Gradient flows on Graphons

Sewoong Oh[1], Soumik Pal[2], Raghav Somani[1] and Raghav Tripathi[2]

[1]UW CSE & [2]UW Math

April 1, 2022




Objective

Study large scale optimization problems that have permutation symmetries.

- Exploiting symmetries allow taking limits of the size of optimization problems.

For $n \in \mathbb{N}$, consider minimizing the following interaction energy $V_n \colon \mathbb{R}^n \to \mathbb{R}_+$

$$V_n(x) \coloneqq \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{1}{2} (x_i - x_j)^2 \, .$$

- Starting from $\{X_{i,0}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \rho_0$, one can perform a gradient flow:

$$\mathrm{d}X_{i,t} = -\frac{1}{n} \sum_{j=1}^{n} (X_{i,t} - X_{j,t}) \, \mathrm{d}t \, , \qquad \forall \, i \in [n], \; t \geq 0 \, .$$

- Notice that V_n is essentially a function of the empirical measure of its inputs!

$$V_n(x) = \mathrm{Var}(\mathrm{Emp}_n(x)) \, .$$

Can we approximate this problem by lifting it over the space of probability measures?

Particle System to Measures

- If a function $V_n \colon \mathbb{R}^n \to \mathbb{R}$ is invariant under permutations of its input, then it can be extended to a function on its empirical measure, and perhaps to a function $V \colon \mathcal{P}(\mathbb{R}) \to \mathbb{R}$.

Particle System to Measures

- If a function $V_n \colon \mathbb{R}^n \to \mathbb{R}$ is invariant under permutations of its input, then it can be extended to a function on its empirical measure, and perhaps to a function $V \colon \mathcal{P}(\mathbb{R}) \to \mathbb{R}$.
- For the interaction energy V_n, we know that $V(\rho) = \mathrm{Var}(\rho)$ for $\rho \in \mathcal{P}(\mathbb{R})$.
- Notice that for all $n \in \mathbb{N}$,

$$\min_{\mathbb{R}^n} V_n = \min_{\mathcal{P}(\mathbb{R})} \mathrm{Var} .$$

- One can solve the latter using *Wasserstein gradient flows*!
- One may also add a noise term.

$$\mathrm{d}X_{i,t} = -\frac{1}{n}\sum_{j=1}^{n}(X_{i,t} - X_{j,t}) + \sqrt{2\beta}\,\mathrm{d}B_{i,t}, \quad \forall\, i \in [n],\ t \geq 0,$$

  where B_t is the standard Brownian motion on $\mathbb{R}^n$, and $\beta \geq 0$.
- This SDE captures the Wasserstein gradient flow of $\mathrm{Var} + \beta\mathrm{Ent} \colon \mathcal{P}(\mathbb{R}) \to \mathbb{R}$, the entropy-regularized optimization.

Benefits

Approximations and universal limits.

Optimization on Large Graphs

Q. What about optimization over dense unlabeled (weighted) graphs?

Optimization on Large Graphs

Q. What about optimization over dense unlabeled (weighted) graphs?

Triangle density

Let G be a finite simple graph with n vertices,

$$h_{\triangle}(G) = \frac{\text{Number of triangles in } G}{n^3} .$$

For a graph with adjacency matrix A one can define

$$\text{Number of triangles in } G = \sum_{\phi\colon [3]\to V(G)} \prod_{\{i,j\}\in E(G)} A_{\phi(i),\phi(j)} .$$

The above formula works even when A is a symmetric matrix of real edge weights.

Optimization on Large Graphs

Scalar Entropy

For a graph G with adjacency matrix A, let $h(p) = p \log p + (1-p)\log(1-p)$,

$$E(G) = \frac{1}{n^2} \sum_{i,j=1}^{n} h(A_{i,j}) \,.$$

- Scalar Entropy is 0 for all unweighted graphs.

Optimization on Large Graphs

Scalar Entropy

For a graph G with adjacency matrix A, let $h(p) = p \log p + (1-p)\log(1-p)$,

$$E(G) = \frac{1}{n^2} \sum_{i,j=1}^{n} h(A_{i,j}) \,.$$

- Scalar Entropy is 0 for all unweighted graphs.

A Problem on Statistics of Exponential Random Graphs

Consider minimizing $h_{\triangle} + E$ over the set of all graphs.

