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

August 23, 2022







Neural Networks (NN) and a permutation symmetry

- Minimize Risk function over weights of the NN.

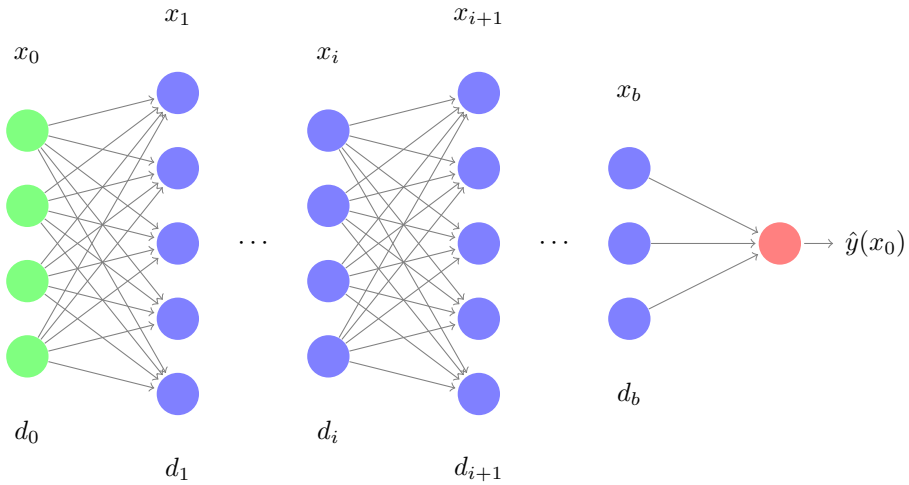


Figure: A feedforward NN as a *computational graph*.

$$x_0 \in \mathbb{R}^{d_0}, \qquad x_i = \sigma(c_i \cdot A_i x_{i-1} + b_i), \quad i \in [b], \qquad \hat{y} = \frac{1}{d_b} \sum_{j=1}^{d_b} x_b(j),$$

Neural Networks (NN) and a permutation symmetry

- Minimize Risk function over weights of the NN.

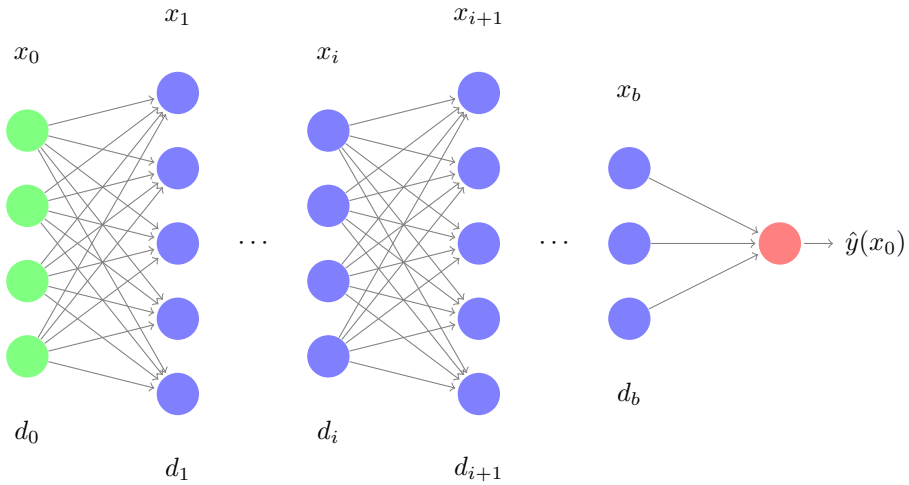


Figure: A feedforward NN as a *computational graph*.

$$x_0 \in \mathbb{R}^{d_0}, \qquad x_i = \sigma(c_i \cdot A_i x_{i-1} + b_i), \quad i \in [b], \qquad \hat{y} = \frac{1}{d_b} \sum_{j=1}^{d_b} x_b(j),$$

- There is a **permutation symmetry** within each layer that keeps the output of the NN unchanged.
- This leads us to treat this problem as an **optimization problem over graphs** invariant under a permutation of vertex labels.

