures

$$\rho_t \approx n^{-1} \sum_{i=1}^{n} \delta_{X_i(t)}.$$

Summary

Particle gradient flow/diffusion

Objective: $R_n : \mathbb{R}^n \to \mathbb{R}$

$$\mathrm{d}X_i(t) = -n\,\partial_i R_n(X(t))\,\mathrm{d}t + \mathrm{d}B_i(t)$$

Wasserstein gradient flow

Objective: $R \colon \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$

$$\partial_t \rho_t = -\nabla_{\mathbb{W}_2}(R + \mathrm{Ent})(\rho_t)$$

Meta Theorem(s)

- Particle system gradient descent approximates the Wasserstein gradient flow of measures

$$\rho_t \approx n^{-1} \sum_{i=1}^{n} \delta_{X_i(t)}.$$

- **Propagation of Chaos**: As n grows, any k randomly chosen particles become independent.

Summary

Particle gradient flow/diffusion

Objective: $R_n : \mathbb{R}^n \to \mathbb{R}$

$$\mathrm{d}X_i(t) = -n\,\partial_i R_n(X(t))\,\mathrm{d}t + \mathrm{d}B_i(t)$$

Wasserstein gradient flow

Objective: $R \colon \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$

$$\partial_t \rho_t = -\nabla_{\mathbb{W}_2}(R + \mathrm{Ent})(\rho_t)$$

Meta Theorem(s)

- Particle system gradient descent approximates the Wasserstein gradient flow of measures

$$\rho_t \approx n^{-1} \sum_{i=1}^{n} \delta_{X_i(t)}.$$

- **Propagation of Chaos**: As n grows, any k randomly chosen particles become independent.
- The dynamics of a randomly chosen particle in is described by McKean-Vlasov equation

$$\mathrm{d}X(t) = b(X(t), \mu_t)\,\mathrm{d}t + \mathrm{d}B_t, \qquad \mu_t = \mathrm{Law}(X_t)$$

An application: Two layer Neural Networks (NNs)

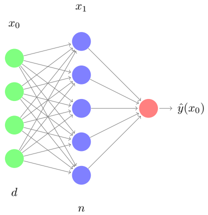


Figure: A 2-layer NN.

$$\Theta = \{\theta_1, \theta_2, \dots, \theta_n\},$$

$$\hat{y}_\Theta(x_0) = \frac{1}{n}\sum_{i=1}^{n} \sigma(\langle \theta_i, x_0 \rangle),$$

$$R_n(\Theta) = \mathbb{E}_{(X,Y)\sim\mu}\left[(Y - \hat{y}_\Theta(X))^2\right].$$

An application: Two layer Neural Networks (NNs)

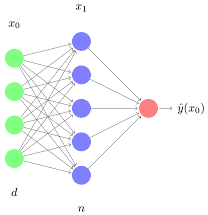


Figure: A 2-layer NN.

$$\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\},$$

$$\hat{y}_\Theta(x_0) = \frac{1}{n}\sum_{i=1}^{n} \sigma(\langle \theta_i, x_0 \rangle),$$

$$R_n(\Theta) = \mathbb{E}_{(X,Y)\sim\mu}\left[(Y - \hat{y}_\Theta(X))^2\right].$$

Minimization Problem(s):

$$R_n(\Theta) = \mathbb{E}\left[Y^2\right] + \frac{2}{n}\sum_{i=1}^{n} V(\theta_i) + \frac{1}{n^2}\sum_{i,j=1}^{n} U(\theta_i, \theta_j)$$

An application: Two layer Neural Networks (NNs)

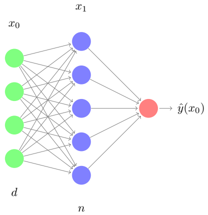


Figure: A 2-layer NN.

$$\Theta = \{\theta_1, \theta_2, \dots, \theta_n\}, \quad \rho_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\theta_i},$$

$$\hat{y}_\Theta(x_0) = \frac{1}{n}\sum_{i=1}^{n} \sigma(\langle \theta_i, x_0\rangle), \quad \hat{y}(x_0) = \int \sigma(\langle \theta, x_0\rangle)\rho_n(\mathrm{d}\theta),$$

$$R_n(\Theta) = \mathbb{E}_{(X,Y)\sim\mu}\Big[\big(Y - \hat{y}_\Theta(X)\big)^2\Big].$$

