Implicit Neural Representations for Variational Problems on Graphons

Taeyoung Kim
School of Computing, KAIST
School of Computational Sciences, KIAS
taeyoungkim21@kaist.ac.kr

Jineon Baek
June E Huh Center for Mathematical Challenges, KIAS
jineon.kias@gmail.com

Joonkyung Lee
School of Computational Sciences, KIAS
joonkyunglee@kias.re.kr

Hongseok Yang
School of Computing, KAIST
School of Computational Sciences, KIAS
hongseokyang@kias.re.kr

Abstract

We propose a neural exploratory framework for *variational problems over graphons*[1], i.e., symmetric measurable functions $[0,1]^2 \to [0,1]$, that arise as limits of dense graph sequences. Such problems are central to two parts of the mathematical literature: extremal graph theory, which studies graph parameter optimisation under given restrictions, often for those graphs with a large number of vertices, and the large deviation theory of dense random graphs, which characterises the structure of rare events through constrained graphon optimisation. We represent graphons by implicit neural networks and optimise graphon objectives by gradient descent. Our design combines three ingredients: a multi-scale sinusoidal residual architecture biased toward sharp, step-like graphons; an embedded solver that enforces a single empirical density constraint and is differentiated by implicit differentiation; and symmetry-aware Monte Carlo estimators. On generalised Turán problems that previously required substantial human effort, our framework rediscovers known optimal graphons without human intervention. Applied to open instances, it produces candidate optima, including a previously unreported family of extremal structures. Applied to the variational problems from the large deviation theory of Erdős–Rényi random graphs, the framework produces new candidate optimal graphons across both upper- and lower-tail regimes for multiple pattern graphs; for the case of the triangle graph and the upper-tail regime, these candidates improve upon the best-known reference construction of Lubetzky and Zhao. The structures found by our framework have also led to new theorems on the optimal graphons for both families of problems.

1 Introduction

Machine learning is opening new directions in mathematical discovery. Recent work spans automatic provers and autoformalisation for Olympiad-style or research-level problems [Polu et al., 2023, Trinh

[1] A variational problem is an optimisation problem over a function space, in which the objective and constraints are expressed by means of functionals, i.e., mappings from the function space to $\mathbb{R}$.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

et al., 2024, Jiang et al., 2023, Wu et al., 2022, Hubert et al., 2026, Huang and Yang, 2025, Alexeev et al., 2026], neural search for combinatorial constructions [Wagner, 2021, Charton et al., 2024, Romera-Paredes et al., 2024, Novikov et al., 2025], and pattern-driven progress on long-standing conjectures in pure mathematics [Davies et al., 2021, He et al., 2025]. In this paper, we extend this line of research to a class of *continuous optimisation problems over graphons*, with two motivating problem families drawn from extremal graph theory and the large deviation theory of dense random graphs.

A central object in extremal graph theory is the *homomorphism density* $t(H, G)$ between finite undirected simple graphs H and G: the probability that a uniformly random map from $V(H)$ to $V(G)$ is a graph homomorphism. As a basic example, the constrained optimisation problem

$$\max_{G} t(K_2, G) \quad \text{subject to} \quad t(K_3, G) = 0 \ \wedge \ |V(G)| = n \tag{1}$$

asks for an n-vertex triangle-free graph that maximises the edge density. Here K_m denotes the complete graph on m vertices, so K_2 and K_3 represent an edge and a triangle, respectively. Although the answer is given by the classical theorem of Mantel [Mantel, 1907], a large family of similar problems involving homomorphism densities remains open [Sidorenko, 1993, Borgs et al., 2006].

A standard approach to such problems is to study their *asymptotic* counterparts, taking $n \to \infty$ and analysing the limit of the optimum. These asymptotic problems can often be reformulated as continuous optimisation over *graphons*, i.e., symmetric measurable functions $[0,1]^2 \to [0,1]$, that arise as limit objects of dense graph sequences [Lovász and Szegedy, 2006, Lovász, 2012]. Another source of graphon optimisation problems comes from the large deviation theory of dense random graphs [Chatterjee and Varadhan, 2011], where the limiting structure of an Erdős–Rényi graph conditioned on an unusually large or small homomorphism density is itself characterised by a constrained optimisation over graphons. In both settings, the objective and constraints are integrals against the graphon W (homomorphism densities or pointwise relative entropy), and known or conjectured optima are step-like.

We attack these graphon-optimisation problems directly with deep learning, representing a graphon W as a neural network in the spirit of implicit neural representations [Park et al., 2019, Mescheder et al., 2019, Chen and Zhang, 2019] and minimising the objective by gradient descent on the network parameters. Three recurring challenges shape our design. First, known and conjectured optima are typically step-like, and standard architectures struggle with such targets because of their bias toward smooth functions. We address this with a *multi-scale sinusoidal residual architecture* that injects sinusoidal encodings of growing frequency into successive layers, biasing the model toward sharp, step-like solutions. Second, the problems often carry a single scalar density constraint (a fixed homomorphism density) that is poorly handled by penalty terms; we embed a numerical solver for an empirical version of the constraint into the forward pass and backpropagate through it via implicit differentiation, as in deep equilibrium models [Bai et al., 2019]. Third, both the objective and the constraints are graphon integrals that must be estimated from samples, including the high-dimensional homomorphism-density integrals; we use symmetry-aware Monte Carlo estimators.

We demonstrate our framework on two families of graphon optimisation problems: **(P1)** generalised Turán problems from extremal graph theory, minimising a homomorphism density under a fixed edge-density constraint; and **(P2)** the variational problems from the large deviation theory of dense random graphs [Chatterjee and Varadhan, 2011], minimising a relative-entropy functional under a fixed homomorphism-density constraint. On **(P1)**, our framework rediscovers nontrivial known optima that previously required substantial human effort [Goodman, 1959, Bollobás, 1976, Lovász and Simonovits, 1983, Fisher, 1989, Razborov, 2008], produces candidate minimisers for open instances such as minimisation of C_7-homomorphism density, and through a systematic sweep over small graphs, uncovers, for a six-vertex graph, a previously unreported family of extremal structures that do not match the complete k-partite pattern of classical examples. Applied to **(P2)**, the framework produces new candidate optimal graphons, in the cases where the optimal structure is not fully understood, for four pattern graphs in the upper-tail regime and K_3 in the lower-tail regime. For K_3, these candidates consistently improve upon the reference construction of Lubetzky and Zhao [2015] in the upper-tail regime. We view the framework as a flexible exploratory tool that complements rigorous lower-bound techniques such as flag algebras [Razborov, 2007, Falgas-Ravry and Vaughan, 2013]: it produces candidate upper bounds, concrete candidate minimisers, and structural conjectures that can be inspected, refined, or subsequently certified. The structures found by our method have also led to new theorems (§D). For problems of **(P1)**, the balanced complete k-partite graphon minimises

the density of every odd cycle at $p = 1 - 1/k$ for every $k \geq 3$, and that of every connected chordal graph whose maximal cliques have a common size $r \geq 3$ at $p = 1 - 1/k$ for every $k \geq r$. For problems of **(P2)**, close to the boundary of the region where the constant graphon is optimal, the upper-tail problem for regular graphs has a unique optimal graphon, which is bipodal.

2 Background

For a positive integer n, write $[n] \coloneqq \{1, 2, \ldots, n\}$. Throughout the paper, a *graph* is an undirected, finite, simple graph (no self-loops and no multiple edges). For a graph G, we write $V(G)$ and $E(G)$ for its vertex and edge sets, and let $v(G) \coloneqq |V(G)|$ and $e(G) \coloneqq |E(G)|$. We use standard notation for familiar graphs: K_n denotes the complete graph on n vertices (i.e., $V(K_n) = [n]$ and $E(K_n) = \{\{i, j\} : i, j \in [n],\, i \neq j\}$), $K_{1,n}$ denotes the complete bipartite graph with part sizes 1 and n (i.e., $V(K_{1,n}) = [n+1]$ and $E(K_{1,n}) = \{\{1, i\} : 2 \leq i \leq n+1\}$), and C_n denotes the cycle on n vertices (i.e., $V(C_n) = [n]$ and $E(C_n) = \{\{i, i+1\} : 1 \leq i \leq n-1\} \cup \{\{n, 1\}\}$).

Let H and G be graphs. A *homomorphism* from H to G is a map $\phi : V(H) \to V(G)$ that preserves adjacency, i.e., $\{\phi(u), \phi(v)\} \in E(G)$ for every $\{u, v\} \in E(H)$. The *homomorphism density* from H to G, denoted $t(H, G)$, is the probability that a uniformly random map $\phi : V(H) \to V(G)$ is a homomorphism. Equivalently,

$$t(H, G) \coloneqq \frac{1}{v(G)^{v(H)}} \sum_{\phi : V(H) \to V(G)} \prod_{\{u,v\} \in E(H)} \mathbb{I}\big[\{\phi(u), \phi(v)\} \in E(G)\big], \tag{2}$$

where $\mathbb{I}[\cdot]$ is the indicator function. Homomorphism densities are a natural language for extremal graph problems: they are size-normalised statistics of fixed patterns, so objectives and constraints are comparable across graphs of different sizes, and they are central to dense graph limit theory, quasirandomness, and regularity-based approximation [Lovász, 2012].

A *graphon* W is a symmetric measurable function $W : [0, 1]^2 \to [0, 1]$. Intuitively, W can be viewed as the (probabilistic) adjacency matrix of a continuum graph with vertex set $[0, 1]$: $W(x, y)$ is the probability of an edge between x and y. Graphons arise as limits of dense graphs: along any graph sequence $(G_n)_{n \in \mathbb{N}}$ whose homomorphism densities $t(H, G_n)$ converge for every fixed graph H with vertex set $[v(H)]$, there is a graphon W with $t(H, G_n) \to t(H, W)$ for every H with vertex set $[v(H)]$, where $t(H, W)$ is the graphon homomorphism density of H in W defined by

$$t(H, W) \coloneqq \mathbb{E}_{x_1, \ldots, x_{v(H)} \overset{\text{iid}}{\sim} \mathrm{Unif}[0,1]} \Big[\prod_{\{i,j\} \in E(H)} W(x_i, x_j) \Big]. \tag{3}$$

Conversely, every graphon arises in this way [Lovász and Szegedy, 2006, Lovász, 2012]. Familiar discrete structure is reflected in W: balanced complete bipartite graphs converge to the two-block graphon equal to 1 on the off-diagonal blocks and 0 on the diagonal blocks; an Erdős–Rényi $G(n, q)$ converges to the constant graphon $W \equiv q$; a stochastic block model converges to a multi-block graphon whose blocks correspond to the communities.

Graphons provide a common language for both problem families studied in this paper: the objective and constraints in each involve graphon homomorphism densities $t(H, W)$ or closely related graphon integrals, and the structure of the optimal graphons is the central object of interest. The first family arises from extremal graph theory, where optimising over graphons is equivalent to optimising the corresponding asymptotic finite-graph problem [Lovász, 2012]; the second arises from the large deviation theory of dense random graphs. We describe each in turn.

Constrained density minimisation. The first problem family is concerned with minimising the homomorphism density of a fixed graph H subject to a fixed edge density p. To state the asymptotic problem formally, we phrase it in terms of graph sequences:

$$\inf \Big\{ \liminf_{n \to \infty} t(H, G_n) \,\Big|\, (G_n) \text{ is a graph sequence with } v(G_n) \to \infty \text{ and } t(K_2, G_n) \to p \Big\}. \tag{4}$$

By the graph-limit theory [Lovász, 2012], this equals $\min_{W :\, t(K_2, W) = p} t(H, W)$, the graphon problem our method optimises directly. It generalises Turán's theorem [Turán, 1941] and the Erdős–Stone–Simonovits theorem [Erdős and Stone, 1946, Erdős and Simonovits, 1966]. These classical results in extremal graph theory say that for every H with at least one edge and $p \leq 1 - 1/(\chi(H) - 1)$, where

$\chi(H)$ is the chromatic number of H, the optimum is 0. But for $p > 1 - 1/(\chi(H) - 1)$, the problem is much harder. It is completely solved only for cliques: $H = K_r$ for all $r \geq 3$ [Reiher, 2016], with earlier results for $r = 3$ [Razborov, 2008] and $r = 4$ [Nikiforov, 2011]. For C_5, the minimiser is known only at the special values $p = 1 - 1/k$ (k a positive integer) [Bennett et al., 2020].

Variational problem from large deviation theory. The large deviation theory of dense random graphs addresses two related questions: at what exponential rate does the Erdős–Rényi random graph $G(n, q)$ have an atypically large or small homomorphism density, and what does the graph typically look like when such a rare event occurs? These questions are of fundamental interest in probability theory and combinatorics, and they carry information-theoretic structure that is natural from an ML perspective: the Chatterjee–Varadhan framework reduces both questions to a single constrained optimisation over graphons whose objective is an averaged KL divergence [Chatterjee and Varadhan, 2011].