Optimization under permutation invariance

As an example consider a simple function $R_n : \mathbb{R}^{n\times n} \to \mathbb{R}$:

$$R_n(A) := \mathrm{Tr}\left[A^3\right]$$

- Matrix A can be thought of as *adjacency matrix* of an edge-weighted graph.
- This is a spectral function (only depends on eigenvalues of A), counts # of $\triangle$s.
- Symmetry: Does not depend on the labeling of vertices.

Optimization under permutation invariance

As an example consider a simple function $R_n : \mathbb{R}^{n\times n} \to \mathbb{R}$:

$$R_n(A) := \mathrm{Tr}\left[A^3\right]$$

- Matrix A can be thought of as *adjacency matrix* of an edge-weighted graph.
- This is a spectral function (only depends on eigenvalues of A), counts # of $\triangle$s.
- Symmetry: Does not depend on the labeling of vertices.

Optimization of R_n

- Can use *Euclidean* gradient descent (GD) or stochastic gradient descent (SGD).

Optimization under permutation invariance

As an example consider a simple function $R_n : \mathbb{R}^{n\times n} \to \mathbb{R}$:

$$R_n(A) := \mathrm{Tr}\left[A^3\right]$$

- Matrix A can be thought of as *adjacency matrix* of an edge-weighted graph.
- This is a spectral function (only depends on eigenvalues of A), counts # of $\triangle$s.
- Symmetry: Does not depend on the labeling of vertices.

Optimization of R_n

- Can use *Euclidean* gradient descent (GD) or stochastic gradient descent (SGD).
- But optimization should be understood on a space where this symmetry is *quotiented out*. I.e., we recognize all permutations of A as the *same* element.
- This need not enjoy the Euclidean geometry.

We are typically interested in the optimization behavior as n grows.

Optimization under permutation invariance

As an example consider a simple function $R_n : \mathbb{R}^{n\times n} \to \mathbb{R}$:

$$R_n(A) := \mathrm{Tr}\left[A^3\right]$$

- Matrix A can be thought of as *adjacency matrix* of an edge-weighted graph.
- This is a spectral function (only depends on eigenvalues of A), counts # of $\triangle$s.
- Symmetry: Does not depend on the labeling of vertices.

Optimization of R_n

- Can use *Euclidean* gradient descent (GD) or stochastic gradient descent (SGD).
- But optimization should be understood on a space where this symmetry is *quotiented out*. I.e., we recognize all permutations of A as the *same* element.
- This need not enjoy the Euclidean geometry.

We are typically interested in the optimization behavior as n grows.

Question

How do Euclidean GD and SGD behave when we scale the problem size ($n \to \infty$) under the presence of such a symmetry?

Graphons

- There is already a space of **graphons** that exactly captures this symmetry[1].

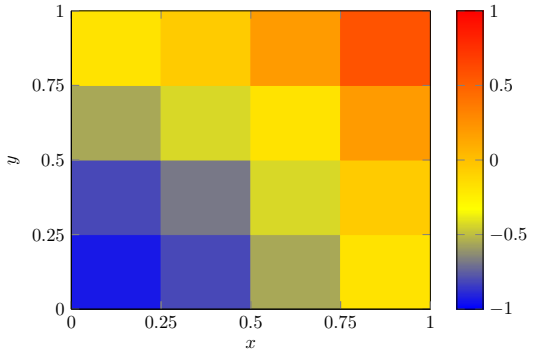
[1] Limits of dense graph sequences - Lovász & Szegedy, 2006

Graphons

- There is already a space of **graphons** that exactly captures this symmetry[1].
- A matrix can be converted to 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

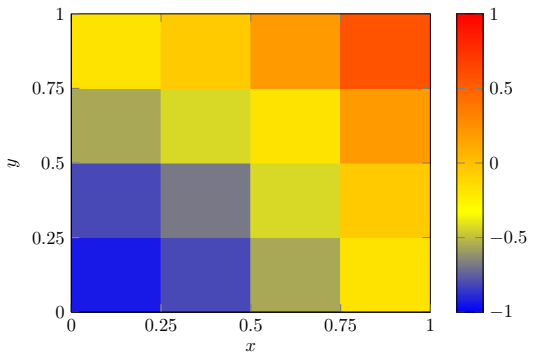
[1] Limits of dense graph sequences - Lovász & Szegedy, 2006

Graphons

- There is already a space of **graphons** that exactly captures this symmetry[1].
- A matrix can be converted to 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

- This allows us to take $n \to \infty$!