See Diaconis and Janson 2008, Chatterjee & Varadhan 2011, Lovász 2012, Lubetzky and Zhao 2015 etc.

Is there a symmetry?

- Notice that unlabeled graphs have a symmetry under vertex relabeling.

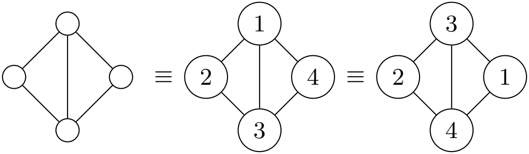


Figure: Symmetry in unlabeled graphs.

- I.e., for an unlabeled graph G with n vertices.
  If A is its adjacency matrix, so is $A_\pi = \left(A_{\pi(i),\pi(j)}\right)_{i,j}$.

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \equiv \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} = A_\pi \ .$$

- This makes these graphs *exchangeable* under this symmetry. See Aldous '81, '82, and Austin '08, '12.

Neural Networks: Another Example

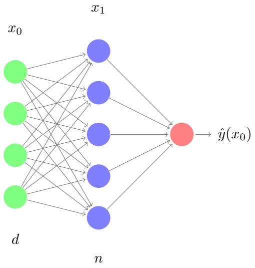


Figure: NN problem is optimization over unlabeled networks.

$$\hat{y}(x_0) = \frac{1}{n}\sum_{i=1}^{d} \sigma(A_{i,j} x_{0,j}) , \quad A \in \mathbb{R}^{n\times d} , \quad R_n(A) := \mathbb{E}_{(X,Y)\sim\mu}[\ell(Y, \hat{y}(X))] .$$

A Mean Field View of the Landscape of Two-Layer Neural Networks - Mei, Montanari & Nguyen, 2018

On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport - Chizat & Bach, 2018

What we need?

- A common embedding that contains all unlabeled graphs
- A suitable topology of ‘graph convergence’
- Completion under a metric
- A notion of ‘differentiable structure’ to define ‘gradient flow’ on this space.

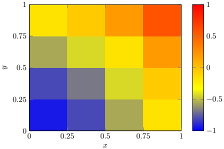
Kernels and Graphons

Kernels $\mathcal{W}$

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

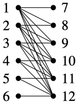
- Symmetric matrices can be converted into a kernel.

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



Symmetric matrix A Kernel representation of A

- (Weighted) Graphs $\Leftrightarrow$ adjacency matrix $\Leftrightarrow$ kernel.



$$\begin{matrix} 0&0&0&0&0&0&1&1&1&1&1&1 \\ 0&0&0&0&0&0&0&1&1&1&1&1 \\ 0&0&0&0&0&0&0&0&1&1&1&1 \\ 0&0&0&0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&0&0&1 \\ 1&0&0&0&0&0&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0&0&0&0&0 \\ 1&1&1&0&0&0&0&0&0&0&0&0 \\ 1&1&1&1&0&0&0&0&0&0&0&0 \\ 1&1&1&1&1&0&0&0&0&0&0&0 \\ 1&1&1&1&1&1&0&0&0&0&0&0 \end{matrix}$$

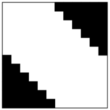


Figure: Example 4.1.6, Graph Theory and Additive Combinatorics, Zhao

Graphons

- Identify two kernels if one can be obtained by 'permuting' the other.
- $W_1 \cong W_2$ if there is a measure preserving transform $\varphi\colon [0,1] \to [0,1]$ such that

$$W_1^{\varphi}(x,y) := W_1(\varphi(x),\varphi(y)) = W_2(x,y)\ .$$

Space of Graphons $\widehat{\mathcal{W}}$ (Lovász & Szegedy, 2006)

$$\widehat{\mathcal{W}} := \mathcal{W}/\cong\ .$$

- For finite labeled graphs, the corresponding graphons are the equivalent classes for identification modulo graph isomorphisms.
- Compare with a measure given by two different pushforwards $T_1, T_2\colon [0,1] \to \mathbb{R}$.