Concretely, fix $q \in (0, 1)$ and let $h_q(W) \coloneqq \int_{[0,1]^2} \mathrm{KL}(\mathrm{Bern}(W(x, y)) \,\|\, \mathrm{Bern}(q))\, \mathrm{d}x\, \mathrm{d}y$, the integrated per-edge KL divergence of W from the base Erdős–Rényi graphon $W \equiv q$. For a fixed graph H and target level $r \in (0, 1)$, the large deviation principle yields a variational characterisation of both tails. In the *upper tail* ($r > q$), $\binom{n}{2}^{-1} \log \Pr[t(H, G(n, q)) \geq r^{e(H)}]$ converges to $-\min_{W : t(H,W) \geq r^{e(H)}} h_q(W)$. In the *lower tail* ($r < q$), $\binom{n}{2}^{-1} \log \Pr[t(H, G(n, q)) \leq r^{e(H)}]$ converges to $-\min_{W : t(H,W) \leq r^{e(H)}} h_q(W)$. In either case, the typical structure of $G(n, q)$ conditioned on the rare event is described by near-minimisers of the corresponding variational problem [Lubetzky and Zhao, 2015]. Thus the minimiser is the graphon closest in KL divergence to $W \equiv q$ that explains the observed *excess* (upper tail) or *deficit* (lower tail) in the H-homomorphism density, in the same spirit as maximum-entropy modelling in ML. Even in the triangle case $H = K_3$, it is not fully understood when the deviation localises (e.g. through a dense clique) versus spreads more uniformly, or what the minimisers look like across (q, r).

Our method applies directly to the boundary form

$$\min_{W : t(H,W) = r^{e(H)}} h_q(W) \tag{5}$$

which is equivalent to the inequality-constrained minimisations above (with $\geq$ in the upper tail $r > q$ and $\leq$ in the lower tail $r < q$).

3 Implicit neural representations of graphons

We tackle the two graphon optimisation problems introduced in §2:

$$\textbf{(P1)} \quad \min_{W : t(K_2,W) = p} t(H, W), \qquad \textbf{(P2)} \quad \min_{W : t(H,W) = r^{e(H)}} h_q(W), \tag{6}$$

for a graph H, reals $p, q, r \in (0, 1)$, where W ranges over graphons. Both problems carry a single scalar equality constraint.

We represent W as a neural network f_θ, recast these graphon optimisations as optimisations over the parameters θ, and apply gradient-based training. A naive instantiation of this idea fails for three reasons. First, known optimal graphons are typically step functions, which standard architectures such as MLPs struggle to represent due to their bias toward smooth functions. Second, the scalar constraints in **(P1)** and **(P2)** are poorly handled by penalty terms, so the architecture must enforce them more directly. Third, each occurrence of $t(\cdot, W)$ involves an intractable integral, and estimating these expectations efficiently requires problem-specific sampling. The remainder of this section describes the architecture and training procedure that address these three challenges.

Architecture. We parameterise graphons by a residual MLP that injects a layer-specific sinusoidal encoding of the input at every layer. To handle constraints in **(P1)** and **(P2)**, the architecture additionally includes a constraint solver.

Fix the number of layers $L \geq 2$, the hidden-layer width d, and a differentiable activation $\phi : \mathbb{R} \to \mathbb{R}$. Also, fix an increasing function $s : [L] \to (0, \infty)$, which determines the scale of weight initialisation at each layer; in our experiments, we use $s(\ell) = 2^{\ell - 1}$. The network $f_\theta[c] : [0, 1]^2 \to [0, 1]$ has a distinguished scalar bias $c \in \mathbb{R}$ and remaining parameters θ, and is defined for $2 \leq \ell \leq L$ by

$$f_\theta[c](x_1, x_2) \coloneqq sm\big(V g_\theta^{(L)}(\min(x_1, x_2), \max(x_1, x_2)) + c\big),$$

$$g_\theta^{(1)}(\mathbf{x}) \coloneqq \sin\big(U^{(1)}\mathbf{x} + b^{(1)}\big),$$
$$g_\theta^{(\ell)}(\mathbf{x}) \coloneqq g_\theta^{(\ell-1)}(\mathbf{x}) + \phi\big(V^{(\ell-1)} g_\theta^{(\ell-1)}(\mathbf{x}) + W^{(\ell)} \sin(U^{(\ell)}\mathbf{x}) + b^{(\ell)}\big).$$

Here sm is the sigmoid, and $g_\theta^{(1)}, g_\theta^{(\ell)}$ are maps from $\{\mathbf{x} = (x_1, x_2) \in [0,1]^2 : x_1 \le x_2\}$ to $\mathbb{R}^d$. The network parameters are $c \in \mathbb{R}$, $b^{(\ell')} \in \mathbb{R}^d$, $V \in \mathbb{R}^{1\times d}$, $U^{(\ell')} \in \mathbb{R}^{d\times 2}$, and $V^{(\ell'')}, W^{(\ell')} \in \mathbb{R}^{d\times d}$, for $\ell' \in [L]$, $\ell'' \in [L-1]$. These parameters are initialised independently by

$$c,\, b_i^{(\ell')},\, V_{1,i},\, V_{i,j}^{(\ell'')},\, W_{i,j}^{(\ell')} \overset{\text{iid}}{\sim} \mathrm{Unif}\big[-1/\sqrt{2d},\, 1/\sqrt{2d}\big], \qquad U_{i,k}^{(\ell')} \overset{\text{iid}}{\sim} \mathrm{Unif}\big[-s(\ell')/\sqrt{2},\, s(\ell')/\sqrt{2}\big],$$

for all $i, j \in [d]$, $k \in [2]$. We collect every parameter except c into θ. Using $\min$ and $\max$ in the definition of $f_\theta[c]$ ensures that $f_\theta[c]$ is symmetric and represents a graphon.

The architecture has two notable features. First, it repeatedly injects a sinusoidal encoding of the input (x_1, x_2) at every layer, with an encoding scale that grows with depth. Increasing $s(\ell)$ with depth shifts higher-frequency content into deeper layers [Tancik et al., 2020] and encourages late layers to capture sharp boundaries, while early layers stabilise the coarse, low-frequency structure, an inductive bias that fits step-like graphons.

Second, in constrained problems, we determine c via a numerical solver rather than learning it directly. To enforce $t(K_2, f_\theta[c]) = p$ in **(P1)**, we use a batch of K i.i.d. samples

$$\mathcal{E} = \big\{(x_1^{(k)}, x_2^{(k)})\big\}_{k\in[K]}, \quad (x_1^{(k)}, x_2^{(k)}) \overset{\text{iid}}{\sim} \mathrm{Unif}([0,1]^2),$$

and define

$$\textsc{Solve}_\theta(\mathcal{E}, p) \coloneqq \text{the unique } c \in \mathbb{R} \text{ such that } \frac{1}{|\mathcal{E}|} \sum_{(x_1,x_2)\in\mathcal{E}} f_\theta[c](x_1, x_2) = p. \tag{7}$$

The left-hand side of the equation in (7) is an unbiased estimator of $t(K_2, f_\theta[c])$, so the equation is the empirical counterpart of the constraint. The solution exists and is unique, since the left-hand side is continuous and strictly increasing in c and tends to 0 and 1 as $c \to -\infty$ and $c \to +\infty$, respectively. We solve (7) with a clipped variant of Newton's method (Algorithm 1) and use the resulting c^* when evaluating $f_\theta[c^*]$ on other inputs. The same idea applies to any single homomorphism-density constraint of the form $t(H, W) = \tau$: one replaces the edge-density estimator in (7) by an estimator of $t(H, f_\theta[c])$. Because increasing the scalar bias c increases every edge probability, this estimator remains monotone in c. For the large-deviation application **(P2)** with $H = K_3$, for instance, this means sampling triples $(x_1^{(k)}, x_2^{(k)}, x_3^{(k)})$, storing those samples in $\mathcal{T}$, and solving the empirical equation

$$\frac{1}{|\mathcal{T}|} \sum_{(x_1,x_2,x_3)\in\mathcal{T}} f_\theta[c](x_1, x_2) f_\theta[c](x_2, x_3) f_\theta[c](x_3, x_1) = r^3.$$

The same construction applies to any other pattern graph H by replacing the triangle-density estimator with the corresponding H-density estimator from (8).

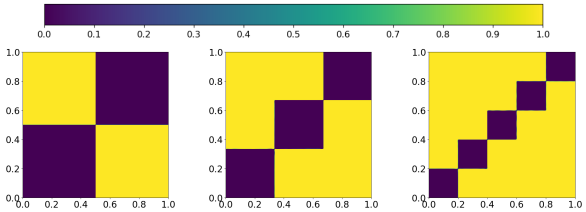
Estimating homomorphism density. Let H be a graph with vertex set $[v(H)]$. By (3), $t(H, f_\theta[c])$ is an expectation that we estimate by Monte Carlo sampling. Our estimator is parameterised by a set $\mathcal{S}$ of permutations of $[v(H)]$. It draws $(N \cdot v(H))$ i.i.d. samples $x_i^{(k)} \overset{\text{iid}}{\sim} \mathrm{Unif}[0,1]$ for $(k, i) \in [N]\times[v(H)]$, and then computes

$$\widehat{t}(H, f_\theta[c]) \coloneqq \frac{1}{N\,|\mathcal{S}|} \sum_{k\in[N],\pi\in\mathcal{S}} \prod_{\{i,j\}\in E(H)} f_\theta[c]\big(x_{\pi(i)}^{(k)}, x_{\pi(j)}^{(k)}\big). \tag{8}$$

The estimator is unbiased for any choice of $\mathcal{S}$ (Lemma E.1); the choice of $\mathcal{S}$ controls variance and computational cost. When $\mathcal{S} = \{\mathrm{id}\}$, our estimator in (8) is the standard Monte Carlo estimator using N independent samples of $(x_1, \ldots, x_{v(H)})$. With multiple permutations, the same N tuples are reused across the $|\mathcal{S}|$ permuted copies, often producing lower variance for the same number of independent tuples. Different permutations also induce repeated evaluations of the network $f_\theta[c]$ on the same unordered pair $(x_i^{(k)}, x_j^{(k)})$ across summands, so caching the evaluations reduces the number of network evaluations from $N \cdot |\mathcal{S}| \cdot e(H)$ to $N \cdot C \cdot e(H)$ for some $C \le |\mathcal{S}|$ that depends on $\mathcal{S}$ and H. We exploit this for larger graphs (e.g., the Petersen graph in §G).

Table 1: K_3 densities in graphons found by our method and optimal construction for K_3 density minimisation under $t(K_2, W) = p$.

p	Optimum	Ours
1/2	0	0.00008130
2/3	0.22222222	0.22232803 (+0.05%)
4/5	0.48	0.48010290 (+0.02%)
4/7	0.10298638	0.10299835 (+0.01%)
5/8	0.17361111	0.17362659 (+0.01%)
5/7	0.31363508	0.31370832 (+0.02%)



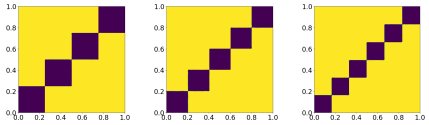
(a) p=1/2. (b) p=2/3. (c) p=4/5.

(d) p=4/7. (e) p=5/8. (f) p=5/7.

Figure 1: Graphons found by our method for K_3 density minimisation under $t(K_2, W) = p$.

Table 2: K_k densities in graphons found by our method and the optimal construction for K_k density minimisation under $t(K_2, W)$=1−1/k.

k	Optimum	Ours (Estimated)
$k = 4$	0.09375	$(0.09403727, 0.09411154)$
$k = 5$	0.0384	$(0.03868024, 0.03872081)$
$k = 6$	0.01543210	$(0.01565511, 0.01568170)$



(a) k=4. (b) k=5. (c) k=6.

Figure 2: Graphons found by our method for K_k density minimisation under $t(K_2, W) = 1 - 1/k$.

Training algorithm. For **(P1)**, fix a graph H on $[v(H)]$ and $p \in (0, 1)$. At each iteration, we sample $(N \cdot v(H))$ i.i.d. uniform values $\{x_i^{(k)}\}_{(i,k)\in[N]\times[v(H)]}$ on $[0, 1]$, and form $\mathcal{D} = \{(x_1^{(k)}, \ldots, x_{v(H)}^{(k)}) : k \in [N]\}$ and the unordered-pair set $\mathcal{E} = \{(x_i^{(k)}, x_j^{(k)}) : k \in [N],\ 1 \le i < j \le v(H)\}$. Then, we set $c_\theta \coloneqq \textsc{Solve}_\theta(\mathcal{E}, p)$ to enforce the empirical constraint, and update

$$\theta_{t+1} \leftarrow \theta_t - \eta_t \nabla_\theta \widehat{t}_{\mathcal{D}}(H, f_\theta[c_\theta])\big|_{\theta=\theta_t},$$

where $\widehat{t}_{\mathcal{D}}$ denotes the estimator in (8) evaluated on the tuples in $\mathcal{D}$. Reusing the $x_i^{(k)}$ across $\mathcal{D}$ and $\mathcal{E}$ saves forward evaluations of $f_\theta[c]$. We resample $(\mathcal{D}, \mathcal{E})$ at every iteration to mitigate overfitting to particular Monte Carlo samples. We set η_t via a cosine-annealing schedule [Loshchilov and Hutter, 2017], which alternates between discovering coarse structure (high learning rate) and refining sharp boundaries (low learning rate).