Kernels $\mathcal{W}$

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

[1] Limits of dense graph sequences - Lovász & Szegedy, 2006

Graphons

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

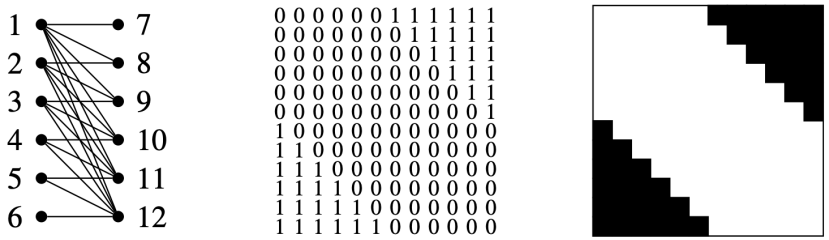


Figure: Example 4.1.6, Graph Theory and Additive Combinatorics, Zhao

Graphons

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

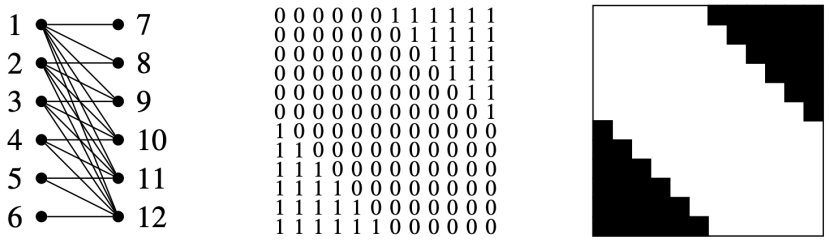


Figure: Example 4.1.6, Graph Theory and Additive Combinatorics, Zhao

- We should recognize a weight matrix/kernel up to 'permutations'.
- Identify $W_1 \cong W_2$ if one can be obtained by relabeling the vertices of the other.

Graphons

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

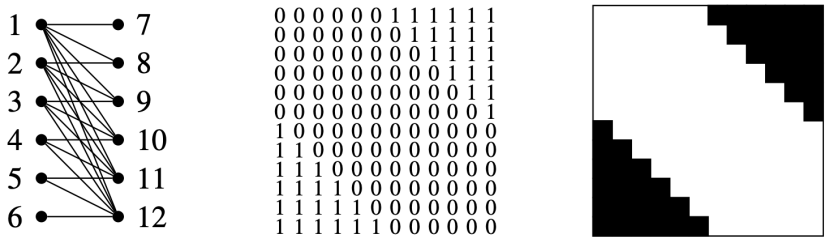


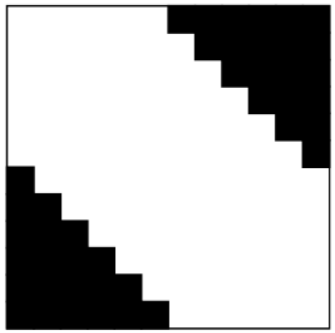
Figure: Example 4.1.6, Graph Theory and Additive Combinatorics, Zhao

- We should recognize a weight matrix/kernel up to 'permutations'.
- Identify $W_1 \cong W_2$ if one can be obtained by relabeling the vertices of the other.

