Scaling limits of SGD over large networks: a Graphon perspective

Zaid Harchaoui Sewoong Oh Soumik Pal Raghav Somani Raghav Tripathi

University of Washington

Motivation

- Study large scale optimization problems over **unlabeled graphs**.

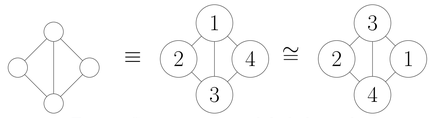


Figure: Symmetry in unlabeled graphs.

- Minimize Risk function over weights of the Neural Network (NN).

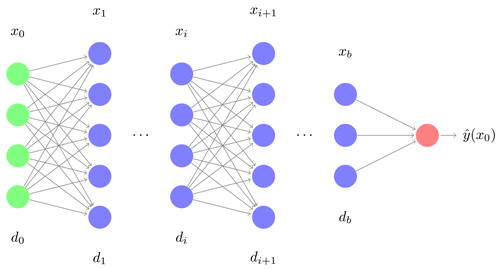


Figure: A feedforward NN as a *computational graph*.

- Do symmetries help our understanding and analysis? Unlabeled graphs, and NNs have **permutation symmetries**.
- Stochastic Gradient Descent (SGD) behave as network grows? Does noise play any role?

Objective

Understand the scaling limits of the standard first-order stochastic optimization algorithms over functions of large dense unlabeled weighted graphs, which are invariant under vertex relabeling.

For 1-hidden layer NNs (evolving neurons = interacting particle)

- System of interacting particles (neurons in a 1-hidden layer NN) have single permutation symmetry.
- Scaling limit - "Mean-field limits" [4], "Wasserstein gradient flow" [1].

For multi-layer NNs (evolving weights = interacting weighted edges)

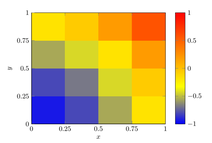
- We attempt to generalize the Wasserstein calculus to higher-order exchangeable structures.

Graphons $\widehat{\mathcal{W}}$

- Space of *graphons* capture this symmetry. Adjacency matrix ≡ 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

- **Kernels**, $\mathcal{W}$: Measurable symmetric function $W\colon [0,1]^{(2)} \to [-1,1]$.
- **Graph isomorphism**: Identify $W_1 \cong W_2$ if one can be obtained by 'relabeling' the vertices of the other.
- **Graphons**: $\widehat{\mathcal{W}} := \mathcal{W}/\cong$.

Topology, metric & differentiable structure over graphons

- **Cut Topology**: Plays a similar topological role as the topology of weak convergence does on probability measures. It captures graph convergence, is *compact* and metrizable by $\delta_\square$.
- **Invariant L^2 metric,** δ_2: Sometimes referred to as the "Gromov-Wasserstein" metric. Plays a similar role as the Wasserstein-2 metric does on probability measures. It is a *geodesic* metric [3].
- **Fréchet-like derivative** [3]: For $R\colon \widehat{\mathcal{W}} \to \mathbb{R}$, the Fréchet-like derivative $DR(W)$ of R is the *first order linear approximation* of R at $W \in \mathcal{W} \subseteq L^2([0,1]^2)$.

Any $n \times n$ symmetric matrix A naturally corresponds to a kernel and hence can be naturally associated with a graphon. Thus, any function $R\colon \widehat{\mathcal{W}} \to \mathbb{R}$ defines a map on bounded $n \times n$ symmetric matrices $\mathcal{M}_n$, denoted by R_n. Spatial scaling leads to the relation: $n^2 \nabla R_n \equiv DR$.

Scaling limit of SGD

Existence of a gradient flow on graphons [3]

Let $R\colon \widehat{\mathcal{W}} \to \mathbb{R}$ be $\delta_\square$-continuous, Fréchet differentiable, and geodesically semiconvex. Then staring from $W_0 \in \widehat{\mathcal{W}}$, there exists a unique gradient flow curve $(W_t)_{t\geq 0}$ of R satisfying

$$W_t = W_0 - \int_0^t DR(W_s)\,\mathrm{d}s, \qquad t \geq 0,$$

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

- For every $n \in \mathbb{N}$, start at $W_0^{(n)} \in \mathcal{M}_n$, take steps towards the negative of the scaled Euclidean gradient $n^2\nabla R_n$, to obtain

$$W_{k+1}^{(n)} = P\Big(W_k^{(n)} - \tau_n n^2 \nabla R_n\Big(W_k^{(n)}\Big)\Big), \qquad k \in \mathbb{Z}_+. \qquad \text{(PGD)}$$

Convergence of Gradient Descents [3]

Let the Fréchet-like derivative be uniformly bounded, i.e., $\|DR(W)\|_\infty < M < \infty$, $\forall\, W \in \widehat{\mathcal{W}}$. If $W_0^{(n)} \xrightarrow{\delta_\square} W_0$, and $\tau_n \to 0$, then as curves, $W^{(n)} \overset{\delta_\square}{\rightrightarrows} W$, as $n \to \infty$.