Minimization Problem(s):

$$R_n(\Theta) = \mathbb{E}\big[Y^2\big] + \frac{2}{n}\sum_{i=1}^{n} V(\theta_i) + \frac{1}{n^2}\sum_{i,j=1}^{n} U(\theta_i, \theta_j)$$

$$R(\rho) := \mathbb{E}\big[Y^2\big] + 2\int V(\theta)\,\mathrm{d}\rho(\theta) + \iint U(\theta_1, \theta_2)\,\mathrm{d}\rho(\theta_1)\,\mathrm{d}\rho(\theta_2).$$

Two layer NN continued...

Minimization Problem(s):

$$R_n(\Theta) = \mathbb{E}\left[Y^2\right] + \frac{2}{n}\sum_{i=1}^{n} V(\theta_i) + \frac{1}{n^2}\sum_{i,j=1}^{n} U(\theta_i, \theta_j)$$

$$R(\rho) := \mathbb{E}\left[Y^2\right] + 2\int V(\theta)\,\mathrm{d}\rho(\theta) + \iint U(\theta_1, \theta_2)\,\mathrm{d}\rho(\theta_1)\,\mathrm{d}\rho(\theta_2).$$

- Consider SGD on R_n with step size τ_n.
- Let $\hat{\rho}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\theta_i(t)}$, for $t = k\tau_n, \quad k \in \mathbb{N}$.

Two layer NN continued...

Minimization Problem(s):

$$R_n(\Theta) = \mathbb{E}\big[Y^2\big] + \frac{2}{n}\sum_{i=1}^{n} V(\theta_i) + \frac{1}{n^2}\sum_{i,j=1}^{n} U(\theta_i,\theta_j)$$

$$R(\rho) := \mathbb{E}\big[Y^2\big] + 2\int V(\theta)\,\mathrm{d}\rho(\theta) + \iint U(\theta_1,\theta_2)\,\mathrm{d}\rho(\theta_1)\,\mathrm{d}\rho(\theta_2).$$

- Consider SGD on R_n with step size τ_n.
- Let $\hat\rho_n(t) = \frac{1}{n}\sum_{i=1}^{n}\delta_{\theta_i(t)}$, for $t = k\tau_n$, $\quad k \in \mathbb{N}$.

Theorem [MMN '18]

If $\hat\rho_n(0) \xrightarrow{n\to\infty} \rho_0$, then $\quad \hat\rho_n(t) \xrightarrow[\substack{n\to\infty \\ \tau_n\to 0}]{\mathbb{W}_2} \rho(t)$, uniformly for $t \in [0,T]$,

where $\rho\colon t \mapsto \rho(t)$ solves

$$\partial_t\rho(t) = -\nabla_{\mathbb{W}_2} R(\rho(t)), \qquad \rho(0) = \rho_0.$$

Two layer NN continued...

Minimization Problem(s):

$$R_n(\Theta) = \mathbb{E}\big[Y^2\big] + \frac{2}{n}\sum_{i=1}^{n} V(\theta_i) + \frac{1}{n^2}\sum_{i,j=1}^{n} U(\theta_i, \theta_j)$$

$$R(\rho) := \mathbb{E}\big[Y^2\big] + 2\int V(\theta)\,\mathrm{d}\rho(\theta) + \iint U(\theta_1, \theta_2)\,\mathrm{d}\rho(\theta_1)\,\mathrm{d}\rho(\theta_2).$$

- Consider SGD on R_n with step size τ_n.
- Let $\hat{\rho}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\theta_i(t)}$, for $t = k\tau_n, \quad k \in \mathbb{N}$.

Theorem [MMN '18]

If $\hat{\rho}_n(0) \xrightarrow{n\to\infty} \rho_0$, then $\hat{\rho}_n(t) \xrightarrow[\substack{n\to\infty \\ \tau_n\to 0}]{\mathbb{W}_2} \rho(t)$, uniformly for $t \in [0, T]$,

where $\rho \colon t \mapsto \rho(t)$ solves

$$\partial_t \rho(t) = -\nabla_{\mathbb{W}_2} R(\rho(t)), \qquad \rho(0) = \rho_0.$$

And, $$\inf_{\Theta\in(\mathbb{R}^d)^n} R_n(\Theta) \xrightarrow{n\to\infty} \inf_{\rho\in\mathcal{P}(\mathbb{R}^d)} R(\rho).$$

A new world

Objective

Study large scale optimization problems over dense weighted **unlabeled graphs**.