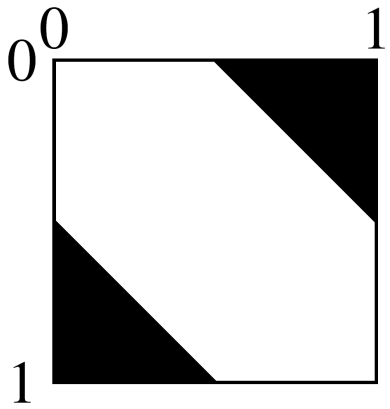
Graphons $\widehat{\mathcal{W}}$ (Lovász & Szegedy, 2006)

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

Convergence of Graph(ons) ($n \to \infty$)

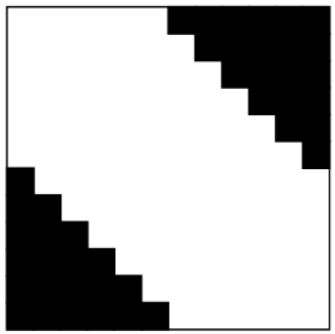


(a) Half Graph (Kernel)

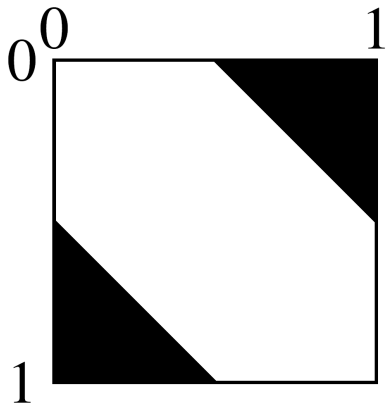


(b) Limit of Half Graph ($n \to \infty$)

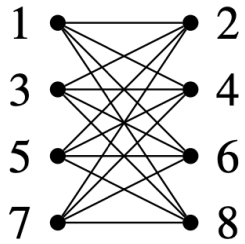
Convergence of Graph(ons) ($n \to \infty$)

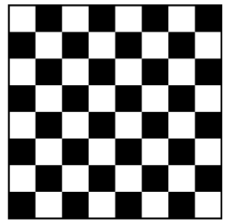


(a) Half Graph (Kernel)



(b) Limit of Half Graph ($n \to \infty$)

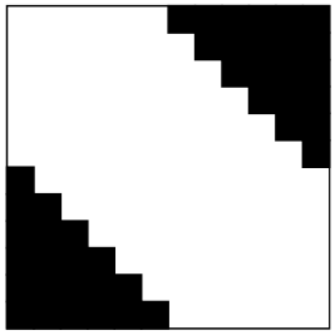




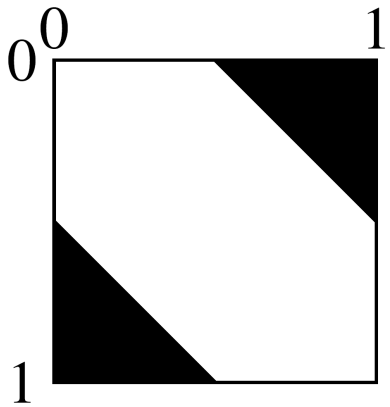
(a) Checkerboard

Q. Where do these graphons converge?

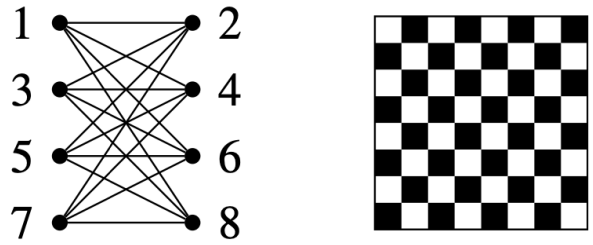
Convergence of Graph(ons) ($n \to \infty$)



(a) Half Graph (Kernel)

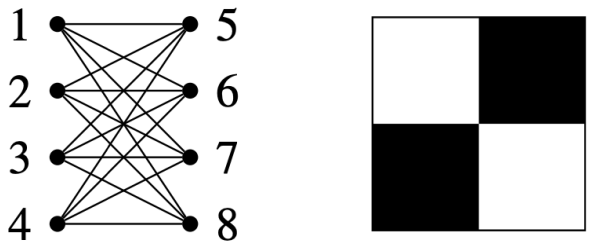


(b) Limit of Half Graph ($n \to \infty$)