Computing $\nabla_\theta \widehat{t}_{\mathcal{D}}(H, f_\theta[c_\theta])$ requires only the standard product rule once $\nabla_\theta f_\theta[c_\theta]$ is known; the only non-routine ingredient is the gradient of the solver output c_θ with respect to θ, which we compute via implicit differentiation. A direct alternative is to differentiate through an unrolled Newton iteration with autodiff, but this requires storing every intermediate iterate and is substantially more expensive in time and memory; the same trade-off motivates implicit-differentiation methods in deep equilibrium models [Bai et al., 2019] and neural ODEs [Chen et al., 2018]. The closed-form gradient is given by:

Lemma 3.1. *For all* $\mathbf{x} \in [0, 1]^2$,

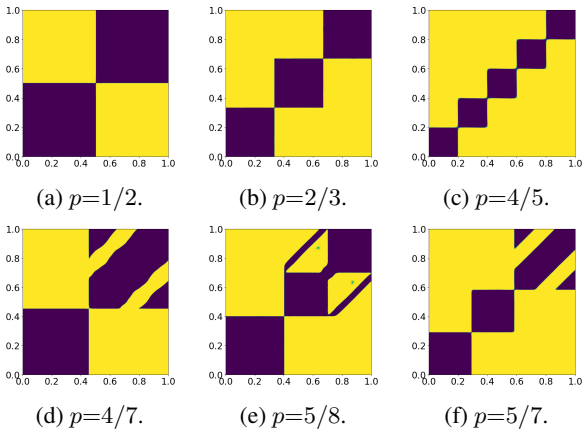
$$\nabla_\theta f_\theta[c_\theta](\mathbf{x}) = \frac{\partial f_\theta[c](\mathbf{x})}{\partial \theta}\bigg|_{c=c_\theta} - sm'\big(h_\theta(\mathbf{x}) + c_\theta\big) \sum_{\mathbf{y}\in\mathcal{E}} w_\theta(\mathbf{y})\, \nabla_\theta h_\theta(\mathbf{y}),$$

$$\textit{where } w_\theta(\mathbf{y}) \coloneqq \frac{sm'(h_\theta(\mathbf{y}) + c_\theta)}{\sum_{\mathbf{z}\in\mathcal{E}} sm'(h_\theta(\mathbf{z}) + c_\theta)}, \quad h_\theta(x_1, x_2) \coloneqq V g_\theta^{(L)}\big(\min(x_1, x_2), \max(x_1, x_2)\big).$$

For **(P2)**, despite its different mathematical origins in large deviation theory, the same constrained loop applies with two substitutions: for a given pattern graph H and $0 <$

Table 3: C_5 densities in graphons found by the known or conjectured optimal construction and our method for C_5 density minimisation under $t(K_2, W) = p$.

p	**Known** or **Conjectured**	**Ours**
1/2	0	0.00004086
2/3	0.12345679	0.12350364 (+0.04%)
4/5	0.3264	0.32644549 (+0.01%)
4/7	0.04587149	0.04590020 (+0.06%)
5/8	0.08653616	0.08657084 (+0.04%)
5/7	0.18200809	0.18206173 (+0.03%)



(a) p=1/2. (b) p=2/3. (c) p=4/5.

(d) p=4/7. (e) p=5/8. (f) p=5/7.

Figure 3: Graphons found by our method for C_5 density minimisation under $t(K_2, W) = p$.

$q < 1$ with target $r \neq q$, the edge-density solver is replaced by the H-density solver targeting $t(H, f_\theta[c]) = r^{e(H)}$, and the objective is replaced by the Monte Carlo estimate of $h_q(f_\theta[c_\theta]) = \int_{[0,1]^2} \mathrm{KL}(\mathrm{Bern}(f_\theta[c_\theta](x,y)) \,\|\, \mathrm{Bern}(q))\,\mathrm{d}x\,\mathrm{d}y$, obtained by averaging $\mathrm{KL}(\mathrm{Bern}(f_\theta[c_\theta](x,y)) \,\|\, \mathrm{Bern}(q))$ over a fresh batch of pairs (x, y). The sigmoid output ensures the KL objective is differentiable in θ, and an analogous implicit-differentiation formula applies to the H-density constraint.

4 Experimental results

We evaluate our framework on instances of **(P1)** and **(P2)** drawn from the extremal-graph-theory and large-deviation literatures. The same architecture, solver, and training loop are used throughout; the only changes between problem families are which constraint is enforced and how the objective is estimated. The experiments are organised to demonstrate two capabilities: (i) rediscovering classical optima that previously required substantial human effort, and (ii) producing mathematically interesting candidates for open instances. An ablation study at the end isolates the contribution of each architectural component.

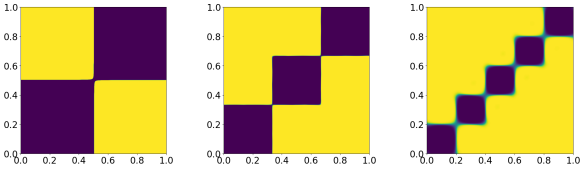
We visualise graphons as Viridis heatmaps (blue $= 0$, yellow $= 1$) (Fig. 1). Reported homomorphism densities are either numerically computed by discretising the learnt graphon into an $M \times M$ weighted adjacency matrix, or estimated using 2^{28} samples with a 95% bootstrap confidence interval (shown in *italic* font) when discretisation is infeasible. Across training runs over the 8 configurations of activation and learning rate $(\phi, \eta) \in \{\mathrm{GELU}, \sin\} \times \{10^{-3}, 5\times10^{-4}, 10^{-4}, 5\times10^{-5}\}$, we report the lowest-objective run; reported values should be read as best-found candidate upper bounds, and the bootstrap intervals quantify Monte Carlo estimation error rather than variability across runs. See §F for details.

Constrained homomorphism-density minimisation. We apply our method to **(P1)** across three classes of instances: clique densities ($H = K_n$, fully solved [Razborov, 2008, Reiher, 2016]), odd-cycle densities ($H = C_5, C_7$, partly open), and a systematic sweep over 117 non-bipartite graphs with $v(H) \leq 6$. The clique experiments validate the method against known optima; the latter two produce new candidate minimisers and, through the exhaustive sweep, a previously unreported family of extremal structures. We describe each in turn.

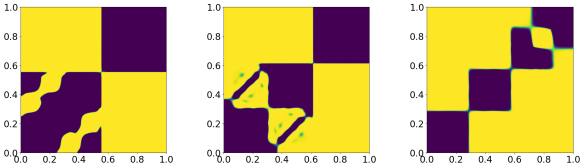
Clique densities ($H = K_n$). We begin with the case $n = 3$. When $p = 1 - 1/k$ for some integer $k \geq 2$, the optimal graphon is the complete k-partite graphon or any of its measure-preserving permutations. For $i \in [k]$, let $X_i = ((i-1)/k, i/k]$. The complete k-partite graphon W_k is defined as follows: $W_k(x_1, x_2) = 1$ if $x_1 \in X_i$ and $x_2 \in X_j$ with $i \neq j$; otherwise, $W_k(x_1, x_2) = 0$. For a measure-preserving ϕ on $[0, 1]$, its permutation is $W_{k,\phi}(x_1, x_2) = W_k(\phi(x_1), \phi(x_2))$. Every optimum is of this form in this case of $p = 1 - 1/k$. When $1 - 1/k < p < 1 - 1/(k+1)$, the minimisation admits multiple types of optima. One example is a variant W_p of W_{k+1}, where the

Table 4: C_7 densities in graphons found by the conjectured optimal construction and our method for C_7 density minimisation under $t(K_2, W) = p$.

p	Conjectured	Ours
1/2	0	0.00001517
2/3	0.05761317	0.05763195 (+0.03%)
4/5	0.209664	0.20967621 (+0.01%)
4/7	0.01713510	0.01715017 (+0.09%)
5/8	0.03610272	0.03611957 (+0.05%)
5/7	0.09455118	0.09457182 (+0.02%)



(a) p=1/2. (b) p=2/3. (c) p=4/5.

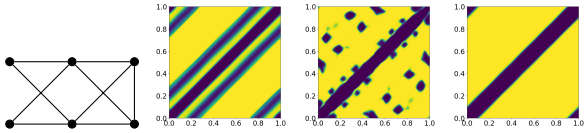


(d) p=4/7. (e) p=5/8. (f) p=5/7.

Figure 4: Graphons found by our method for C_7 density minimisation under $t(K_2, W) = p$.

Table 5: H_6 densities in graphons found by our method, the k-partite construction, and the constant graphon, for H_6 density minimisation under $t(K_2, W)$=p.

p	Ours	k-partite	Constant
2/3	0.02476414	0.03292181	0.02601229
3/4	0.07218717	0.07617188	0.07508469
4/5	0.12985603	0.13056	0.13421773



(a) H_6. (b) p=2/3. (c) p=3/4. (d) p=4/5.

Figure 5: The graph H_6 (six vertices, nine edges) and graphons found by our method for H_6 density minimisation under $t(K_2, W) = p$.

first k parts are equal in length but the last part X_{k+1} is shorter. Again, its permutations $W_{p,\phi}$ by measure-preserving ϕ on $[0, 1]$ are also optimal, though the $W_{p,\phi}$ do not exhaust all optima.

We considered $p \in \{1/2, 2/3, 4/5, 4/7, 5/8, 5/7\}$. For the three cases with $p = 1 - 1/k$ ($p \in \{1/2, 2/3, 4/5\}$), our method produced near-optimal graphons (top row of Fig. 1) within 0.1% of the theoretical optima, closely matching the complete k-partite graphons W_k; among the equivalent measure-preserving permutations, our method consistently converged to these simpler representatives. For $p \in \{4/7, 5/8\}$, which satisfy $1 - 1/k < p < 1 - 1/(k + 1)$, the learnt graphons (Fig. 1d and Fig. 1e) capture the core structure of the known optimal W_p. For $p = 5/7$, the learnt graphon (Fig. 1f) does not match W_p, but its bipartite middle-block structure matches the characterisation of all optimal graphons in [Pikhurko and Razborov, 2017].

Turning to larger cliques ($k > 3$), we ran $\min_W t(K_k, W)$ s.t. $t(K_2, W)$=$1-1/k$ for $k \in \{4, 5, 6\}$. Even though the larger motif size increases the variance of our estimator and the optimal graphon involves more blocks, our method recovers the complete k-partite structure (Fig. 2).

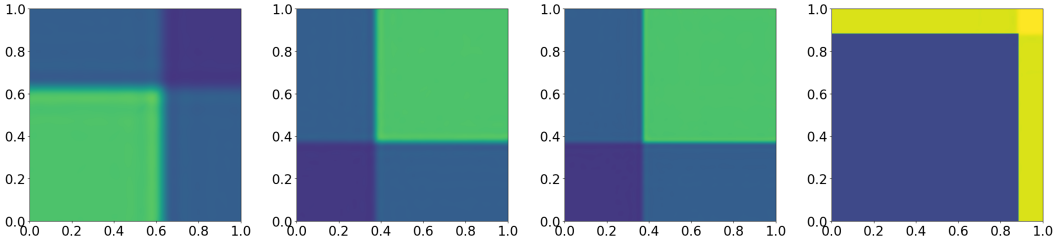


Figure 6: Learnt graphons for upper-tail variational problems with $q = 0.1, r = 0.5$ for $H = K_3, C_4, C_5, K_{1,3}$ (left to right).

Table 6: Relative-entropy $h_q(W)$ at our learnt graphon (W_{ours}) and at the tripartite construction W_{LZ} of Lubetzky and Zhao [2015], for **(P2)** with $H = K_3$ on $(q, r) \in \{0.05, 0.1\} \times \{0.4, 0.5, 0.6\}$, together with a lower bound on the optimum (§G). Lower is better.

(q, r)	$(0.05, 0.4)$	$(0.05, 0.5)$	$(0.05, 0.6)$	$(0.10, 0.4)$	$(0.10, 0.5)$	$(0.10, 0.6)$
$h_q(W_{\text{LZ}})$	0.556057	0.830366	1.144945	0.311239	0.510826	0.750684
$h_q(W_{\text{ours}})$	0.464208	0.727881	1.051203	0.309109	0.506060	0.747694
Lower bound	0.460284	0.725284	1.049172	0.306561	0.504311	0.746006

Open problems: C_5 and C_7. We next apply our method to two graphon optimisation problems that, to the best of our knowledge, remain open in the tested regimes: minimising the homomorphism densities of C_5 and C_7, each under a fixed edge density p. In §G, we also report results for the Petersen graph, for which no prior optima or conjectured optima are known to us.

For C_5, Bennett et al. [2020] characterised the minimisers when $p = 1 - 1/k$ and conjectured them for other p. We applied our method for $p \in \{1/2, 2/3, 4/5, 4/7, 5/8, 5/7\}$ (Fig. 3); the resulting values agree with the conjectured optima within 0.1% (Tab. 3). When $p = 1 - 1/k$, the learnt graphon recovers the complete k-partite structure (top row); for other p, it exhibits pronounced diagonal (banded) structure (bottom row), suggesting a qualitatively different extremal configuration away from the breakpoints $p = 1 - 1/k$. For C_7, no prior results are known to us. Motivated by the C_5 results of Bennett et al. [2020], we propose a conjecture for the structure and optimum value of C_7-minimising graphons under a fixed edge density; see §C. Empirically, our learnt graphons match our conjectured optima within 0.1% across all tested p (Tab. 4). For $p \in \{1/2, 2/3, 4/5\}$ ($p = 1 - 1/k$), they recover the complete k-partite structure (top row of Fig. 4); for $p \in \{4/7, 5/8\}$, they show diagonal patterns similar to C_5, suggesting that this is a common feature of odd-cycle minimisation; for $p = 5/7$, the structure is qualitatively different, with block-like rather than diagonal patterns (bottom-right of Fig. 4).