A new world

Objective

Study large scale optimization problems over dense weighted **unlabeled graphs**.

Let $G = (V, E)$ be a graph and let A be an adjacency matrix of G.

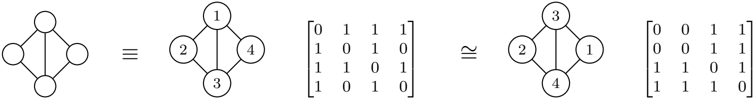


Figure: Symmetry in unlabeled graphs.

Examples

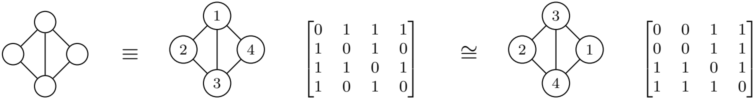
- Edge density: $h_{-}(G) = (\#\text{ of edges in } G)/\binom{n}{2}$.
- Triangle density: $h_{\triangle}(G) = (\#\text{ of }\triangle\text{s in } G)/\binom{n}{3}$.

A new world

Objective

Study large scale optimization problems over dense weighted **unlabeled graphs**.

Let $G = (V, E)$ be a graph and let A be an adjacency matrix of G.



$$\equiv \quad \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \quad \cong \quad \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

Figure: Symmetry in unlabeled graphs.

Examples

- Edge density: $h_{-}(G) = (\#\text{ of edges in } G)/\binom{n}{2}$.
- Triangle density: $h_{\triangle}(G) = (\#\text{ of } \triangle\text{s in } G)/\binom{n}{3}$.

Invariant functions

A function $F\colon \mathcal{M}_n \to \mathbb{R}$ is said to be *invariant function/graph function* if $F(A) = F(A^\sigma)$ for all permutations $\sigma \in S_n$ and $A \in \mathcal{M}_n$, where $A^\sigma(i,j) = A(\sigma(i), \sigma(j))$.

General plan and analogies

Let F be graph function. Our goal is to minimize F over large graphs.

Can perform gradient descent on finite graphs/symmetric matrices.

Exploiting the symmetry

- Think of the problem as an optimization problem on the space of ‘graphons’.
- Hope-Pray-Prove! The gradient descent process on finite graphs/symmetric matrices converge to a limit as $n \to \infty$.

General plan and analogies

Let F be graph function. Our goal is to minimize F over large graphs.

Can perform gradient descent on finite graphs/symmetric matrices.

Exploiting the symmetry

- Think of the problem as an optimization problem on the space of 'graphons'.
- Hope-Pray-Prove! The gradient descent process on finite graphs/symmetric matrices converge to a limit as $n \to \infty$.
- Can we show that the limit of GD is a gradient flow on graphons?

General plan and analogies

Let F be graph function. Our goal is to minimize F over large graphs.

Can perform gradient descent on finite graphs/symmetric matrices.

Exploiting the symmetry

- Think of the problem as an optimization problem on the space of 'graphons'.
- Hope-Pray-Prove! The gradient descent process on finite graphs/symmetric matrices converge to a limit as $n \to \infty$.
- Can we show that the limit of GD is a gradient flow on graphons?

Graphons vs Wasserstein space

- Given a graph on n vertices is akin to particle ensemble
- Think of every edge as a *particle* and edge-weights are evolving

Setup and Results

Graphons

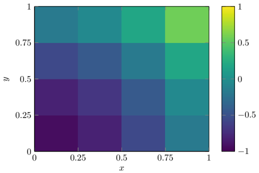
Kernels $\mathcal{W}$

A kernel is a measurable function $W : [0,1]^2 \to [-1,1]$ such that $W(x,y) = W(y,x)$.

- Adjacency matrix $\equiv$ *kernel*.

$$\frac{1}{16}\begin{bmatrix} -16 & -15 & -12 & -7 \\ -15 & -14 & -11 & 1 \\ -12 & -11 & -6 & 4 \\ -7 & 1 & 4 & 9 \end{bmatrix}$$

Symmetric matrix A



Kernel representation of A

Graphons

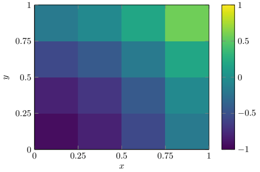
Kernels $\mathcal{W}$

A kernel is a measurable function $W\colon [0,1]^2 \to [-1,1]$ such that $W(x,y) = W(y,x)$.