Invariant functions on Kernels = functions on graphons

- Recall the triangle density function

$$h_{\triangle}(G) = \frac{\text{Number of triangles in G}}{n^3} = \frac{1}{n^3} \sum_{\phi:\ [3]\to V(G)} \prod_{\{i,j\}\in E(G)} A_{\phi(i),\phi(j)}.$$

- For a kernel W, the triangle density can be defined as

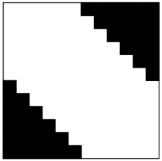
$$h_{\triangle}(W) = \int_{[0,1]^3} W(x_1, x_2)W(x_2, x_3)W(x_3, x_1)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3\ .$$

- $h_{\triangle}$ is a function on the corresponding graphon. That is,

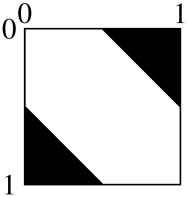
$$h_{\triangle}(V) = h_{\triangle}(W),$$

if V can be obtained from W by vertex permutations.

Convergence of Graph(ons)



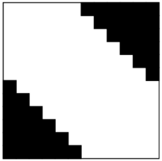
(a) Half Graph (Kernel)



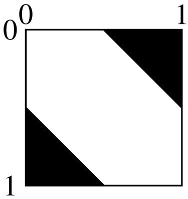
(b) Limit of Half Graph

Graph Theory and Additive Combinatorics, Yufei Zhao

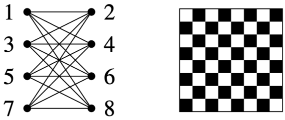
Convergence of Graph(ons)



(a) Half Graph (Kernel)



(b) Limit of Half Graph

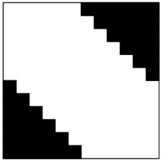


(a) Checkerboard

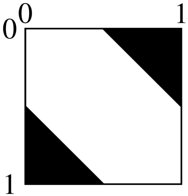
Q. Where does this sequence of graphons converge?

Graph Theory and Additive Combinatorics, Yufei Zhao

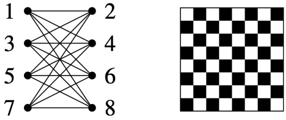
Convergence of Graph(ons)



(a) Half Graph (Kernel)

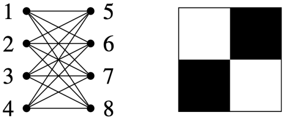


(b) Limit of Half Graph



(a) Checkerboard

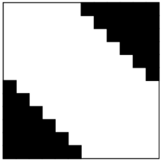
Q. Where does this sequence of graphons converge?



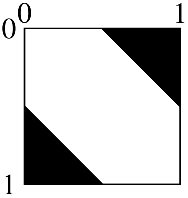
(b) Checkerboard after vertex relabeling

Graph Theory and Additive Combinatorics, Yufei Zhao

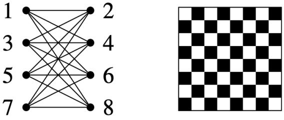
Convergence of Graph(ons)



(a) Half Graph (Kernel)

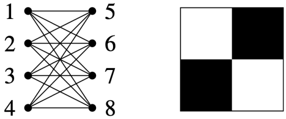


(b) Limit of Half Graph



(a) Checkerboard

Q. Where does this sequence of graphons converge?



(b) Checkerboard after vertex relabeling

A. Both (a) and (b) are the *same* graphon, but two different kernel representations.

Graph Theory and Additive Combinatorics, Yufei Zhao

Metrics on Graphons

- Recall: $W_1 \cong W_2$ if there is a measure preserving transform $\varphi\colon [0,1] \to [0,1]$ such that
$$W_1^{\varphi}(x,y) \coloneqq W_1(\varphi(x), \varphi(y)) = W_2(x,y)\ .$$
- How to define metrics for graphon convergence?

A general recipe

Start with any norm $\|\cdot\|$ on functions $[0,1]^2 \to [-1,1]$. Define δ as
$$\delta(W_1, W_2) = \inf_{\varphi} \|W_1^{\varphi} - W_2\|.$$

Cut Metric: $\delta_\square$

$$\|W\|_\square \coloneqq \sup_{S,T} \left| \int_{S\times T} W(x,y)\,\mathrm{d}x\,\mathrm{d}y \right| .$$

- Cut metric (Frieze & Kannan, 1999) metrizes *graph convergence* (Lovász & Szegedy, 2006).
  - $(G_n)_n$ converges in $\delta_\square$ if
    $$\lim_{n\to\infty} h_F(G_n)$$
    exists for all simple graphs $F \in \{-, \wedge, \triangle, \curlywedge, \sqcup, \square, \boxtimes, \ltimes, \bowtie, \ldots\}$.
- $(\widehat{\mathcal{W}}, \delta_\square)$ is compact.[1]
- Analogous to the weak topology over probabilities.
- Example: Almost surely, random graph $G(n, 1/2)$ converges to constant graphon

$$W(x,y) = 1/2 \ , \ \forall\,(x,y) \in [0,1]^2 \ .$$

[1] uses Szemerédi's regularity lemma

Invariant L^2 metric δ_2

For $\|\cdot\| = \|\cdot\|_{L^2([0,1]^2)}$, we get the Invariant L^2 metric δ_2.