Exhaustive study over small graphs. We now describe a systematic study enabled by our method. We applied our method to the problem **(P1)** in (6) for all non-bipartite graphs H with $v(H) \leq 6$ and with no isolated vertices or isolated edges, for a total of 117 graphs; isolated vertices and isolated edges are excluded because they only multiply $t(H, W)$ by 1 and p, respectively. For each H, we trained our network under a single fixed hyperparameter configuration (GELU activation, learning rate 10^{-3}) with three random seeds to minimise $t(H, W)$ subject to $t(K_2, W) = p$, for the six densities $p = 1 - 1/(k - 1 + j)$ with $j \in \{1/3, 1/2, 2/3, 1, 2, 3\}$, where k is the chromatic number of H. These values of p are chosen to probe the structure of the curve $x \mapsto \min_{W: t(K_2, W) = x} t(H, W)$: the values with $j \in \{1, 2, 3\}$ correspond to breakpoints of piecewise-defined curves established for several classes of graphs [Razborov, 2008, Reiher, 2016, Bennett et al., 2020], while those with $j < 1$ lie inside the segment $(k - 2)/(k - 1) < p < (k - 1)/k$. To assess the optimality of our solutions, we computed lower bounds on the minimum H-density using flag algebras [Razborov, 2007] on 7-vertex graphs.

Taking the best of the three seeds, the H-density of the learnt graphon is within 1% of the flag-algebra lower bound in 72.6% of the 702 (H, p) instances, and a graphon whose H-density is more than 1% smaller is known in 20.8% of them; neither holds for the remaining 6.6%.

Using the lower bounds, we also examined whether the minimum is convex or concave on the segment containing the three densities with $j < 1$. For 104 of the 117 graphs, including every graph with $k = 3$ except K_3, the minimum appears to be convex on this segment, as conjectured for odd cycles [Bennett et al., 2020]; for 7 graphs, it appears to be concave, as for cliques [Reiher, 2016]; and for 5 graphs, including K_6 minus an edge, it is neither convex nor concave. Our experiments also suggest a novel family of extremal graphon structures. When $p = 1 - 1/k$, the minimisers for cliques and C_5 are known to be complete k-partite [Reiher, 2016, Bennett et al., 2020], and we conjecture the same for C_7; in all three cases, our method recovers this structure (Figs. 1, 3 and 4). However, this is not universal: for the graph H_6 in Fig. 5a at the same p, our method finds a candidate minimiser that is neither complete k-partite nor constant (Fig. 5); the learnt graphons exhibit a geometric pattern in which edge probabilities appear to depend on the circular distance when $[0, 1]$ is interpreted as a circle.

Variational problem from large deviation theory. We apply our framework to **(P2)**, the variational problem of Chatterjee and Varadhan [2011]. We test $(q, r) \in \{0.05, 0.1\} \times \{0.4, 0.5, 0.6\}$ for upper-tail variational problems with $H \in \{K_3, C_4, C_5, K_{1,3}\}$, and $(q, r) \in \{0.7, 0.8, 0.9\} \times \{0.05, 0.1\}$ for lower-tail variational problems with $H = K_3$; in all these regimes, the structure of optimisers is, to our knowledge, not fully understood. As a reference for $H = K_3$, we use the tripartite construction W_{LZ} proposed by Lemma 3.4 of Lubetzky and Zhao [2015]; Tab. 6 compares $h_q(W_{\mathrm{LZ}})$ to $h_q(W_{\mathrm{ours}})$ at our learnt graphon W_{ours} in the upper-tail regime. Qualitatively, our method consistently finds two-block structures for all graphs and in both upper- and lower-tail regimes. For the upper-tail problems, the 2-regular graphs K_3, C_4, C_5 share close edge densities and block sizes in their respective learnt graphons, while the non-2-regular graph $K_{1,3}$ shows a qualitatively different structure with a much smaller block size; see Fig. 6 for the learnt graphons with $q = 0.1, r = 0.5$ and §G for all upper-tail regimes. For the lower-tail problems, we obtain a qualitatively different structure of nearly-bipartite graphons. Quantitatively, in the upper-tail problems for $H = K_3$, our method consistently finds graphons with lower relative entropy than the tripartite construction W_{LZ}, showing that the latter is suboptimal in these regimes; see §G for details.

Ablation study, baselines, and multi-task variant. To assess the contribution of each architectural component and to compare the neural approach with simpler alternatives, we ran an ablation study and four non-neural baselines on triangle-density minimisation under $t(K_2, W) = 7/9$. Replacing the sinusoidal features with GELU features, the constraint solver with a regulariser, or the cosine-annealing schedule with a constant learning rate increases the triangle density, and using the sinusoidal encoding only in the first layer degrades the recovered block structure; the best constant encoding scale reaches a similar density. The four non-neural baselines tailored to step graphons (a fixed-block SBM with 256 blocks; SBMs with trainable block sizes; a decision tree; a tree ensemble) all reach higher triangle densities than our method. To explore whether a single model can handle multiple problem instances simultaneously, we also trained a FiLM-conditioned multi-task variant over $p \in [1/2, 2/3]$ for $H = K_3$ [Perez et al., 2018], which demonstrates that the architecture extends naturally to this setting. Full details are deferred to §I.

Acknowledgments and Disclosure of Funding

We thank Sangho Lim, Gyeongwon Jung, Seonghun Park, Seonghyuk Im, and Kyeongsik Nam for helpful discussions. T. Kim and H. Yang were supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. RS-2023-00279680). H. Yang was also supported by a KIAS Individual Grant (CG109601) at the Korea Institute for Advanced Study. J. Lee was supported by Samsung STF Grant (SSTF-BA2201-02) and the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. NRF-2022R1C1C1010300) . J. Baek was supported by KIAS Individual Grant (HP105001).

References

Boris Alexeev, Kevin Barreto, Yanyang Li, Jared Duker Lichtman, Liam Price, Jibran Iqbal Shah, Quanyu Tang, and Terence Tao. Primitive sets and von Mangoldt chains: Erdős problem #1196 and beyond, 2026. URL `https://arxiv.org/abs/2605.00301`.

Ali Azizpour, Nicolas Zilberstein, and Santiago Segarra. Scalable implicit graphon learning. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3952–3960. PMLR, 2025. URL `https://proceedings.mlr.press/v258/azizpour25a.html`.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/01386bd6d8e091c2ab4c7c7de644d37b-Abstract.html`.

Patrick Bennett, Andrzej Dudek, Bernard Lidický, and Oleg Pikhurko. Minimizing the number of 5-cycles in graphs with given edge-density. *Combinatorics, Probability and Computing*, 29(1): 44–67, 2020. doi: 10.1017/S0963548319000257.

Béla Bollobás. Relations between sets of complete subgraphs. In C. St. J. A. Nash-Williams and J. Sheehan, editors, *Proceedings of the Fifth British Combinatorial Conference*, pages 79–84, Winnipeg, 1976. Utilitas Mathematica.

Christian Borgs, Jennifer Chayes, László Lovász, Vera T. Sós, and Katalin Vesztergombi. Counting graph homomorphisms. In *Topics in Discrete Mathematics*, pages 315–371. Springer, 2006. doi: 10.1007/3-540-33700-8_18.

François Charton, Jordan S. Ellenberg, Adam Zsolt Wagner, and Geordie Williamson. PatternBoost: Constructions in mathematics with a little help from AI, 2024. URL https://arxiv.org/abs/2411.00566.

Sourav Chatterjee and S. R. S. Varadhan. The large deviation principle for the Erdős–Rényi random graph. *European Journal of Combinatorics*, 32(7):1000–1017, 2011. doi: 10.1016/j.ejc.2011.03.014.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. doi: 10.1109/CVPR.2019.00609.

Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, Marc Lackenby, Geordie Williamson, Demis Hassabis, and Pushmeet Kohli. Advancing mathematics by guiding human intuition with AI. *Nature*, 600(7887):70–74, 2021. doi: 10.1038/s41586-021-04086-x.

Paul Erdős and Miklós Simonovits. A limit theorem in graph theory. *Studia Scientiarum Mathematicarum Hungarica*, 1:51–57, 1966.

Paul Erdős and A. H. Stone. On the structure of linear graphs. *Bulletin of the American Mathematical Society*, 52:1087–1091, 1946.

Victor Falgas-Ravry and Emil R. Vaughan. Applications of the semi-definite method to the Turán density problem for 3-graphs. *Combinatorics, Probability and Computing*, 22(1):21–54, 2013. doi: 10.1017/S0963548312000508.

David C. Fisher. Lower bounds on the number of triangles in a graph. *Journal of Graph Theory*, 13 (4):505–512, 1989. doi: 10.1002/jgt.3190130411.

A. W. Goodman. On sets of acquaintances and strangers at any party. *The American Mathematical Monthly*, 66(9):778–783, 1959. doi: 10.1080/00029890.1959.11989408.

Yang-Hui He, Kyu-Hwan Lee, Thomas Oliver, and Alexey Pozdnyakov. Murmurations of elliptic curves. *Experimental Mathematics*, 34(3):528–540, 2025. doi: 10.1080/10586458.2024.2382361.

Yichen Huang and Lin F. Yang. Winning gold at IMO 2025 with a model-agnostic verification-and-refinement pipeline, 2025. URL https://arxiv.org/abs/2507.15855.

Thomas Hubert, Rishi Mehta, Laurent Sartran, Miklós Z. Horváth, Goran Žužić, Eric Wieser, Aja Huang, Julian Schrittwieser, Yannick Schroecker, Hussain Masoom, Ottavia Bertolli, Tom Zahavy, Amol Mandhane, Jessica Yung, Iuliya Beloshapka, Borja Ibarz, Vivek Veeriah, Lei Yu, Oliver Nash, Paul Lezeau, Salvatore Mercuri, Calle Sönne, Bhavik Mehta, Alex Davies, Daniel Zheng, Fabian Pedregosa, Yin Li, Ingrid von Glehn, Mark Rowland, Samuel Albanie, Ameya Velingker, Simon Schmitt, Edward Lockhart, Edward Hughes, Henryk Michalewski, Nicolas Sonnerat, Demis Hassabis, Pushmeet Kohli, and David Silver. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 651(8106):607–613, 2026. doi: 10.1038/s41586-025-09833-y.

Seonghyuk Im, Taeyoung Kim, and Hongseok Yang. Goodman-type lower bounds for odd cycles I: The dense regime, 2026. URL https://arxiv.org/abs/2610.18277.

Albert Qiaochu Jiang, Sean Welleck, Jin Peng Zhou, Timothée Lacroix, Jiacheng Liu, Wenda Li, Mateja Jamnik, Guillaume Lample, and Yuhuai Wu. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. In *Proceedings of the 11th International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SMa9EAovKMC.

Sangho Lim, Seonghyuk Im, Taeyoung Kim, Kyeongsik Nam, and Hongseok Yang. Bipodal optimizers in the upper-tail variational problem for regular subgraph densities, 2026. URL https://arxiv.org/abs/2609.15222.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *Proceedings of the 5th International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

László Lovász. *Large Networks and Graph Limits*, volume 60 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2012. ISBN 978-0-8218-9085-1. doi: 10.1090/coll/060.

László Lovász and Miklós Simonovits. On the number of complete subgraphs of a graph II. In *Studies in Pure Mathematics: To the Memory of Paul Turán*, pages 459–495. Birkhäuser, Basel, 1983. doi: 10.1007/978-3-0348-5438-2_41.

László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006. doi: 10.1016/j.jctb.2006.05.002.

Eyal Lubetzky and Yufei Zhao. On replica symmetry of large deviations in random graphs. *Random Structures & Algorithms*, 47(1):109–146, 2015. doi: 10.1002/rsa.20536.

Willem Mantel. Vraagstuk XXVIII. *Wiskundige Opgaven met de Oplossingen*, 10:60–61, 1907.

Abbas Mehrabian, Ankit Anand, Hyunjik Kim, Nicolas Sonnerat, Matej Balog, Gheorghe Comanici, Tudor Berariu, Andrew Lee, Anian Ruoss, Anna Bulanova, Daniel Toyama, Sam Blackwell, Bernardino Romera-Paredes, Petar Veličković, Laurent Orseau, Joonkyung Lee, Anurag Murty Naredla, Doina Precup, and Adam Zsolt Wagner. Finding increasingly large extremal graphs with AlphaZero and tabu search. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 6985–6993, 2024. doi: 10.24963/ijcai.2024/772.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. doi: 10.1109/CVPR.2019.00459.

Konrad Mundinger, Max Zimmer, Aldo Kiem, Christoph Spiegel, and Sebastian Pokutta. Neural discovery in mathematics: Do machines dream of colored planes? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 45236–45255. PMLR, 2025. URL https://proceedings.mlr.press/v267/mundinger25a.html.

Vladimir Nikiforov. The number of cliques in graphs of given order and size. *Transactions of the American Mathematical Society*, 363(3):1599–1618, 2011. doi: 10.1090/S0002-9947-2010-05189-X.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery, 2025. URL https://arxiv.org/abs/2506.13131.