- Adjacency matrix $\equiv$ *kernel*.

$$\frac{1}{16}\begin{bmatrix} -16 & -15 & -12 & -7 \\ -15 & -14 & -11 & 1 \\ -12 & -11 & -6 & 4 \\ -7 & 1 & 4 & 9 \end{bmatrix}$$

Symmetric matrix A



Kernel representation of A

- Identify adjacency matrix/kernel up to ‘permutations’.
- Identify $W_1 \cong W_2$ if one can be obtained by ‘relabeling’ the vertices of the other, i.e.,

$$W_1(\varphi(x), \varphi(y)) = W_2(x,y), \qquad x, y \in [0,1].$$

Graphons

Graphons $\widehat{\mathcal{W}}$ (Lovász & Szegedy, 2006): $\widehat{\mathcal{W}} := \mathcal{W}/\cong$

Cut metric :: Weak convergence

- Cut metric, $\delta_\square$, metrizes graph convergence.
- $(\widehat{\mathcal{W}}, \delta_\square)$ is **compact**.

[1] Gradient flows on graphons - Oh, Pal, Somani, Tripathi, 2021

[2] Gradient Flows: In Metric Spaces and in the Space of Probability Measures - Ambrosio, Gigli, Savaré, 2008

Graphons

Graphons $\widehat{\mathcal{W}}$ (Lovász & Szegedy, 2006): $\widehat{\mathcal{W}} := \mathcal{W}/\cong$

Cut metric :: Weak convergence

- Cut metric, $\delta_\square$, metrizes graph convergence.
- $(\widehat{\mathcal{W}}, \delta_\square)$ is **compact**.

Invariant L^2 metric δ_2 :: 2-Wasserstein metric $\mathbb{W}_2$

- Stronger than the cut metric (i.e., $\delta_\square$ convergence $\not\Rightarrow$ δ_2 convergence).
- **Gromov-Wasserstein distance** between $([0,1], \mathrm{Leb}, W_1)$ and $([0,1], \mathrm{Leb}, W_2)$.

[1] Gradient flows on graphons - Oh, Pal, Somani, Tripathi, 2021

[2] Gradient Flows: In Metric Spaces and in the Space of Probability Measures - Ambrosio, Gigli, Savaré, 2008

Graphons

Graphons $\widehat{\mathcal{W}}$ (Lovász & Szegedy, 2006): $\widehat{\mathcal{W}} := \mathcal{W}/\cong$

Cut metric :: Weak convergence

- Cut metric, $\delta_\square$, metrizes graph convergence.
- $(\widehat{\mathcal{W}}, \delta_\square)$ is **compact**.

Invariant L^2 metric δ_2 :: 2-Wasserstein metric $\mathbb{W}_2$

- Stronger than the cut metric (i.e., $\delta_\square$ convergence $\not\Rightarrow$ δ_2 convergence).
- **Gromov-Wasserstein distance** between $([0,1], \text{Leb}, W_1)$ and $([0,1], \text{Leb}, W_2)$.

We show[1]

- The metric δ_2 is **geodesic** (just like $\mathbb{W}_2$). Geodesic convexity on $(\widehat{\mathcal{W}}, \delta_2)$.
- Notion of 'gradient' on $(\widehat{\mathcal{W}}, \delta_2)$ called 'Frechét-like derivative'!
- Construction of 'gradient flows' on $(\widehat{\mathcal{W}}, \delta_2)$[2].