- Stronger than the cut metric (i.e., $\delta_\square \leq \delta_2$).
- **Gromov-Wasserstein distance** between the metric measure spaces $([0,1], \mathrm{Leb}, W_1)$ and $([0,1], \mathrm{Leb}, W_2)$.
- Provides geodesic metric structure on $\widehat{\mathcal{W}}$.
- Allows notion of geodesic convexity.
- Analogous to the Wasserstein-2 metric over measures.

Borgs, Chayes, Lovász, Sós & Vesztergombi, 2008

What is a 'gradient flow' on a metric space?

On $\mathbb{R}^d$

The 'gradient flow' u of a function $F\colon \mathbb{R}^d \to \mathbb{R}$ is given by solutions of

$$u'(t) = -\nabla F(u(t)) \ ,$$

$$\frac{\mathrm{d}}{\mathrm{d}t} F(u(t)) = \langle u'(t), \nabla F(u(t)) \rangle$$

$$\geq -\frac{1}{2}|u'|^2(t) - \frac{1}{2}|\nabla F(u(t))|^2 \ .$$

A curve u is a gradient flow of F if

$\frac{\mathrm{d}}{\mathrm{d}t} F(u(t)) \leq -\frac{1}{2}|u'|^2(t) - \frac{1}{2}|\nabla F(u(t))|^2.$

Gradient Flows in Metric Spaces and in the Space of Probability Measures - Ambrosio, Gigli & Savaré, 2005

What is a 'gradient flow' on a metric space?

On $\mathbb{R}^d$

The 'gradient flow' u of a function $F\colon \mathbb{R}^d \to \mathbb{R}$ is given by solutions of

$$u'(t) = -\nabla F(u(t)) \ ,$$

$$\frac{\mathrm{d}}{\mathrm{d}t} F(u(t)) = \langle u'(t), \nabla F(u(t)) \rangle$$

$$\geq -\frac{1}{2} \left| u' \right|^2 (t) - \frac{1}{2} \left| \nabla F(u(t)) \right|^2 \ .$$

A curve u is a gradient flow of F if

$$\frac{\mathrm{d}}{\mathrm{d}t} F(u(t)) \leq -\tfrac{1}{2} \left| u' \right|^2 (t) - \tfrac{1}{2} \left| \nabla F(u(t)) \right|^2 .$$

On $(\widehat{\mathcal{W}}, \delta_2)$

Consider a curve ω and a function F on $\widehat{\mathcal{W}}$.

- Speed of ω: Metric derivative $|\omega'|$

Metric Derivative of ω

$$\left| \omega' \right| (t) = \lim_{s \to t} \frac{\delta_2(\omega_t, \omega_s)}{|t - s|} \ .$$

- Gradient of F: Fréchet-like derivative

Fréchet-like derivative of F: DF

Provides a local linear approximation of F.

A curve u is a gradient flow of F if

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\omega(t)) \leq -\tfrac{1}{2} \left| \omega' \right|^2 (t) - \tfrac{1}{2} \left| DF(\omega(t)) \right|^2 .$$

Gradient Flows in Metric Spaces and in the Space of Probability Measures - Ambrosio, Gigli & Savaré, 2005

Fréchet-like derivative and existence of gradient flow

Theorem [OPST '21]

If F

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

then starting from any $W_0 \in \widehat{\mathcal{W}}$, there exists a unique gradient flow curve $(W_t)_{t \in \mathbb{R}_+}$ for F.
The curve satisfies ODE

$$W_t := W_0 - \int_0^t DF(W_s)\,\mathrm{d}s\ ,$$

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

Gradient flows on graphons

- For the triangle density function $h_\triangle$,

$$h_\triangle(W) = \int_{[0,1]^3} W(x_1, x_2)W(x_2, x_3)W(x_3, x_1)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3,$$

  its Fréchet-like derivative is

$$(Dh_\triangle)(W)(x, y) = 3\int_0^1 W(x, z)W(z, y)\,\mathrm{d}z\ .$$

- Example of "potential energy". Similarly, one has interaction energy and internal energy.