Olaf Parczyk, Sebastian Pokutta, Christoph Spiegel, and Tibor Szabó. New Ramsey multiplicity bounds and search heuristics. *Foundations of Computational Mathematics*, 25(5):1777–1814, 2025. doi: 10.1007/s10208-024-09675-6.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. doi: 10.1109/CVPR.2019.00025.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018. doi: 10.1609/aaai.v32i1.11671.

Oleg Pikhurko and Alexander Razborov. Asymptotic structure of graphs with the minimum number of triangles. *Combinatorics, Probability and Computing*, 26(1):138–160, 2017. doi: 10.1017/S0963548316000110.

Stanislas Polu, Jesse Michael Han, Kunhao Zheng, Mantas Baksys, Igor Babuschkin, and Ilya Sutskever. Formal mathematics statement curriculum learning. In *Proceedings of the 11th International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=-P7G-8dmSh4.

Alexander A. Razborov. Flag algebras. *Journal of Symbolic Logic*, 72(4):1239–1282, 2007. doi: 10.2178/jsl/1203350785.

Alexander A. Razborov. On the minimal density of triangles in graphs. *Combinatorics, Probability and Computing*, 17(4):603–618, 2008. doi: 10.1017/S0963548308009085.

Alexander A. Razborov. Flag algebras: An interim report. In *The Mathematics of Paul Erdős II*, pages 207–232. Springer, New York, NY, 2013. doi: 10.1007/978-1-4614-7254-4_16.

Christian Reiher. The clique density theorem. *Annals of Mathematics*, 184(3):683–707, 2016. doi: 10.4007/annals.2016.184.3.1.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024. doi: 10.1038/s41586-023-06924-6.

Alexander Sidorenko. A correlation inequality for bipartite graphs. *Graphs and Combinatorics*, 9 (2–4):201–204, 1993. doi: 10.1007/BF02988307.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/55053683268957697aa39fba6f231c68-Abstract.html.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024. doi: 10.1038/s41586-023-06747-5.

Pál Turán. Eine Extremalaufgabe aus der Graphentheorie. *Matematikai és Fizikai Lapok*, 48:436–452, 1941.

Adam Zsolt Wagner. Constructions in combinatorics via neural networks, 2021. URL https://arxiv.org/abs/2104.14516.

Yuhuai Wu, Albert Qiaochu Jiang, Wenda Li, Markus Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. Autoformalization with large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/d0c6bc641a56bebee9d985b937307367-Abstract-Conference.html.

Xinyue Xia, Gal Mishne, and Yusu Wang. Implicit graphon neural representation. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 10619–10634. PMLR, 2023. URL https://proceedings.mlr.press/v206/xia23b.html.

A Related work

Implicit neural representations of graphons were introduced by Xia et al. [2023] for *graphon learning*, and Azizpour et al. [2025] proposed a scalable variant. In graphon learning, one observes a collection of graphs assumed to be sampled from a common unknown graphon and estimates this graphon from the samples. Recovering sharp details of the true graphon is not the focus of these methods, and such details are often lost in the learnt graphons; see, e.g., Azizpour et al. [2025, Figs. 3(b) and 9]. Our setting is different: there is neither a graph dataset nor a target graphon, and we search directly over graphons to minimise a homomorphism-density or relative-entropy functional under a density constraint. Applying graphon learning to these problems would first require near-optimal finite graphs, which is itself a hard discrete optimisation problem.

Razborov's *flag algebra* framework [Razborov, 2007] enables proofs of inequalities between graph homomorphism densities by means of sums of squares (ultimately, applications of the Cauchy–Schwarz inequality) in an algebra of graphs, where each finite graph can be interpreted as a functional on the space of graphons. This viewpoint also enables the automation of Cauchy–Schwarz-style arguments via semidefinite programming (SDP), and has been implemented in the FLAGMATIC package [Falgas-Ravry and Vaughan, 2013]. Flag algebras and Flagmatic have resolved many open problems in extremal combinatorics over the past decade (for a survey, see Razborov [2013]), but their computational cost can be prohibitive. [2] This limits applicability to some instances we consider, such as problems involving the Petersen graph. We view our method as *complementary* to flag algebras rather than a replacement: flag algebras provide rigorous *lower* bounds together with certificates, whereas our method searches directly for candidate graphons and yields candidate *upper* bounds and concrete candidate minimisers. When an upper bound from our method nearly matches a rigorous lower bound, this gives strong evidence for the correct optimum and produces a graphon whose structure can be inspected and turned into mathematical conjectures (as in our exhaustive sweep in §4).

For fixed values of n (i.e., without taking the limit in Eq. (4) or related formulations), one may instead attempt to solve the discrete optimisation problem directly. A classical approach is local (e.g., Tabu) search, which remains useful in combination with reinforcement learning; see, e.g., Parczyk et al. [2025], Mehrabian et al. [2024]. As a direct reinforcement-learning approach, Mehrabian et al. [2024] used ALPHAZERO to search for n-vertex graphs with as many edges as possible and no copies of K_3 and C_4. The so-called *deep cross entropy method* was used by Wagner [2021] for analysing discrete problems in combinatorics. It iteratively trains a neural network to learn a policy that promotes the generation of high-quality examples, which succeeded in disproving conjectures regarding the largest eigenvalue of the adjacency matrix and the matching number. Finally, PATTERNBOOST [Charton et al., 2024] takes a hybrid approach that combines local search with a learned generative model; one illustrative example is generating lower-bound examples related to Mantel's theorem by solving finite-n instances of Eq. (4) for $H = K_3$ (for small n) up to edge density $p \leq 1/2$. For discrete objectives, these approaches can be more versatile than our method because they can optimise non-differentiable quantities (e.g., the matching number). In contrast, they operate at a fixed graph size, whereas our method targets asymptotic problems in the dense-graph limit. Because many combinatorial objects are highly symmetric, solutions produced by such discrete approaches can be difficult to interpret or distil into human-understandable structure. In contrast, as demonstrated by visualisations of learnt graphons, our method often captures structure that is amenable to inspection and interpretation.

From the perspective of implicit neural representations, the work most closely related to ours is Mundinger et al. [2025], which applies implicit neural representations to the Hadwiger–Nelson problem: colouring the plane $\mathbb{R}^2$ with the minimum number of colours such that no two points at unit distance share a colour. Analogously to our approach, they represent a colouring by a neural network and recover a 6-colouring by minimising a monochromatic edge-density objective. With human post-processing, they construct two new 6-colourings that expand the known range of realisable distances from $[0.415, 0.447]$ to $[0.354, 0.657]$. In contrast, we focus on asymptotic extremal problems in graph theory that can be phrased as constrained optimisation over graphons. Moreover, our approach integrates problem structure into the learning procedure, including an architectural bias toward discontinuous, step-like solutions, explicit constraint enforcement via an embedded solver, and

[2] For example, to derive a nontrivial inequality for a graph G, FLAGMATIC typically enumerates all graphs H with $v(H) = k \geq v(G)$ and solves an SDP over vectors indexed by such graphs. Graphs H with more than eight vertices are currently unsupported due to computational complexity.

variance reduction in homomorphism density estimation, with the aim of supporting systematic conjecture generation and exploration across diverse extremal problems in graph theory.

B Societal impacts and limitations

Our work presents a method for finding structured candidate optima of graphon variational problems, which is of interest to mathematicians working on extremal graph theory and the large deviation theory of dense random graphs. Compared with existing baselines, our method produces candidate optima that are markedly more interpretable, offering insight into the structure of optimal graphons and thereby supporting mathematical conjecture generation and discovery. We do not foresee any major negative societal impacts arising from this work.

Our method has several limitations. First, it is tailored to graphon optimisation and therefore does not apply to combinatorial problems on sparse graphs, for which the graphon limit is degenerate. Second, although the candidate optima it produces are highly interpretable, the method itself does not provide rigorous guarantees of optimality; it should therefore be used in conjunction with rigorous lower-bound techniques such as flag algebras [Razborov, 2007, Falgas-Ravry and Vaughan, 2013] to verify optimality and provide certificates. Third, evaluating our method requires estimating homomorphism densities, for which we use either discretisation-based computation or Monte Carlo estimation; both can be computationally expensive for large motifs. Discretisation-based computation of $t(H, W)$ on an $M \times M$ grid becomes infeasible for motifs with many vertices, such as the Petersen graph; Monte Carlo estimation can suffer from large relative error when the motif density is very small.

C Background on graphons, odd cycles, and large deviations

In this section, we give an informal introduction to the graphon formulation of §2 and state the conjectured optima for odd cycles used in §4.

Graphons as limits of graphs. Extremal graph theory and the large deviation theory of random graphs ask questions about very large graphs, such as how few triangles a graph with a given edge density can have. Graphons replace a sequence of growing graphs by a single function on the unit square, and our method optimises over these functions.

A graph G with vertex set $[n]$ can be drawn as an image on the unit square through its adjacency matrix:

$$W_G(x,y) \coloneqq \begin{cases} 1 & \text{if } x \in \left(\frac{i-1}{n}, \frac{i}{n}\right], \ y \in \left(\frac{j-1}{n}, \frac{j}{n}\right], \text{ and } \{i,j\} \in E(G), \\ 0 & \text{otherwise.} \end{cases}$$

Note that isomorphic graphs give images that differ only by a relabelling, i.e., a permutation of rows and columns. The heatmaps in our figures show graphons in the same way.

Consider the limit of such images when the number of vertices grows. For the complete graph K_n, the image takes the value 1 except on the n diagonal squares, whose total area is $1/n$, so it converges to the constant function 1 as $n \to \infty$. For the Erdős–Rényi random graph $G(n,q)$, the image is noise with entries 0 and 1, which never approaches the constant function q pointwise. Nevertheless, the constant graphon $W \equiv q$ is the natural limit of $G(n,q)$. This limit is characterised by the so-called cut distance. For a symmetric measurable function $W : [0,1]^2 \to [0,1]$, we consider a measure-preserving bijection $\sigma : [0,1] \to [0,1]$, which is a continuous counterpart of a relabelling of vertices, and write $W^\sigma(x,y) \coloneqq W(\sigma(x), \sigma(y))$. Then the cut distance is the infimum of the difference between two graphons at a cut $S \times T$ over such transformed graphons:

$$\delta_\square(U,W) \coloneqq \inf_\sigma \sup_{S,T \subseteq [0,1]} \left| \int_{S \times T} \big(U(x,y) - W^\sigma(x,y)\big) \, \mathrm{d}x \, \mathrm{d}y \right|.$$

A sequence of graphs (G_n) converges to a graphon W if $\delta_\square(W_{G_n}, W) \to 0$. The space of graphons, considered up to measure-preserving transformations (more precisely, with graphons at cut distance zero identified), is compact in the cut distance, and the images W_G of graphs are dense in it [Lovász, 2012]. In particular, every sequence of graphs has a subsequence that converges to a graphon.

Homomorphism densities are compatible with this limit. For every graph H and graphons U and W, we have

$$|t(H,U) - t(H,W)| \ \le \ e(H) \, \delta_\square(U,W),$$

so the H-density of a large graph is close to that of any graphon near it in the cut distance. Conversely, for a sequence of graphs (G_n), if $t(H, W_{G_n}) \to t(H, W)$ for every graph H, then $W_{G_n} \to W$ in the cut distance [Lovász, 2012]. By the continuity of homomorphism densities and the compactness of the space of graphons, an optimisation problem over large graphs whose objective and constraints are continuous in the cut distance, such as homomorphism densities, becomes an optimisation problem over graphons: the asymptotic optimum over graph sequences equals the optimum over graphons, and the latter is attained. For example, the asymptotic minimum in (4) equals the minimum in **(P1)**. The optimum is also attained for objectives that are only lower semicontinuous in the cut distance, such as the relative entropy h_q in **(P2)** [Chatterjee and Varadhan, 2011].

Conjectured optima for odd cycles. In Bennett et al. [2020], the conjectured optimal graphon is constructed as follows. Let k be the integer satisfying $1 - \frac{1}{k} < p < 1 - \frac{1}{k+1}$. First, we partition the interval $[0, 1]$ into $(k + 1)$ sets, $X_1, \dots, X_{k-1}, Y_1, Y_2$. We set the lengths of $X_1, \dots, X_{k-1}$ to be x, and the lengths of Y_1 and Y_2 to be $\frac{y}{2}$, where $y = 1 - (k - 1)x$. We then define the graphon by

$$W(x_1, x_2) = \begin{cases} 1 & \text{if } x_1 \in X_m,\ x_2 \in X_n,\ m \neq n \\ 1 & \text{if } x_1 \in X_m,\ x_2 \in Y_n, \text{ or vice versa} \\ 0 & \text{if } x_1 \in X_m,\ x_2 \in X_m \\ \rho & \text{if } x_1 \in Y_1,\ x_2 \in Y_2 \text{ or vice versa} \\ 0 & \text{if } x_1 \in Y_m,\ x_2 \in Y_m \end{cases}$$

where ρ is chosen to satisfy the edge-density constraint.

