Gradient flows on graphons: existence, convergence, continuity equations

Sewoong Oh Soumik Pal Raghav Somani Raghav Tripathi

University of Washington

NEURAL INFORMATION PROCESSING SYSTEMS

We study the gradient flow of functions on large graphs that are invariant under vertex relabeling. Suppose $W \in \mathbb{R}^{[n]^2}$ is the adjacency matrix of a labeled graph (V, E). Let f_n be a permutation-invariant function, that is,

$$f_n(W_{\pi_i \pi_j},\ i, j \in [n]) = f_n(W_{ij},\ i, j \in [n]) \ , \quad \forall \text{ permutations } \pi \text{ of } [n] \ .$$

We are interested in the behavior of the gradient flow Cauchy problem

$$\dot{W}_{ij}(t) = -\nabla_{ij} f_n(W_{ij},\ i, j \in [n]) \ , \quad i, j \in [n] \ . \tag{1}$$

Question

Continuum limits of large graphs is called *graphons*. Is there a scaling limit of the Cauchy problem (1) as a curve on the space of graphons?

Motivation

When f_n is a permutation invariant function of 1-d arrays, the scaling limit for problem (1) is called the *Wasserstein gradient flows* [1]. This has been successfully used in the context of single hidden layer Neural Networks (NN) [2, 3], where in problem (1), exchangeable particles are weight vectors corresponding to neurons in the hidden layer. For NNs with multiple hidden layers, one can consider the edge weights of the network as exchangeable particles and ask the above question in this context.

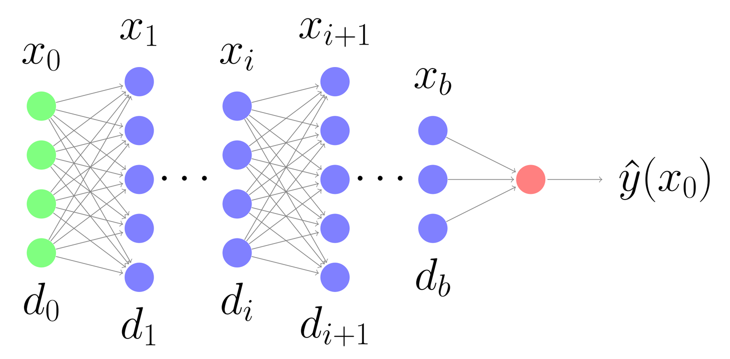


Figure 1: Finite width Neural Network with multiple hidden layers.

Introduction

Kernels and Graphons

A *kernel* is a Borel measurable function $W\colon [0,1]^2 \to [-1, 1]$ that is symmetric, i.e., $W(x, y) = W(y, x)$. The set of all kernels is denoted by $\mathcal{W}$. Kernels identified up to measure preserving transformation are called *graphons*. The set of all graphons is denoted by $\widehat{\mathcal{W}}$. The equivalence class of $W \in \mathcal{W}$ is denoted by $[W] \in \widehat{\mathcal{W}}$.

Metrics over the space of Graphons

There are two natural metrics on $\widehat{\mathcal{W}}$:

1. **Invariant L^2 metric, δ_2**: Also sometimes referred to as the "Gromov-Wasserstein" metric, the invariant L^2 metric δ_2 plays a similar role as the Wasserstein-2 metric does on probability measures.
2. **Cut metric, $\delta_\square$**: The cut metric plays a similar topological role as the metric of weak convergence does on probability measures.

Functions on Graphons

Any function $F\colon \widehat{\mathcal{W}} \to \mathbb{R}$ naturally extends to an *invariant function* $f\colon \mathcal{W} \to \mathbb{R}$ by setting $f(W) = F([W])$ for $W \in \mathcal{W}$. Any symmetric matrix A naturally corresponds to a kernel and hence can be naturally associated with a graphon. Thus, F defines a map on $k \times k$ symmetric matrices, denoted by f_k.

Block Graphons

For $k \in \mathbb{N}$, let $\mathcal{W}_k \subset \mathcal{W}$ be the set of kernels that are constant a.e. over rectangles $[i/k, (i+1)/k] \times [j/k, (j+1)/k]$, for $i, j \in [k-1]$. Let the corresponding quotient set be $\widehat{\mathcal{W}}_k$. The gradient flows (a.k.a curves of maximal slope) on $\widehat{\mathcal{W}}_k$ can be obtained from a suitable scaling of Euclidean gradient flow on the space of $k \times k$ symmetric matrices.