Example

- For the scalar entropy function

$$E(W) = \int_{[0,1]^2} h(W(x,y))\,\mathrm{d}x\,\mathrm{d}y\ ,\ \ h(p) = p\log(p) + (1-p)\log(1-p),$$

  if $0 < W < 1$, its Fréchet-like derivative is

$$(DE)(W)(x,y) = \log\left(\frac{W(x,y)}{1 - W(x,y)}\right)\ .$$

- Gradient flow

$$W_t'(x,y) = -(DE)(W_t)(x,y)\ ,$$

  converges to the constant $W \equiv 1/2$.

Example

- Given Dh_F and DE, we can now perform a gradient flow to minimize $h_\triangle + E$ on the space of graphons.
- Given initial conditions, one needs to solve for all $x, y \in [0,1]$,

$$W_t'(x,y) = -\left[3\int_0^1 W(x,z)W(z,y)\,\mathrm{d}z + \log\left(\frac{W(x,y)}{1-W(x,y)}\right)\right].$$

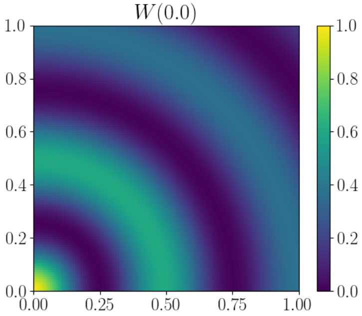


Figure: Gradient flow of $h_\triangle + 10^{-1}E$

Euclidean Gradient flow and Gradient flow on $\widehat{\mathcal{W}}$

Consider a function $F\colon \widehat{\mathcal{W}} \to \mathbb{R}$ that has following gradient flow

$$W(t) = W_0 - \int_0^t DF(W(s))\,\mathrm{d}s\ .$$

- Note that the function F can be regarded as a function on symmetric matrices $F_n\colon \mathcal{M}_n \to R$. Suppose that F_n has a gradient flow. It is then given by

$$V^{(n)}(t) = V_0^{(n)} - \int_0^t \nabla_n F_n\Big(V^{(n)}(s)\Big)\,\mathrm{d}s\ .$$

Question?

Are the curves $V^{(n)}$ and W close (if n is large)?

Euclidean Gradient and Fréchet-like derivative

Fréchet-like derivative [OPST '21]

A symmetric measurable function $\phi \in L^\infty([0,1]^2)$ is said to be Fréchet-like derivative $DF(W)$ of F at $W \in \widehat{\mathcal{W}}$ if

$$\lim_{\substack{U \in \mathcal{W}, \\ \|U-W\|_2 \to 0}} \frac{F(U) - F(W) - \langle \phi, U - W \rangle_{L^2([0,1]^2)}}{\|U - W\|_2} = 0 \, .$$

- Recall that $F \colon \widehat{\mathcal{W}} \to \mathbb{R}$ can be regarded as a function $F_n \colon \mathcal{M}_n \to \mathbb{R}$.
- Let $\nabla_n F_n$ be Euclidean derivative of $F_n \colon \mathcal{M}_n \to \mathbb{R}$.

$\lim_{n\to\infty} n^2 \nabla_n F_n(W) = DF(W)$ as graphons.

Scalings of derivatives

Scaling derivatives for mean

$$F_n\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\nabla F_n = \frac{1}{n}\mathbf{1}$$
$$F(\mu) = \int x \,\mathrm{d}\mu$$
$$\nabla_W F(\mu) \equiv 1$$
$$\lim_{n\to\infty} n\nabla F_n = \nabla_W F(\mu)$$

Scaling derivatives for edge density

$$F_n(A_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_n(i,j)$$
$$\nabla F_n = \frac{1}{n^2}\mathbf{1}$$
$$F(W) = \int_{[0,1]^2} W(x,y)\,\mathrm{d}x\,\mathrm{d}y$$
$$DF(W) \equiv 1$$
$$\lim_{n\to\infty} n^2\nabla F_n = DF$$