One can compute the C_5 density of this graphon, which can be written as a polynomial $f_5(x, y, \rho)$ given by

$$\begin{aligned} & f_5(x, y, \rho) \\ &= \left(\frac{(k-1)_5}{10} + \frac{(k-1)_4}{2} + \frac{(k-1)_3}{2} \right) x^5 \\ &+ \left(\frac{(k-1)_4}{2} + \frac{3(k-1)_3}{2} + \frac{(k-1)_2}{2} \right) x^4 y \\ &+ \left(\left(\frac{1}{2} + \frac{\rho}{2} \right) (k-1)_3 + \left(1 + \frac{\rho}{2} \right) (k-1)_2 \right) x^3 y^2 \\ &+ \left(\left(\frac{\rho}{2} + \frac{\rho^2}{2} \right) (k-1)_2 + \frac{\rho}{2}(k-1) \right) x^2 y^3 \\ &+ \frac{\rho^3}{2}(k-1) x y^4, \end{aligned}$$

where $(k-1)_r$ is a falling factorial, $(k-1)_r = (k-1) \cdot (k-2) \cdots (k-r)$. With these definitions, the conjectured optimal value is obtained by optimising over x and ρ subject to the constraints below (with y determined by x via $(k-1)x + y = 1$). The contribution of the Y_1–Y_2 rectangles to the ordered edge density is $\rho y^2/2$, since the two off-diagonal rectangles have total area $y^2/2$. Equivalently, one can compute the conjectured optimum by solving

$$\begin{aligned} & \min f_5(x, y, \rho) \\ \text{subject to } & (k-1)x + y = 1, \\ & (k-1)_2 x^2 + 2(k-1)xy + \frac{\rho y^2}{2} = p, \\ & x, y \geq 0, \quad 0 \leq \rho \leq 1. \end{aligned}$$

Using the same structure, we can generalise the conjecture to arbitrary odd cycles $C_{2\ell+1}$ by replacing f_5 with the corresponding $f_{2\ell+1}$. For instance, f_7 is given as follows.

$$
\begin{aligned}
&f_7(x,y,\rho)\\
&= \left(\frac{(k-1)_7}{14} + (k-1)_6 + 4(k-1)_5 + 5(k-1)_4 + \frac{3(k-1)_3}{2}\right) x^7\\
&+ \left(\frac{(k-1)_6}{2} + 5(k-1)_5 + \frac{25(k-1)_4}{2} + \frac{15(k-1)_3}{2} + \frac{(k-1)_2}{2}\right) x^6 y\\
&+ \left((k-1)_5 + (3\rho+7)(k-1)_4 + \left(\frac{7\rho}{2}+10\right)(k-1)_3 + \left(\frac{\rho}{2}+2\right)(k-1)_2\right) x^5 y^2\\
&+ \left(\left(\frac{3\rho}{2}+\frac{1}{2}\right)(k-1)_4 + \left(\frac{3\rho^2}{2}+6\rho+\frac{5}{2}\right)(k-1)_3 + \left(\frac{\rho^2}{2}+3\rho+2\right)(k-1)_2\right) x^4 y^3\\
&+ \left(\left(\frac{\rho^3}{2}+\frac{3\rho^2}{2}+\frac{\rho}{2}\right)(k-1)_3 + \left(\frac{\rho^3}{2}+3\rho^2+\frac{3\rho}{2}\right)(k-1)_2 + \frac{\rho}{2}(k-1)\right) x^3 y^4\\
&+ \left(\left(\frac{\rho^4}{2}+\rho^3\right)(k-1)_2 + \rho^3(k-1)\right) x^2 y^5\\
&+ \frac{\rho^5}{2}(k-1) x y^6.
\end{aligned}
$$

Large deviations and the relative entropy. For $a, q \in (0,1)$, write

$$\mathrm{KL}(a \,\|\, q) \coloneqq \mathrm{KL}\big(\mathrm{Bern}(a) \,\|\, \mathrm{Bern}(q)\big) = a \log \frac{a}{q} + (1-a) \log \frac{1-a}{1-q}.$$

In $G(n,q)$, for each of the $N = \binom{n}{2}$ vertex pairs, the two vertices are adjacent independently with probability q. If aN is an integer, Stirling's formula gives $\log \binom{N}{aN} = -N\big(a \log a + (1-a)\log(1-a)\big) + o(N)$. Hence,

$$
\begin{aligned}
\log \Pr_{G \sim G(n,q)}\big[e(G) = aN\big] &= \log\left(\binom{N}{aN} q^{aN} (1-q)^{(1-a)N}\right)\\
&= -N\left(a \log \frac{a}{q} + (1-a)\log\frac{1-a}{1-q}\right) + o(N)\\
&= -N\,\mathrm{KL}(a \,\|\, q) + o(N).
\end{aligned}
$$

Thus the probability that $G(n,q)$ has edge density a rather than the typical value q decays exponentially in N, at rate $\mathrm{KL}(a \,\|\, q)$. The large deviation principle of Chatterjee and Varadhan [2011] extends this from a single edge density to the whole profile of edge probabilities: informally, the probability that $G(n,q)$ is close in cut distance to a graphon W is $\exp(-N h_q(W) + o(n^2))$, where $h_q(W) = \int_{[0,1]^2} \mathrm{KL}(W(x,y) \,\|\, q)\,\mathrm{d}x\,\mathrm{d}y$ as in §2. The probability of an event such as $t(H, G(n,q)) \geq r^{e(H)}$ is dominated by the graphons that satisfy this constraint with the smallest relative entropy, which gives the variational formulas in §2, and conditioned on the event, $G(n,q)$ is close to these minimisers [Lubetzky and Zhao, 2015].

Optimal solutions of large deviation problems. Consider the upper-tail problem in **(P2)**, where $q < r$:

$$\min_W h_q(W) \quad \text{s.t.} \quad t(H,W) = r^{e(H)}.$$

The constant graphon $W \equiv r$ satisfies the constraint $t(H,W) = r^{e(H)}$, so it is a feasible solution, with objective value $h_q(W) = \mathrm{KL}(r \,\|\, q)$. For a d-regular graph H with $d \geq 2$, Lubetzky and Zhao [2015] characterised when this constant graphon is optimal: there is a curve $q_c(r)$ such that $W \equiv r$ is the unique optimum of **(P2)** if $q > q_c(r)$, and some graphon has a smaller objective value than $W \equiv r$ if $q < q_c(r)$. The region where the constant graphon $W \equiv r$ is optimal is called the *replica-symmetric phase*, and the region where it is not optimal is called the *symmetry-breaking phase*.

D Proven results on the learnt structures

The graphons found by our method motivated the following three theorems. We state them in the notation of this paper and explain how each relates to our experiments. Recall the balanced complete k-partite graphon W_k from §4.

Theorem D.1 (Im et al., 2026)**.** *Let $m \geq 3$ be odd, and let W be a graphon with $t(K_2, W) = p \geq 2/3$. Then*

$$t(C_m, W) \geq p^m - p(1-p)^{m-1}.$$

In particular, when $p = 1 - 1/k$ for some integer $k \geq 3$, the balanced complete k-partite graphon W_k minimises the density of every odd cycle.

The second statement holds because $t(C_m, W_k) = p^m - p(1-p)^{m-1}$ at $p = 1 - 1/k$. Hence the learnt C_7 graphons at $p \in \{2/3, 4/5\}$ (top row of Fig. 4), which recover the complete k-partite structure, are optimal.

A graph is *chordal* if no cycle of length greater than 3 appears as an induced subgraph of it.

Theorem D.2. *Let H be a connected chordal graph whose maximal cliques all have the same size $r \geq 3$. Then, for every integer $k \geq r$, the graphon W_k minimises $t(H, W)$ subject to $t(K_2, W) = 1 - 1/k$.*

A proof of Theorem D.2, formalised in Lean 4, is available at `https://github.com/mekty2012/PureChordalGoodman`. Theorem D.2 applies to 16 graphs of our exhaustive study in §4 other than cliques, and settles 48 of its instances.

A graphon W is *bipodal* if there is a measurable set $A \subseteq [0,1]$ such that, with $B = [0,1] \setminus A$, the graphon W is almost everywhere constant on each of $A \times A$, $(A \times B) \cup (B \times A)$, and $B \times B$. Recall the curve $q_c(r)$ of Lubetzky and Zhao [2015] from §C.

Theorem D.3 (Lim et al., 2026, Theorem 1.5)**.** *Let H be a d-regular graph with $d \geq 2$, and let $r_0 \in (0,1) \setminus \{(d-1)/d\}$. There is an open neighbourhood U of $(q_c(r_0), r_0)$ such that, for every $(q, r) \in U$ with $q < q_c(r)$, the minimiser of $h_q(W)$ subject to $t(H, W) \geq r^{e(H)}$ is unique up to relabelling, nonconstant, and bipodal.*

This partially resolves the conjecture that the upper-tail large deviation problem for regular graphs has bipodal graphons as optimal solutions. The learnt upper-tail graphons for the 2-regular graphs K_3, C_4, and C_5 have this two-block form (Fig. 6 and §G).

E Properties of the density estimator and the implicit gradient

Lemma E.1. *The estimator in Eq.* (8) *is unbiased, i.e., $\mathbb{E}[\widehat{t}(H, f_\theta[c])] = t(H, f_\theta[c])$.*

Proof. We compute the claimed equality as follows:

$$\begin{aligned}
\mathbb{E}\left[\widehat{t}(H, f_\theta[c])\right] &= \mathbb{E}\left[\frac{1}{N|\mathcal{S}|} \sum_{k\in[N],\pi\in\mathcal{S}} \left(\prod_{\{i,j\}\in E(H)} f_\theta[c]\left(x^{(k)}_{\pi(i)}, x^{(k)}_{\pi(j)}\right)\right)\right] \\
&= \frac{1}{N|\mathcal{S}|} \sum_{k\in[N],\pi\in\mathcal{S}} \mathbb{E}\left[\prod_{\{i,j\}\in E(H)} f_\theta[c]\left(x^{(k)}_{\pi(i)}, x^{(k)}_{\pi(j)}\right)\right] \\
&= \frac{1}{N|\mathcal{S}|} \sum_{k\in[N],\pi\in\mathcal{S}} \mathbb{E}_{x_1,\ldots,x_{v(H)} \overset{\text{iid}}{\sim} \text{Unif}[0,1]}\left[\prod_{\{i,j\}\in E(H)} f_\theta[c](x_i, x_j)\right] \\
&= \frac{1}{N|\mathcal{S}|} \sum_{k\in[N],\pi\in\mathcal{S}} t(H, f_\theta[c]) = t(H, f_\theta[c]).
\end{aligned}$$

□

Lemma E.2 (Restatement of Lemma 3.1)**.** *For all* $\mathbf{x} \in [0,1]^2$*,*

$$\nabla_\theta\, f_\theta[c_\theta](\mathbf{x}) \;=\; \frac{\partial f_\theta[c](\mathbf{x})}{\partial \theta}\Big|_{c=c_\theta} \;-\; sm'\big(h_\theta(\mathbf{x}) + c_\theta\big) \sum_{\mathbf{y}\in\mathcal{E}} w_\theta(\mathbf{y})\, \nabla_\theta h_\theta(\mathbf{y}),$$

$$\textit{where } w_\theta(\mathbf{y}) \;\coloneqq\; \frac{sm'(h_\theta(\mathbf{y}) + c_\theta)}{\sum_{\mathbf{z}\in\mathcal{E}} sm'(h_\theta(\mathbf{z}) + c_\theta)}.$$

Proof. Let $\mathbf{x} \in [0,1]^2$. Note that

$$\begin{aligned}\nabla_\theta\; f_\theta[c_\theta](\mathbf{x}) &= \frac{\partial f_\theta[c](\mathbf{x})}{\partial \theta}\Big|_{c=c_\theta} + \frac{\partial f_\theta[c](\mathbf{x})}{\partial c}\Big|_{c=c_\theta} \nabla_\theta c_\theta \\ &= \frac{\partial f_\theta[c](\mathbf{x})}{\partial \theta}\Big|_{c=c_\theta} + sm'(h_\theta(\mathbf{x}) + c_\theta)\nabla_\theta c_\theta\end{aligned}$$

Thus, it suffices to show that

$$\nabla_\theta c_\theta = -\sum_{\mathbf{y}\in\mathcal{E}} \frac{sm'(h_\theta(\mathbf{y}) + c_\theta)}{\sum_{\mathbf{z}\in\mathcal{E}} sm'(h_\theta(\mathbf{z}) + c_\theta)} \nabla_\theta\, h_\theta(\mathbf{y}).$$

We prove the above equality by applying implicit differentiation to the following equation about c_θ, which holds by the choice of c_θ:

$$\frac{1}{|\mathcal{E}|} \sum_{\mathbf{y}\in\mathcal{E}} f_\theta[c_\theta](\mathbf{y}) = p.$$

First, we multiply both sides of the equation by $|\mathcal{E}|$, and differentiate them with respect to θ:

$$\left(\sum_{\mathbf{y}\in\mathcal{E}} \frac{\partial f_\theta[c](\mathbf{y})}{\partial \theta}\Big|_{c=c_\theta}\right) + \left(\sum_{\mathbf{z}\in\mathcal{E}} \frac{\partial f_\theta[c](\mathbf{z})}{\partial c}\Big|_{c=c_\theta} \nabla_\theta c_\theta\right) = 0$$

Thus,

$$\begin{aligned}\nabla_\theta c_\theta &= -\left(\sum_{\mathbf{y}\in\mathcal{E}} \frac{\partial f_\theta[c](\mathbf{y})}{\partial \theta}\Big|_{c=c_\theta}\right) \Bigg/ \left(\sum_{\mathbf{z}\in\mathcal{E}} \frac{\partial f_\theta[c](\mathbf{z})}{\partial c}\Big|_{c=c_\theta}\right) \\ &= -\sum_{\mathbf{y}\in\mathcal{E}} \frac{sm'(h_\theta(\mathbf{y}) + c_\theta)}{\sum_{\mathbf{z}\in\mathcal{E}} sm'(h_\theta(\mathbf{z}) + c_\theta)} \nabla_\theta\, h_\theta(\mathbf{y}),\end{aligned}$$

as desired. □

The gradient in Lemma 3.1 is the exact gradient for the empirical constraint (7). Since $\nabla_\theta c_\theta$ is a ratio of two sample averages, it is in general a biased estimator of the gradient of the offset that enforces the population constraint $t(K_2, f_\theta[c]) = p$. It is, however, a consistent estimator, i.e., it converges to this gradient in the limit of infinitely many samples.