[1] Gradient flows on graphons - Oh, Pal, Somani, Tripathi, 2021

[2] Gradient Flows: In Metric Spaces and in the Space of Probability Measures - Ambrosio, Gigli, Savaré, 2008

Existence of gradient flow on Graphons

Theorem [OPST '21]

If $R\colon \widehat{\mathcal{W}} \to \mathbb{R}$

- has a Fréchet-like derivative,
- is geodesically semiconvex in δ_2,

then starting from any $W_0 \in \widehat{\mathcal{W}}$, $\exists!$ gradient flow curve $(W_t)_{t\in\mathbb{R}_+}$ for R

Existence of gradient flow on Graphons

Theorem [OPST '21]

If $R \colon \widehat{\mathcal{W}} \to \mathbb{R}$

- has a Fréchet-like derivative,
- is geodesically semiconvex in δ_2,

then starting from any $W_0 \in \widehat{\mathcal{W}}$, $\exists!$ gradient flow curve $(W_t)_{t\in\mathbb{R}_+}$ for R satisfying

$$W_t := W_0 - \int_0^t DR(W_s)\,\mathrm{d}s, \qquad t \in \mathbb{R}_+,$$

inside $\widehat{\mathcal{W}}$. At the boundary $\{-1, 1\}$ of $\widehat{\mathcal{W}}$, add constraints to contain it.

Scaling limits of GD [OPST '21 + HOPST '22]

Euclidean GD/SGD of R_n over $n \times n$ symmetric matrices, converges to the 'gradient flow' of R on the metric space of graphons.

Scaling limit of Noisy SGD

For $n \in \mathbb{N}$, let $\quad R_n(A) = \mathbb{E}_\xi[\ell_n(A;\xi)] \quad$ for $A \in \mathcal{M}_n$.

SGD

Given the k-th iterate $W_k^{(n)} \in \mathcal{M}_n$, sample ξ,

$$W_{k+1}^{(n)} = W_k^{(n)} \quad - \quad \tau_n \cdot n^2 \underbrace{\nabla \ell_n(W_k^{(n)};\xi)}_{\text{stochastic Euclidean gradient}}$$

Scaling limit of Noisy SGD

For $n \in \mathbb{N}$, let $R_n(A) = \mathbb{E}_\xi[\ell_n(A;\xi)]$ for $A \in \mathcal{M}_n$.

Noisy SGD

Given the k-th iterate $W_k^{(n)} \in \mathcal{M}_n$, sample ξ,

$$W_{k+1}^{(n)} = W_k^{(n)} - \tau_n \cdot n^2 \underbrace{\nabla \ell_n(W_k^{(n)};\xi)}_{\text{stochastic Euclidean gradient}} + \tau_n^{1/2} \cdot \underbrace{N(0,\mathrm{id})}_{\text{added noise}}$$

Scaling limit of Noisy SGD

For $n \in \mathbb{N}$, let $R_n(A) = \mathbb{E}_\xi[\ell_n(A;\xi)]$ for $A \in \mathcal{M}_n$.

Noisy SGD

Given the k-th iterate $W_k^{(n)} \in \mathcal{M}_n$, sample ξ,

$$W_{k+1}^{(n)} = P\left(W_k^{(n)} - \tau_n \cdot n^2 \underbrace{\nabla \ell_n(W_k^{(n)};\xi)}_{\text{stochastic Euclidean gradient}} + \tau_n^{1/2} \cdot \underbrace{N(0,\mathrm{id})}_{\text{added noise}}\right)$$

Scaling limit of Noisy SGD

For $n \in \mathbb{N}$, let $\quad R_n(A) = \mathbb{E}_\xi[\ell_n(A;\xi)] \quad$ for $A \in \mathcal{M}_n$.

Noisy SGD

Given the k-th iterate $W_k^{(n)} \in \mathcal{M}_n$, sample ξ,

$$W_{k+1}^{(n)} = P\left(W_k^{(n)} \quad - \quad \tau_n \cdot n^2 \underbrace{\nabla \ell_n(W_k^{(n)};\xi)}_{\text{stochastic Euclidean gradient}} \quad + \quad \tau_n^{1/2} \cdot \underbrace{N(0,\mathrm{id})}_{\text{added noise}} \right)$$

If $W_0^{(n)} \xrightarrow{\delta_2} W_0$, and $\tau_n \to 0$, as $n \to \infty$, then a.s.

$$W^{(n)} \overset{\delta_\square}{\rightrightarrows} \Gamma, \qquad \text{as } n \to \infty,$$

where $\Gamma \colon t \mapsto \Gamma(t)$ is the curve described by the McKean-Vlasov equation.