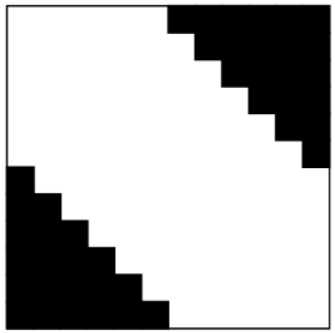
(a) Checkerboard

Q. Where do these graphons converge?

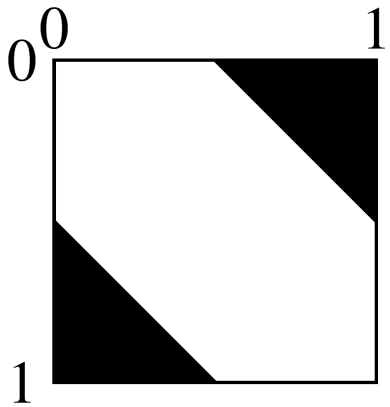


(b) Checkerboard after vertex relabeling

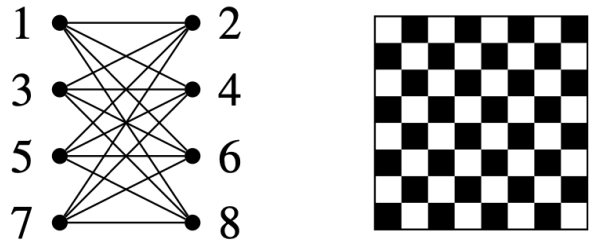
Convergence of Graph(ons) ($n \to \infty$)



(a) Half Graph (Kernel)

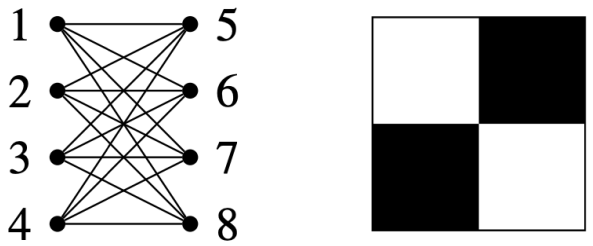


(b) Limit of Half Graph ($n \to \infty$)



(a) Checkerboard

Q. Where do these graphons converge?



(b) Checkerboard after vertex relabeling

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

Topology and Metric over Graphons

The set of Graphons comes with nice topological and geometrical properties.

- Topology induced by **Cut metric**:[2]
  - Provides a natural topology over graphons that **captures graph convergence**.
  - Provides **compactness** when edge-weights are bounded, say in $[-1, 1]$.

[2] Convergent sequences of dense graphs I - Borgs, Chayes, Lovász, Sós, Vesztergombi, 2008.
[3] Graphons and cut metric on sigma-finite measure spaces - Janson, 2016
[4] Gradient flows on graphons - Oh, Pal, Somani, Tripathi, 2021

Topology and Metric over Graphons

The set of Graphons comes with nice topological and geometrical properties.

- Topology induced by **Cut metric**:[2]
  - Provides a natural topology over graphons that **captures graph convergence**.
  - Provides **compactness** when edge-weights are bounded, say in $[-1, 1]$.

- **Invariant L^2 metric δ_2**:[3]
  - We show that it is a **geodesic** metric.[4]

[2] Convergent sequences of dense graphs I - Borgs, Chayes, Lovász, Sós, Vesztergombi, 2008.
[3] Graphons and cut metric on sigma-finite measure spaces - Janson, 2016
[4] Gradient flows on graphons - Oh, Pal, Somani, Tripathi, 2021

Topology and Metric over Graphons

The set of Graphons comes with nice topological and geometrical properties.

- Topology induced by **Cut metric**:[2]
  - Provides a natural topology over graphons that **captures graph convergence**.
  - Provides **compactness** when edge-weights are bounded, say in $[-1, 1]$.

- **Invariant L^2 metric δ_2**:[3]
  - We show that it is a **geodesic** metric.[4]
  - Allows defining geodesic convexity on $(\widehat{\mathcal{W}}, \delta_2)$.
  - Allows a notion of 'gradient' on $(\widehat{\mathcal{W}}, \delta_2)$.
  - Allows construction of 'gradient flows' on $(\widehat{\mathcal{W}}, \delta_2)$.