Lemma E.3 (Consistency of the implicit gradient)**.** *Assume that* $h_\theta(\mathbf{y})$ *and* $\nabla_\theta h_\theta(\mathbf{y})$ *are continuous in* $(\theta, \mathbf{y})$*. Fix* θ *and* $p \in (0,1)$*, and let* c^*_θ *be the unique* $c \in \mathbb{R}$ *with* $t(K_2, f_\theta[c]) = p$*. Let* $\mathbf{y}_1, \mathbf{y}_2, \ldots \overset{\text{iid}}{\sim} \text{Unif}([0,1]^2)$*, and for* $K \geq 1$ *let* $c_{\theta,K} \coloneqq \textsc{Solve}_\theta(\{\mathbf{y}_1, \ldots, \mathbf{y}_K\}, p)$*. Then, almost surely,* $c_{\theta,K} \to c^*_\theta$ *and*

$$\begin{aligned}\nabla_\theta c_{\theta,K} &= -\frac{\sum_{k=1}^K sm'\big(h_\theta(\mathbf{y}_k) + c_{\theta,K}\big)\, \nabla_\theta h_\theta(\mathbf{y}_k)}{\sum_{k=1}^K sm'\big(h_\theta(\mathbf{y}_k) + c_{\theta,K}\big)} \\ \longrightarrow\; \nabla_\theta c^*_\theta \;&=\; -\frac{\mathbb{E}\big[sm'\big(h_\theta(\mathbf{y}) + c^*_\theta\big)\, \nabla_\theta h_\theta(\mathbf{y})\big]}{\mathbb{E}\big[sm'\big(h_\theta(\mathbf{y}) + c^*_\theta\big)\big]}\end{aligned}$$

as $K \to \infty$*, where* $\mathbf{y} \sim \text{Unif}([0,1]^2)$*.*

Proof. Let $M \coloneqq \max_{\mathbf{y}\in[0,1]^2} |h_\theta(\mathbf{y})|$, which is finite because h_θ is continuous on the compact set $[0,1]^2$, and let $\ell \coloneqq sm^{-1}(p)$. For $c \in \mathbb{R}$, define

$$F_K(c) \coloneqq \frac{1}{K}\sum_{k=1}^{K} sm\big(h_\theta(\mathbf{y}_k)+c\big), \qquad F(c) \coloneqq \mathbb{E}\big[sm\big(h_\theta(\mathbf{y})+c\big)\big] = t(K_2, f_\theta[c]).$$

Both functions are continuous and strictly increasing in c. If $c < \ell - M$, then $sm(h_\theta(\mathbf{y})+c) < sm(\ell) = p$ for every $\mathbf{y}$, and if $c > \ell + M$, then $sm(h_\theta(\mathbf{y})+c) > p$ for every $\mathbf{y}$. Hence $c_{\theta,K}$ and c^*_θ lie in $[\ell - M, \ell + M]$.

Convergence of the offsets. Let $\varepsilon > 0$. Since F is strictly increasing, $F(c^*_\theta - \varepsilon) < p < F(c^*_\theta + \varepsilon)$. By the strong law of large numbers, $F_K(c^*_\theta \pm \varepsilon) \to F(c^*_\theta \pm \varepsilon)$ almost surely, so almost surely $F_K(c^*_\theta - \varepsilon) < p < F_K(c^*_\theta + \varepsilon)$ for all sufficiently large K, and then $|c_{\theta,K} - c^*_\theta| < \varepsilon$ because F_K is increasing. Taking $\varepsilon = 1/m$ for every $m \in \mathbb{N}$ gives $c_{\theta,K} \to c^*_\theta$ almost surely.

Convergence of the gradients. The formula for $\nabla_\theta c_{\theta,K}$ is the one derived in the proof of Lemma 3.1. Let $g : [0,1]^2 \to \mathbb{R}^{m'}$ be bounded and measurable. Since $|sm''| \le 1$, the function sm' is 1-Lipschitz, so

$$\begin{aligned}
&\Big\| \frac{1}{K}\sum_{k=1}^{K} sm'\big(h_\theta(\mathbf{y}_k)+c_{\theta,K}\big)\, g(\mathbf{y}_k) - \mathbb{E}\big[sm'\big(h_\theta(\mathbf{y})+c^*_\theta\big)\, g(\mathbf{y})\big] \Big\| \\
&\qquad \le \|g\|_\infty\, |c_{\theta,K} - c^*_\theta| + \Big\| \frac{1}{K}\sum_{k=1}^{K} sm'\big(h_\theta(\mathbf{y}_k)+c^*_\theta\big)\, g(\mathbf{y}_k) - \mathbb{E}\big[sm'\big(h_\theta(\mathbf{y})+c^*_\theta\big)\, g(\mathbf{y})\big] \Big\|.
\end{aligned}$$

The first term tends to 0 almost surely by the convergence of the offsets, and the second by the strong law of large numbers. We apply this with $g \equiv 1$ and with $g = \nabla_\theta h_\theta$, which is bounded because it is continuous on $[0,1]^2$. The limit of the denominator is positive: since $h_\theta(\mathbf{y}) + c^*_\theta \in [\ell - 2M, \ell + 2M]$ and $sm'(z)$ decreases in $|z|$,

$$\mathbb{E}\big[sm'\big(h_\theta(\mathbf{y})+c^*_\theta\big)\big] \ge \min\big\{sm'(\ell-2M),\, sm'(\ell+2M)\big\} > 0.$$

Hence the ratio converges almost surely to $-\mathbb{E}[sm'(h_\theta(\mathbf{y})+c^*_\theta)\,\nabla_\theta h_\theta(\mathbf{y})]/\mathbb{E}[sm'(h_\theta(\mathbf{y})+c^*_\theta)]$.

Finally, this limit equals $\nabla_\theta c^*_\theta$. The map $(\theta, c) \mapsto \mathbb{E}[sm(h_\theta(\mathbf{y})+c)]$ is continuously differentiable, since its partial derivatives are expectations of functions continuous in $(\theta, c, \mathbf{y})$ over a compact domain, and its derivative in c is the positive denominator above. The implicit function theorem applied to $\mathbb{E}[sm(h_\theta(\mathbf{y})+c^*_\theta)] = p$ then gives the stated formula for $\nabla_\theta c^*_\theta$. □

When we extend our constraint solver beyond edge-density constraints to H-density constraints, we sample $x_i^{(k)}$ for $k \in [N]$ and $i \in [v(H)]$ independently from $\mathrm{Unif}[0,1]$, and estimate the H-density constraint by the Monte Carlo estimator

$$\widehat{t}(H, f_\theta[c]) = \frac{1}{N|\mathcal{S}|}\sum_{k\in[N]}\sum_{\pi\in\mathcal{S}}\prod_{\{i,j\}\in E(H)} f_\theta[c]\Big(x^{(k)}_{\pi(i)}, x^{(k)}_{\pi(j)}\Big),$$

where $\mathcal{S}$ is a set of permutations of $V(H)$.

Lemma E.4 (Implicit differentiation for the H-density estimator)**.** *Fix the samples $\{x_i^{(k)} : k \in [N],\ i \in [v(H)]\}$ and permutation set $\mathcal{S}$. Suppose c_θ is chosen so that $\widehat{t}(H, f_\theta[c_\theta]) = p$, and that $\partial_c \widehat{t}(H, f_\theta[c])\big|_{c=c_\theta} \neq 0$. Then for all $x \in [0,1]^2$,*

$$\nabla_\theta f_\theta[c_\theta](x) = \frac{\partial f_\theta[c](x)}{\partial\theta}\Big|_{c=c_\theta} - \frac{\partial f_\theta[c](x)}{\partial c}\Big|_{c=c_\theta} \cdot \frac{\nabla_\theta \widehat{t}(H, f_\theta[c])\big|_{c=c_\theta}}{\partial_c \widehat{t}(H, f_\theta[c])\big|_{c=c_\theta}},$$

where

$$\begin{aligned}
&\nabla_\theta \widehat{t}(H, f_\theta[c]) \\
=&\frac{1}{N|\mathcal{S}|} \sum_{k\in[N],\, \pi\in\mathcal{S}} \sum_{\{i,j\}\in E(H)} \left(\prod_{\substack{\{i',j'\}\in E(H) \\ \{i',j'\}\neq\{i,j\}}} f_\theta[c]\big(x^{(k)}_{\pi(i')}, x^{(k)}_{\pi(j')}\big) \right) \nabla_\theta f_\theta[c](x^{(k)}_{\pi(i)}, x^{(k)}_{\pi(j)}), \\
&\partial_c \widehat{t}(H, f_\theta[c]) \\
=&\frac{1}{N|\mathcal{S}|} \sum_{k\in[N],\, \pi\in\mathcal{S}} \sum_{\{i,j\}\in E(H)} \left(\prod_{\substack{\{i',j'\}\in E(H) \\ \{i',j'\}\neq\{i,j\}}} f_\theta[c]\big(x^{(k)}_{\pi(i')}, x^{(k)}_{\pi(j')}\big) \right) \frac{\partial f_\theta[c](x^{(k)}_{\pi(i)}, x^{(k)}_{\pi(j)})}{\partial c}.
\end{aligned}$$

Proof. Let $x \in [0,1]^2$. By the chain rule,

$$\nabla_\theta\, f_\theta[c_\theta](x) = \frac{\partial f_\theta[c](x)}{\partial \theta}\Big|_{c=c_\theta} + \frac{\partial f_\theta[c](x)}{\partial c}\Big|_{c=c_\theta} \nabla_\theta c_\theta.$$

Thus, it suffices to derive an expression for $\nabla_\theta c_\theta$.

By assumption, c_θ is chosen so that $\widehat{t}(H, f_\theta[c_\theta]) = p$. Differentiating both sides with respect to θ yields

$$\nabla_\theta\, \widehat{t}(H, f_\theta[c_\theta]) = \nabla_\theta p = 0.$$

Viewing $\widehat{t}(H, f_\theta[c])$ as a function of (θ, c), another application of the chain rule gives

$$0 = \nabla_\theta\, \widehat{t}(H, f_\theta[c])\Big|_{c=c_\theta} + \partial_c \widehat{t}(H, f_\theta[c])\Big|_{c=c_\theta} \nabla_\theta c_\theta.$$

Since $\partial_c \widehat{t}(H, f_\theta[c])\big|_{c=c_\theta} \neq 0$, we can solve for

$$\nabla_\theta c_\theta = -\frac{\nabla_\theta \widehat{t}(H, f_\theta[c])\big|_{c=c_\theta}}{\partial_c \widehat{t}(H, f_\theta[c])\big|_{c=c_\theta}}.$$

Substituting this into the first display gives the claimed expression for $\nabla_\theta f_\theta[c_\theta](x)$.

Finally, we derive the displayed expressions for $\nabla_\theta \widehat{t}(H, f_\theta[c])$ and $\partial_c \widehat{t}(H, f_\theta[c])$. For fixed $k \in [N]$ and $\pi \in \mathcal{S}$, write

$$P_{k,\pi}(c) \coloneqq \prod_{\{i,j\}\in E(H)} f_\theta[c]\big(x^{(k)}_{\pi(i)}, x^{(k)}_{\pi(j)}\big).$$

Then $\widehat{t}(H, f_\theta[c]) = \frac{1}{N|\mathcal{S}|} \sum_{k,\pi} P_{k,\pi}(c)$, and hence

$$\begin{aligned}
&\nabla_\theta \widehat{t}(H, f_\theta[c]) \\
=&\frac{1}{N|\mathcal{S}|} \sum_{k\in[N],\, \pi\in\mathcal{S}} \nabla_\theta P_{k,\pi}(c) \\
=&\frac{1}{N|\mathcal{S}|} \sum_{k\in[N],\, \pi\in\mathcal{S}} \sum_{\{i,j\}\in E(H)} \left(\prod_{\substack{\{i',j'\}\in E(H) \\ \{i',j'\}\neq\{i,j\}}} f_\theta[c]\big(x^{(k)}_{\pi(i')}, x^{(k)}_{\pi(j')}\big) \right) \nabla_\theta f_\theta[c]\big(x^{(k)}_{\pi(i)}, x^{(k)}_{\pi(j)}\big),
\end{aligned}$$

where the second line is the product rule. The expression for $\partial_c \widehat{t}(H, f_\theta[c])$ is identical, replacing $\nabla_\theta f_\theta[c](\cdot)$ by $\partial_c f_\theta[c](\cdot)$. □

F Details of experiments

All experiments were conducted on a single NVIDIA RTX A5000 GPU. The execution time depends on the size of the graph, from 30 minutes for a triangle to a few hours for the Petersen graph.