Euclidean gradient flow and gradient flow on Graphons

Gradient flow on $\widehat{\mathcal{W}}$

$$\frac{\mathrm{d}}{\mathrm{d}t}W(t) = -DF(W(t)) = -n^2\nabla_n F(W(t))$$

Gradient flow on $\mathcal{M}_n$

$$\frac{\mathrm{d}}{\mathrm{d}t}V(t) = -\nabla_n F(V(t))$$

- The curve $\tilde{W}(t) := V(n^2 t)$ satisfies
$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{W}(t) = -n^2\nabla_n F(\tilde{W}(t)) = -DF(\tilde{W}(t)) \; .$$
- That is, it is reasonable to expect that the gradient flow on Graphons can be obtained by a scaling limit of Euclidean gradient flows.

Convergence of Euclidean Gradient Flow

Theorem [OPST '21]

- Let $F\colon \widehat{\mathcal{W}} \to \mathbb{R}$ be a function with gradient flow $W(t)$, $t \geq 0$.
- Consider the Euclidean gradient flow of $F_n\colon \mathcal{M}_n \to \mathbb{R}$ starting at $V_0^{(n)}$, i.e.,

$$V^{(n)}(t) \coloneqq V_0^{(n)} - \int_0^t \nabla_n F_n\Big(V^{(n)}(s)\Big)\,\mathrm{d}s,$$

  with adjustments at the boundary.
- Set $W^{(n)}(t) = V^{(n)}(n^2 t)$.

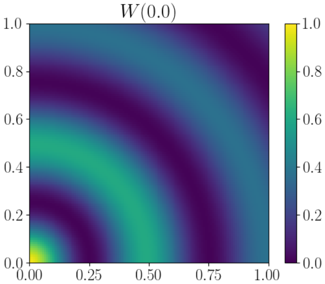
If $W_0^{(n)} \xrightarrow{\delta_\square} W_0$, then

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

uniformly over compact time intervals in $[0, \infty)$.

Simulations

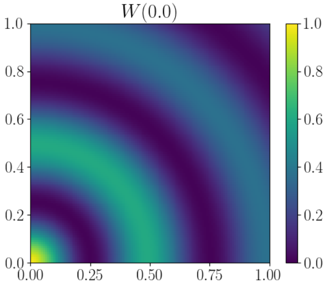
- By Turán's theorem: The n-vertex triangle-free graph with the maximum number of edges is a complete bipartite graph.



Simulations

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

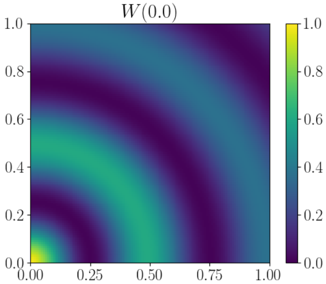
Q. Can one hope to recover this theorem through an optimization problem on graphons?



Simulations

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

Q. Can one hope to recover this theorem through an optimization problem on graphons?

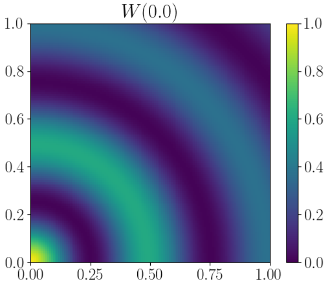


(a) Gradient flow of $10h_{\triangle} - h_{-}$

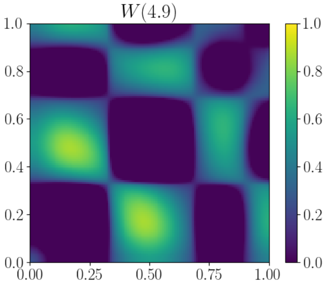
Simulations

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

Q. Can one hope to recover this theorem through an optimization problem on graphons?



(a) Gradient flow of $10h_{\triangle} - h_{-}$



(b) Approximate complete bipartite graphon

Ongoing and Future directions

- Study convergence of stochastic gradient descent with and without added noise.
- Specialize the theory on optimization over multiple layer NNs.
- Limiting curves for other “mean-field interactions” on graphs.

- Optimization on graphs is hard due to discreteness.
- However, gradient flows exist on graphons, their infinite limiting space.
- Analysis is similar to calculus in Wasserstein-2 spaces.
- Approximated by finite dimensional gradient flows on matrices.

Thank you!

- ArXiv version: https://arxiv.org/abs/2111.09459