- Stochastic approximation algorithms like SGD also converges to the *same* gradient flow on graphons.
- The stochastic noise smoothens out due to the regularity of the cut topology.

Examples of functions

1. **Scalar Entropy function**: Let $h\colon p \mapsto p\log p + (1-p)\log(1-p)$, and $\epsilon > 0$. Sample $\{Z_i\}_{i=1}^2 \overset{\text{i.i.d.}}{\sim} \mathrm{Uni}[0,1]$, and define

$$\mathcal{E}(W) := \mathbb{E}[h(W(Z_1,Z_2))]\,, \qquad \text{for } \epsilon \leq W \leq 1-\epsilon.$$

2. **Homomorphism functions**: Let $F = (V,E)$ be a simple graph with k vertices. Sample $\{Z_i\}_{i=1}^k \overset{\text{i.i.d.}}{\sim} \mathrm{Uni}[0,1]$, and define

$$H_F(W) := \mathbb{E}\left[\prod_{\{i,j\}\in E} W(Z_i,Z_j)\right], \qquad \text{for } W \in \widehat{\mathcal{W}}.$$

Scaling limit of Noisy SGD & Graphon McKean-Vlasov eqns.

For every $n \in \mathbb{N}$, start at $W_0^{(n)} \in \mathcal{M}_n$. Take step towards negative of the stochastic gradient. Add scaled variance bounded noise. Project every coordinate on $[-1,1]$. Define

$$W_{k+1}^{(n)} = P\Big(W_k^{(n)} - \tau_n \cdot n^2 g_{n,k+1} + \tau_n^{1/2}\cdot G_{n,k}\Big), \qquad k \in \mathbb{Z}_+, \qquad \text{(PNSGD)}$$

where

$$\mathbb{E}\left[g_{n,k+1} \,\middle|\, W_k^{(n)}\right] = \nabla R_n(W_{n,k}),$$

$$\mathbb{E}\left[\frac{1}{n^2}\left\|n^2 g_{n,k+1} - n^2\nabla R_n\Big(W_k^{(n)}\Big)\right\|_{\mathrm{F}}^2 \,\middle|\, W_k^{(n)}\right] \leq \sigma^2 < \infty,$$

$$\mathbb{E}[G_{n,k}] = 0, \text{ and } \mathbb{E}[G_{n,k}(i,j)^2] < M^2 < \infty \;\forall\, (i,j) \in [n]^{(2)}.$$

Convergence of Noisy Stochastic Gradient Descents [2]

Let the Fréchet-like derivative be uniformly bounded, i.e., $\|DR(W)\|_\infty < M < \infty$, $\forall\, W \in \widehat{\mathcal{W}}$. If $W_0^{(n)} \xrightarrow{\delta_2} W_0$, and $\tau_n \to 0$, then as curves, $W^{(n)} \overset{\delta_\square}{\rightrightarrows} \Gamma$, a.s. as $n \to \infty$.

- Given a probability space with Brownian Motion B, & $(U,V) \overset{\text{i.i.d.}}{\sim} \mathrm{Uni}[0,1]$.
- $(X(t), \Gamma(t))$ solves the McKean-Vlasov type SDE. On $\{U = u, V = v\}$,

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

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

- **Mean-field interaction**: For any edge-weight, the effect of all others edge-weights on its evolution is invariant under vertex relabeling.
- Γ is deterministic and absolutely continuous, but is not the gradient flow of R on graphons.

Simulations

Turán's theorem (extremal graph theory): The n-vertex triangle-free graph with the maximum number of edges is a complete bipartite graph.

Q. Can we recover this theorem through an optimization problem on graphons?

A. Say we minimize $H_\triangle - H_-$ (triangle density minus edge density).

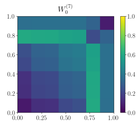


Figure: GD ($n = 7$)

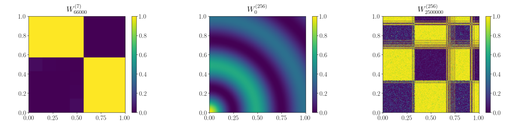


Figure: Noisy SGD ($n = 256$)

- Both the approximate minimizers represent the balanced bipartite graph!

References

[1] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3040–3050, Red Hook, NY, USA, 2018. Curran Associates Inc.

[2] Zaid Harchaoui, Sewoong Oh, Soumik Pal, Raghav Somani, and Raghavendra Tripathi. Stochastic optimization on matrices and a graphon McKean-Vlasov limit. arXiv preprint arXiv:2210.00422, 2022.

[3] Sewoong Oh, Soumik Pal, Raghav Somani, and Raghav Tripathi. Gradient flows on graphons: existence, convergence, continuity equations. arXiv preprint arXiv:2111.09459, 2021.

[4] Mei Song, Andrea Montanari, and P Nguyen. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences*, 115:E7665–E7671, 2018.