McKean-Vlasov equation

- Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with a Brownian Motion $B(t)$, and $(U, V) \overset{\text{i.i.d.}}{\sim} \text{Uni}[0, 1]$.
- Consider the process $(X(t), \Gamma(t))$ such that

Existence + uniqueness when DR is L^2 Lipschitz - [HOPST '22]

McKean-Vlasov equation

- Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with a Brownian Motion $B(t)$, and $(U, V) \overset{\text{i.i.d.}}{\sim} \text{Uni}[0,1]$.
- Consider the process $(X(t), \Gamma(t))$ such that on $\{U = u, V = v\}$,

$$\mathrm{d}X(t) = -(DR)(\Gamma(t))(u,v)\,\mathrm{d}t + \mathrm{d}B(t) \quad \underbrace{+\,\mathrm{d}L^-(t) - \mathrm{d}L^+(t)}_{\text{constrain in } [-1,1]}, \qquad \text{(McKean-Vlasov)}$$

$$\Gamma(t)(x,y) = \mathbb{E}[X(t) \mid (U,V) = (x,y)], \quad \forall\, (x,y) \in [0,1]^2.$$

Existence + uniqueness when DR is L^2 Lipschitz - [HOPST '22]

McKean-Vlasov equation

- Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with a Brownian Motion $B(t)$, and $(U, V) \overset{\text{i.i.d.}}{\sim} \text{Uni}[0,1]$.
- Consider the process $(X(t), \Gamma(t))$ such that on $\{U = u, V = v\}$,

$$\mathrm{d}X(t) = -(DR)(\Gamma(t))(u,v)\,\mathrm{d}t + \mathrm{d}B(t) \quad \underbrace{+\,\mathrm{d}L^{-}(t) - \mathrm{d}L^{+}(t)}_{\text{constrain in } [-1,1]}, \qquad \text{(McKean-Vlasov)}$$

$$\Gamma(t)(x,y) = \mathbb{E}[X(t) \mid (U,V) = (x,y)], \quad \forall\,(x,y) \in [0,1]^2.$$

Expected to arise as limit of large number of graph dynamics:

- "Mean-field interaction": For any edge-weight, the effect of all others edge-weights on its evolution is invariant under vertex relabeling.
- "Propagation of chaos": Every edge-weight between a set of m randomly chosen vertices evolves independently in the limit.

Existence + uniqueness when DR is L^2 Lipschitz - [HOPST '22]

Future directions

- Stronger but natural topology? Measure-valued graphons? In progress.
- Extension to **Deep NNs**. Use a graphon for each layer (bipartite graph), respecting all joint layerwise permutation symmetries - In progress.

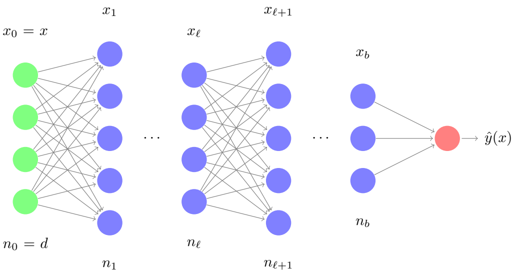
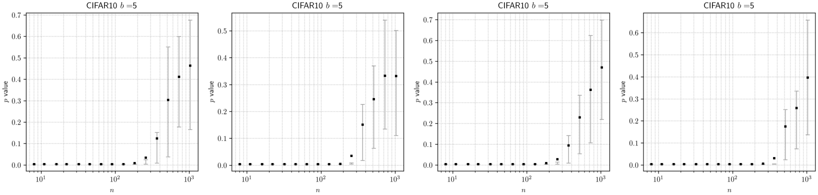


Figure: A b-layer NN.

- How does data distribution propagate across depth? Control theory, optimal transport - Open.

Propagation of Chaos experiments

- SGD training of a 5 layer deep feedforward ReLU networks. $\sigma : x \mapsto \max\{0, x\}$.
- Test joint independence of elements in random 2×2 submatrices.
- Null hypothesis: All the 4 random variables are jointly independent.



(a) Dataset: CIFAR10. x-axis: n, y-axis: p-value with interquartile range.

- For small n ($\lesssim 300$): The p value is $< 0.05 \implies$ reject null hypothesis.
- Monotonic increase in p value as n increases, in all layers.

Thank you!

Thank you!

ArXiv version[3]: https://arxiv.org/abs/2210.00422

[3] Stochastic optimization on matrices and a graphon McKean-Vlasov limit - Harchaoui, Oh, Pal, Somani, Tripathi, 2022