[2] Convergent sequences of dense graphs I - Borgs, Chayes, Lovász, Sós, Vesztergombi, 2008.
[3] Graphons and cut metric on sigma-finite measure spaces - Janson, 2016
[4] Gradient flows on graphons - Oh, Pal, Somani, Tripathi, 2021

Our Results

Algorithms over Euclidean spaces allow us to access this geometry suitably.
Under suitable assumptions, we show that as the step size goes to zero and $n \to \infty$:

Scaling limits of GD

Euclidean GD over finite dimensional Euclidean spaces, converge to a 'gradient flow' on the metric space of graphons as step-size goes to zero and $n \to \infty$.

Our Results

Algorithms over Euclidean spaces allow us to access this geometry suitably.
Under suitable assumptions, we show that as the step size goes to zero and $n \to \infty$:

Scaling limits of GD

Euclidean GD over finite dimensional Euclidean spaces, converge to a 'gradient flow' on the metric space of graphons as step-size goes to zero and $n \to \infty$.

Scaling limits of SGD

Euclidean SGD converges to the same gradient flow on the metric space of graphons.

- Noise in SGD smoothens out due to the regularity of the topology on graphons.

Our Results

Algorithms over Euclidean spaces allow us to access this geometry suitably.
Under suitable assumptions, we show that as the step size goes to zero and $n \to \infty$:

Scaling limits of GD

Euclidean GD over finite dimensional Euclidean spaces, converge to a 'gradient flow' on the metric space of graphons as step-size goes to zero and $n \to \infty$.

Scaling limits of SGD

Euclidean SGD converges to the same gradient flow on the metric space of graphons.

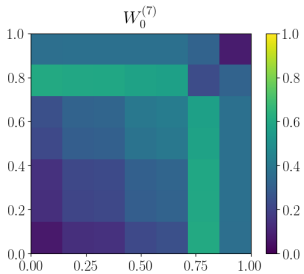
- Noise in SGD smoothens out due to the regularity of the topology on graphons.

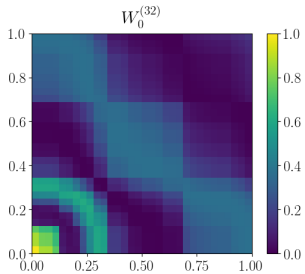
Scaling limits of SGD with added non-trivial noise

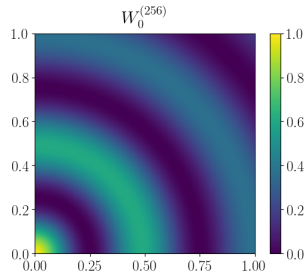
Euclidean SGD with non-trivial added noise converge to curves that are not gradient flows, but can be characterized as fixed points of an iteration scheme.

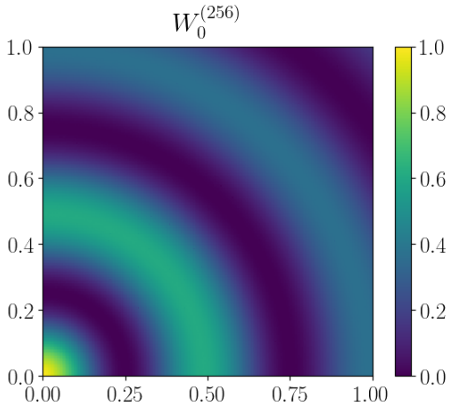
Simulations

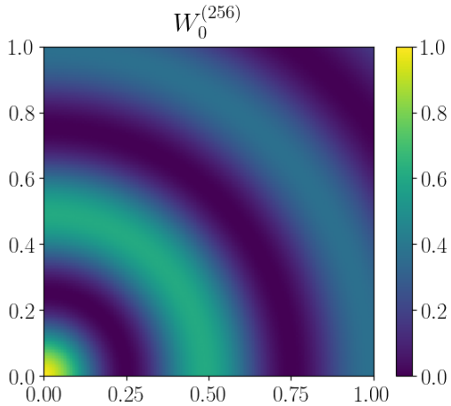
- Turán's theorem: The n-vertex triangle-free graph with the maximum number of edges is a complete bipartite graph.