Fréchet-like derivative and λ-semiconvexity

Let $f\colon \mathcal{W} \to \mathbb{R}$ be the *invariant extension* of $F\colon \widehat{\mathcal{W}} \to \mathbb{R}$. The Fréchet-like derivative of f at $V \in \mathcal{W}$ is given by $\phi \in L^\infty([0,1]^2)$ if

$$\lim_{\substack{W \in \mathcal{W},\\ \|W - V\|_2 \to 0}} \frac{f(W) - f(V) - \langle \phi, W - V\rangle}{\|W - V\|_2} = 0 \ .$$

F is λ-semiconvex w.r.t. δ_2 for $\lambda \in \mathbb{R}$ if there is a geodesic $([W_t])_{t \in [0,1]} \subset \widehat{\mathcal{W}}$ such that

$$F([W_t]) \leq (1-t)F([W_0]) + tF([W_1]) - \frac{\lambda}{2} t(1-t)\delta_2^2([W_0], [W_1]) \ .$$

Result

We obtain sufficient conditions under which the scaling limit of the Euclidean gradient flow of f_n exists and converges as $n \to \infty$ to a curve of maximal slope on $(\widehat{\mathcal{W}}, \delta_2)$. This convergence is w.r.t. $\delta_\square$. We show that $(\widehat{\mathcal{W}}, \delta_2)$ is a geodesic metric space.

Existence and Convergence of Gradient flows

Let F be continuous w.r.t. $\delta_\square$, and λ-semiconvex w.r.t. $(\widehat{\mathcal{W}}, \delta_2)$ for some $\lambda \in \mathbb{R}$. For every $k \in \mathbb{N}$, consider the gradient flow $\omega^{(k)}\colon [0,1] \to \widehat{\mathcal{W}}_k$ of F starting at $\omega^{(k)}(0) = [U_{k,0}] \in \widehat{\mathcal{W}}_k$. If $([U_{k,0}])_{k \in \mathbb{N}} \xrightarrow{\delta_\square} [U_0] \in \widehat{\mathcal{W}}$, and $\limsup_{k \to \infty} \|\phi(U_{k,0})\|_2 < \infty$ and $\|\phi(U_0)\|_2 < \infty$, then

$$\limsup_{k \to \infty} \sup_{t \in [0,1]} \delta_\square\left(\omega^{(k)}(t), \omega(t)\right) = 0 \ ,$$

where ω is the unique curve of maximal slope of F on $\widehat{\mathcal{W}}$ starting at $\omega(0) = [U_0]$.

Examples

1. **Scalar Entropy function**: Let $h\colon p \mapsto p \log p + (1-p)\log(1-p)$, and $\epsilon > 0$.

$$\mathcal{E}([W]) := \mathbb{E}[h(W(Z_i, Z_j))] \ , \quad \{Z_i\}_{i=1}^2 \overset{\text{i.i.d.}}{\sim} \text{Uni}[0,1] \ , \ \text{for } \epsilon \leq W \leq 1 - \epsilon \ .$$

2. **Homomorphism functions**: Let $F = (V, E)$ be a simple graph with k vertices.

$$H_F([W]) := \mathbb{E}\left[\prod_{\{i,j\} \in E} W(Z_i, Z_j)\right] \ , \quad \{Z_i\}_{i=1}^k \overset{\text{i.i.d.}}{\sim} \text{Uni}[0,1] \ , \ \text{for } W \in \mathcal{W} \ .$$

We simulate gradient flows on $\mathcal{W}_{24}$ over the function $H_F + \alpha\mathcal{E}$ for $\alpha \in \{0, 2^{-3}\}$, where F is the 5-cycle.