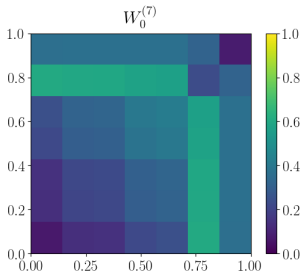


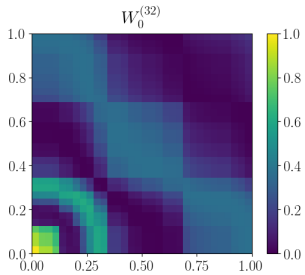


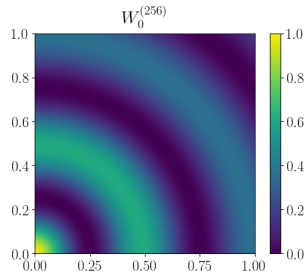
Simulations

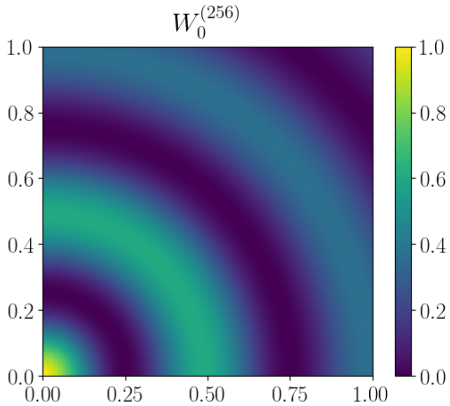
- Turán's theorem: The n-vertex triangle-free graph with the maximum number of edges is a complete bipartite graph.

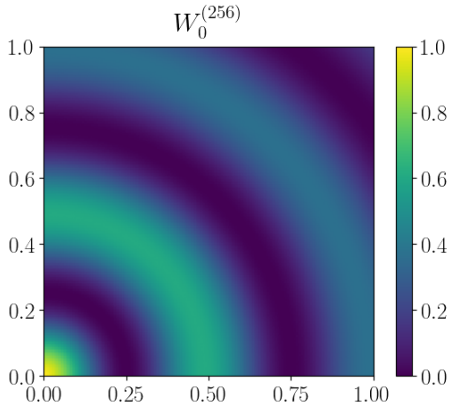
Q. Can we recover this theorem through an optimization problem on graphons?







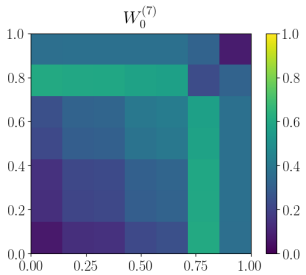




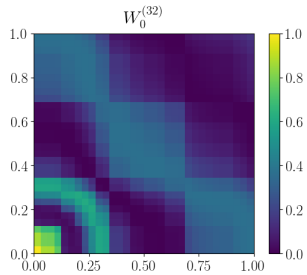
Simulations

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

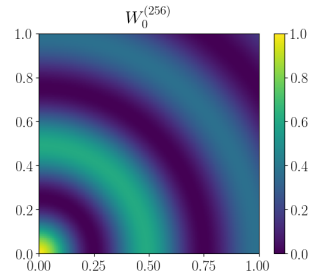
Q. Can we recover this theorem through an optimization problem on graphons?



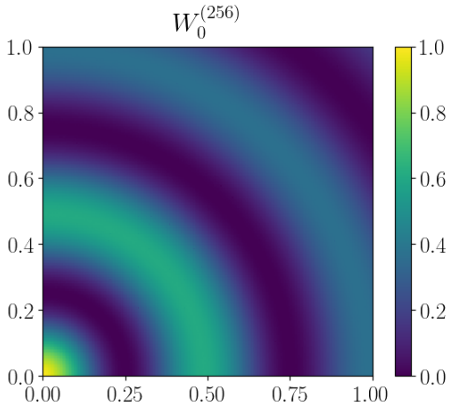
(a) GD ($n = 7$)

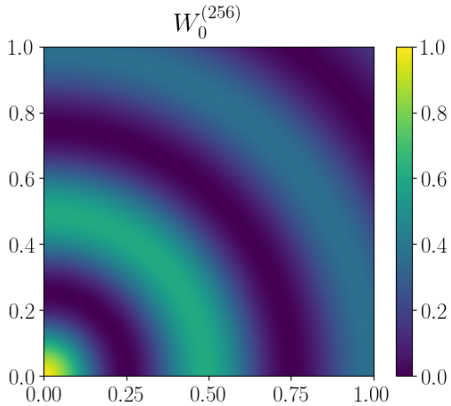


(b) GD ($n = 32$)



(c) GD ($n = 256$)

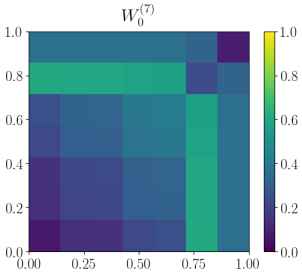




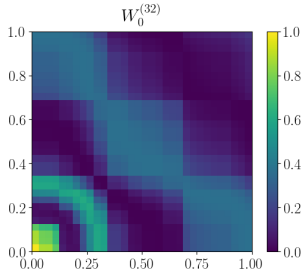
Simulations

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

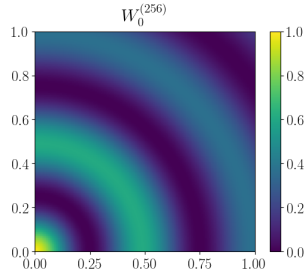
Q. Can we recover this theorem through an optimization problem on graphons?



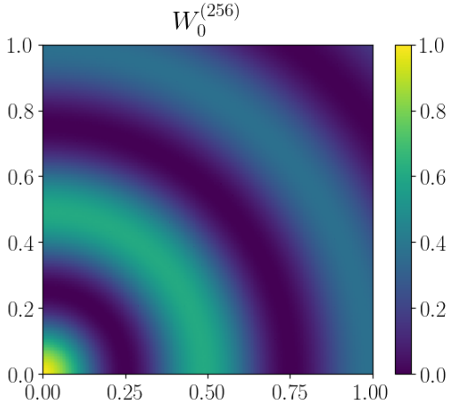
(a) GD ($n = 7$)



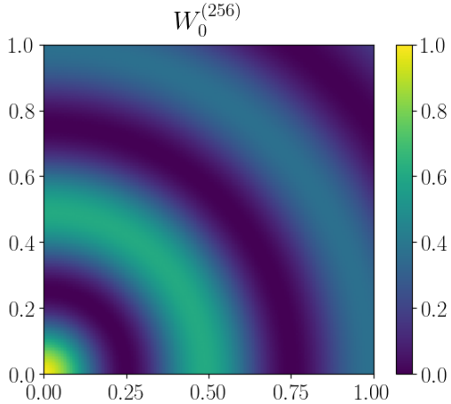
(b) GD ($n = 32$)



(c) GD ($n = 256$)



(a) SGD ($n = 256$)



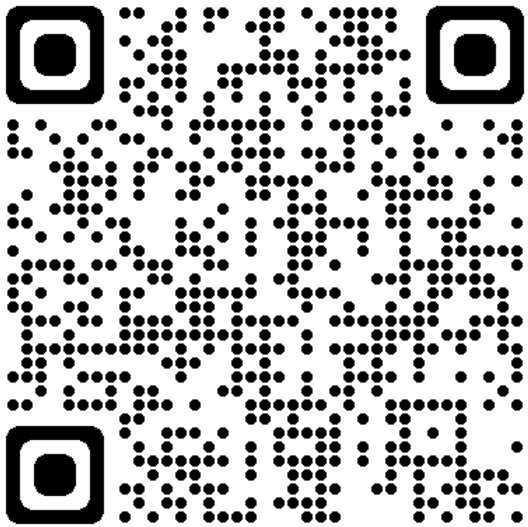
(b) SGD + Brownian noise ($n = 256$)

Future directions

- Extend the geometry to consider symmetry groups in **Deep NNs**.
- What about processes on non-weighted graphs? Sparse graphs? Open.

Thank you!

- ArXiv version[5]: https://arxiv.org/abs/2111.09459



[5] Gradient flows on graphons: existence, convergence, continuity equations - Oh, Pal, Somani, Tripathi, 2021