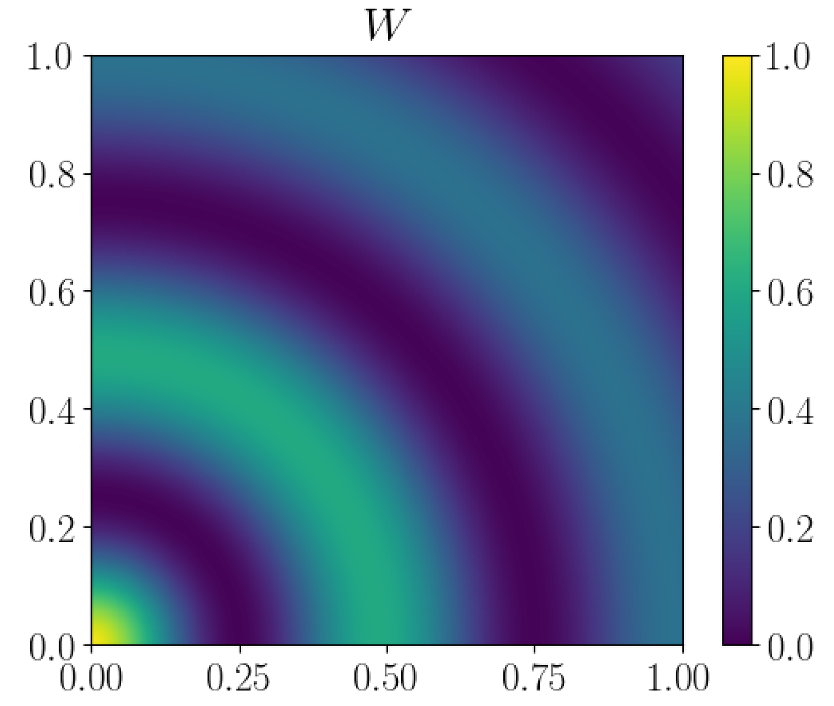
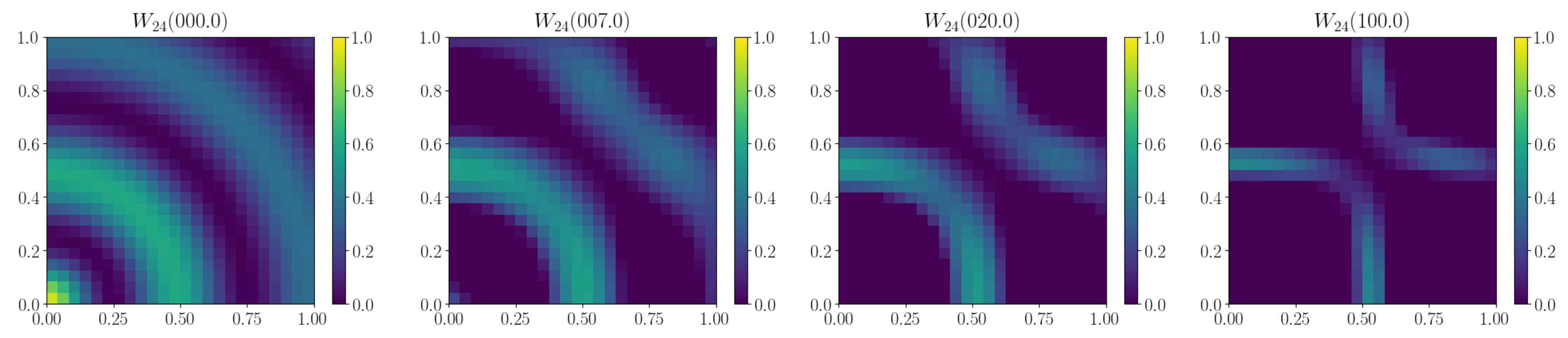
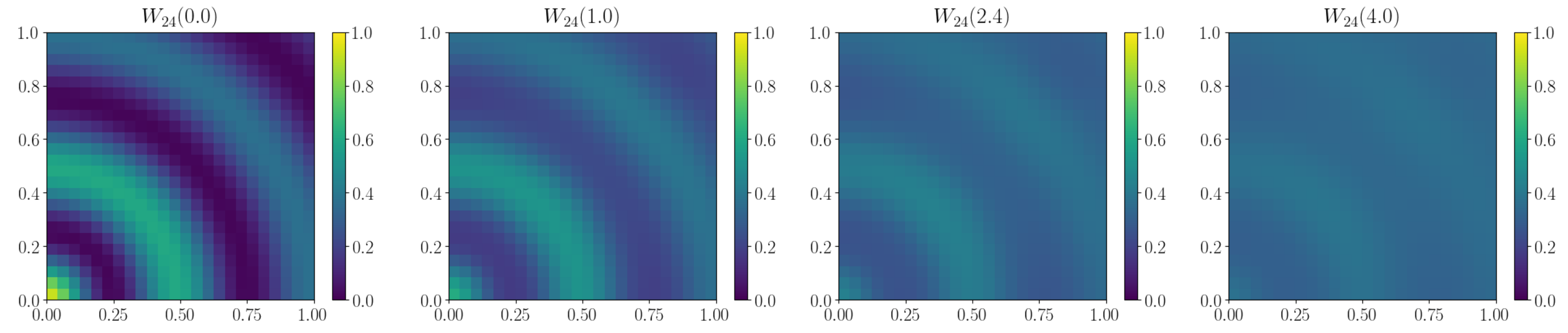


Figure 2: $W \in \mathcal{W}$



(a) $W_{24}(0)$ (b) $W_{24}(7)$ (c) $W_{24}(20)$ (d) $W_{24}(100)$

Figure 3: Euclidean Gradient descent on $\mathcal{W}_{24}$ for $\alpha = 0$.



(a) $W_{24}(0)$ (b) $W_{24}(1)$ (c) $W_{24}(2.4)$ (d) $W_{24}(4)$

Figure 4: Euclidean Gradient descent on $\mathcal{W}_{24}$ for $\alpha = 2^{-3}$.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures. Second Edition.* Lectures in Mathematics. ETH Zürich. Birkhäuser Verlag AG, 2008.

[2] 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.

[3] 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.