In all experiments, we used a network of width $d = 256$ and depth $L = 6$, which was sufficient to represent sharp boundaries and structural properties. The activation function ϕ was either $\sin$ or the GELU nonlinearity; we did not observe a consistent advantage of one over the other. In all experiments, we used the increasing input-encoding scale $s(\ell) = 2^{\ell-1}$. The constant-scale schedules in the ablation study are described in §I.

The set $\mathcal{S}$ in Eq. (8) was chosen according to the graph. For example, for complete graphs we used $\mathcal{S} = \{\mathrm{id}\}$, since any permutation yields the same summand. For cycles, we used a set of representatives for the cosets $S_{V(C_k)}/\mathrm{Aut}(C_k)$, where $\mathrm{Aut}(C_k)$ is the automorphism group of C_k, which gives $|\mathcal{S}| = (k-1)!/2$. For the other graphs, we used a set $\mathcal{S} = \{\pi^j : j = 0, 1, \ldots, m-1\}$ generated by a permutation $\pi \in S_{V(H)}$, where π and m are chosen so that the set

$$\{\{\pi^j(v), \pi^j(w)\} : \{v, w\} \in E(H) \text{ and } j = 0, 1, \ldots, m-1\}$$

equals the set of all unordered pairs of distinct vertices $\{\{v, w\} : v, w \in V(H) \text{ and } v \neq w\}$.

For training, we used the largest number of samples that fit in GPU memory, ranging from 2^{12} to 2^{16}. For large graphs such as the Petersen graph, we repeated multiple forward passes to accumulate gradients, yielding an effective sample size of 2^{16}. For evaluation, we used 2^{28} evaluation samples. With caching of pairwise network evaluations, one training iteration takes $O(Nv(H)^2Ld^2 + N|\mathcal{S}|e(H))$ time and $O(Nv(H)^2Ld)$ memory. Larger motifs therefore reduce the feasible batch size under a fixed memory budget. The Monte Carlo estimator retains the usual $O(N^{-1/2})$ absolute-error rate for bounded integrands, but relative error becomes much harder to control when the true density is extremely small.

For optimisation, we used the Adam optimiser with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We repeated the experiments with learning rates $\eta \in \{10^{-3}, 5\times10^{-4}, 10^{-4}, 5\times10^{-5}\}$, and set the minimum learning rate in the scheduler to 1% of the initial learning rate. We trained for 20,000 epochs, corresponding to five cosine-annealing cycles.

For the constraint solver, we found that applying Newton's method directly (without modification) can be unstable, owing to vanishing derivatives in many regimes. In particular, since many optimal graphons take values close to 0 or 1, the sigmoid derivative $sm(x)\,(1 - sm(x))$ is close to zero over much of the domain. Consequently, the denominator in the Newton update can become very small, leading to excessively large updates and unstable convergence. To address this, we use a clipped variant of Newton's method. To set the clipping range, we compute upper and lower bounds on the solution c^* and use their difference as the initial clipping range. We then gradually decrease the clipping range as the iteration proceeds, which improves stability. The detailed algorithms are described in Algorithm 1 (edge-density constraint) and Algorithm 2 (homomorphism-density constraint).

Algorithm 1 Edge-Density Constraint Solver

Require: $p \in (0, 1), y \in \mathbb{R}^N$
Ensure: $\frac{1}{N}\sum_{i=1}^{N} sm(y_i + c^*) \approxeq p$
  $c_1 \leftarrow sm^{-1}(p) - \frac{1}{N}\sum_{i=1}^{N} y_i$ {Initialise with guess}
  $D \leftarrow \max_i\{y_i\} - \min_i\{y_i\}$ {Compute range of y values for clipping}
  for $t = 1, 2, \ldots, T_{\max}$ **do**
    $f(c_t) \leftarrow \frac{1}{N}\sum_{i=1}^{N} sm(y_i + c_t) - p$
    if $|f(c_t)| < \epsilon$ **then**
      break
    end if
    $f'(c_t) \leftarrow \frac{1}{N}\sum_{i=1}^{N} sm(y_i + c_t)(1 - sm(y_i + c_t))$
    $c_{t+1} \leftarrow c_t - \mathrm{Clip}_{[-D/t, D/t]}\left(\frac{f(c_t)}{f'(c_t)}\right)$ {Clipping range D/t shrinks with the iteration index t.}
  end for
  Return c_t

For the variants in the ablation study, we performed hyperparameter searches for each variant. When using constant per-layer input encoding scale s, we searched over $s(\ell) \in \{1, 2, 4, 8\}$. For the strength of the regulariser, we searched over $\lambda \in \{0.01, 0.1, 1, 10, 100\}$. For the constant learning rate, we used two values, 10^{-4} (low) and 10^{-3} (high).

Algorithm 2 Homomorphism-Density Constraint Solver

Require: $p \in (0,1)$, $y \in \mathbb{R}^{B \times m}$
Ensure: $\frac{1}{B}\sum_{b=1}^{B}\prod_{j=1}^{m} sm(y_{b,j}+c^*) \cong p$
  $c_1 \leftarrow sm^{-1}(p^{1/m}) - \frac{1}{Bm}\sum_{b=1}^{B}\sum_{j=1}^{m} y_{b,j}$ {Initialise with guess}
  $D \leftarrow \max_{b,j}\{y_{b,j}\} - \min_{b,j}\{y_{b,j}\}$ {Compute range of y values for clipping}
  for $t = 1, 2, \ldots, T_{\max}$ **do**
    $S \leftarrow \big(sm(y_{b,j}+c_t)\big)_{b\in[B],\, j\in[m]}$
    $P \leftarrow \Big(\prod_{j=1}^{m} S_{b,j}\Big)_{b\in[B]}$
    $f(c_t) \leftarrow \frac{1}{B}\sum_{b=1}^{B} P_b - p$
    if $|f(c_t)| < \epsilon$ **then**
      break
    end if
    $f'(c_t) \leftarrow \frac{1}{B}\sum_{b=1}^{B} P_b \sum_{j=1}^{m}(1 - S_{b,j})$
    $c_{t+1} \leftarrow c_t - \mathrm{Clip}_{[-D/t,\, D/t]}\left(\frac{f(c_t)}{f'(c_t)}\right)$
  end for
  Return c_t

Table 7: Homomorphism densities of the Petersen graph in graphons under the edge-density constraint $t(K_2, W) = p$, found by our method and by a hand-crafted construction.

p	Ours (Estimated)	Hand-crafted
$1/2$	$(0.00000000, 0.00000000)$	0
$2/3$	$(0.00184293, 0.00184619)$	0.0020322
$4/5$	$(0.03412074, 0.03413641)$	0.0340869
$4/7$	$(0.00006195, 0.00006209)$	0.0001736
$5/8$	$(0.00054270, 0.00054310)$	0.0007531
$5/7$	$(0.00571939, 0.00572465)$	0.0059452

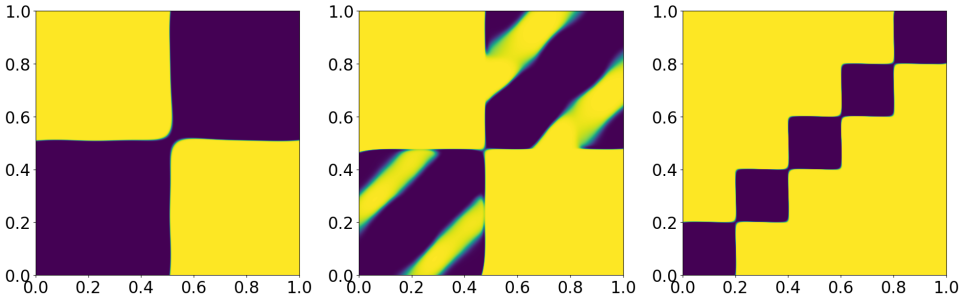
When reporting the homomorphism density computed by discretisation, we use a $(2^{14} \times 2^{14})$ grid for K_3, C_5, C_7 and a $(2^{10} \times 2^{10})$ grid for H_6, and 64-bit floating point precision.

The code to replicate our experiments is available at `https://github.com/mekty2012/INR4GraphonVariationalProblem`.

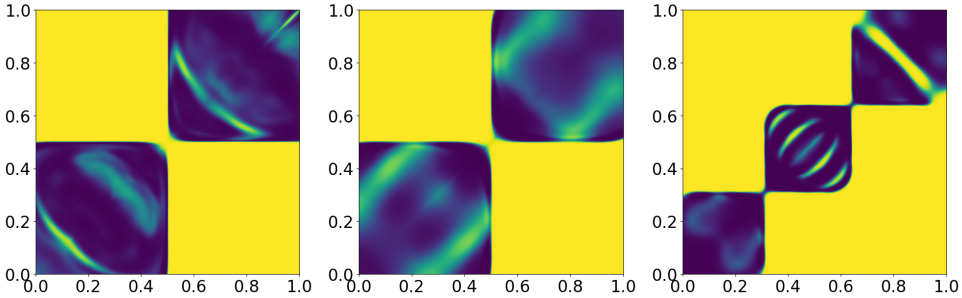
G Additional experimental results

We consider the problem of minimising the homomorphism density of the Petersen graph under the edge-density constraint $t(K_2, W) = p$ for various values of p. For minimising the Petersen homomorphism density, there are (to the best of our knowledge) no known results or conjectures, likely due to the graph size and the lack of an obvious extremal structure. We applied our method for $p \in \{1/2, 2/3, 4/5, 4/7, 5/8, 5/7\}$; the learnt graphons are shown in Fig. 7. As there is no known conjectured optimum, we compare our results to those obtained by a hand-crafted graphon motivated by Razborov [2008]; see §H for details.

We also report the complete results of the variational problem discussed in §4; see Figs. 8 to 12 for the found graphons. We compute the results by solving the constraint after the discretisation of the learnt graphon, which gives a step graphon that satisfies the constraint exactly. For the quantitative comparison, we include the tripartite construction W_{LZ} of Lubetzky and Zhao [2015] only in the cases where this construction is available in the literature, namely the *upper-tail* problem for regular graphs. Accordingly, Tab. 8 compares $h_q(W_{\mathrm{ours}})$ with $h_q(W_{\mathrm{LZ}})$ for $H \in \{C_4, C_5\}$ under constraints $t(C_k, W) \geq r^k$. For the $K_{1,3}$ upper-tail problem, Tab. 8 reports only our learnt solutions (no corresponding W_{LZ} construction is available in the literature for this case). For the *lower-tail* problem for K_3, we likewise report only our learnt solutions; see Tab. 9.



(a) Found graphon, $p = 1/2$. (b) Found graphon, $p = 2/3$. (c) Found graphon, $p = 4/5$.



(d) Found graphon, $p = 4/7$. (e) Found graphon, $p = 5/8$. (f) Found graphon, $p = 5/7$.

Figure 7: Graphons found for the Petersen graph density minimisation under the edge-density constraint $t(K_2, W) = p$.

Table 8: Relative-entropy $h_q(W)$ at our learnt graphon (W_{ours}) and, for $H \in \{C_4, C_5\}$, at the tripartite construction W_{LZ} of Lubetzky and Zhao [2015], for **(P2)** on $(q, r) \in \{0.05, 0.1\} \times \{0.4, 0.5, 0.6\}$. We also report a lower bound on the optimum for $H \in \{C_4, C_5\}$, and W_{ours} for $H = K_{1,3}$ (no corresponding W_{LZ} construction is available in the literature for this case). Lower is better.

	$q = 0.05$			$q = 0.10$		
r	0.4	0.5	0.6	0.4	0.5	0.6
$h_q(W_{\text{LZ}})$, $H = C_4$	0.556057	0.830366	1.144945	0.311239	0.510826	0.750684
$h_q(W_{\text{ours}})$, $H = C_4$	0.464162	0.727364	1.050351	0.309184	0.506141	0.747042
$h_q(W_{\text{LZ}})$, $H = C_5$	0.556057	0.830366	1.144945	0.311239	0.510826	0.750684
$h_q(W_{\text{ours}})$, $H = C_5$	0.465644	0.727618	1.052023	0.308858	0.506122	0.746845
Lower bound, $H \in \{C_4, C_5\}$	0.460284	0.725284	1.049172	0.306561	0.504311	0.746006
$h_q(W_{\text{ours}})$, $H = K_{1,3}$	0.357819	0.665939	1.076113	0.263666	0.484864	0.748724

When H is d-regular, Lubetzky and Zhao [2015] also prove a lower bound on the optimum of the upper-tail problem. Three graphs in our experiments, K_3, C_4, and C_5, satisfy this assumption, and the computed lower bound is reported in Tabs. 6 and 8; its values are the same for the three graphs, as the bound is determined by d.

H Details of hand-crafted graphons

We consider edge densities satisfying $1 - 1/k < p < 1 - 1/(k+1)$, where k is a positive integer.

The first two instances are graphons arising from the optimal solution to edge-density-constrained K_3 homomorphism-density minimisation. We construct a complete $(k+1)$-partite graphon by

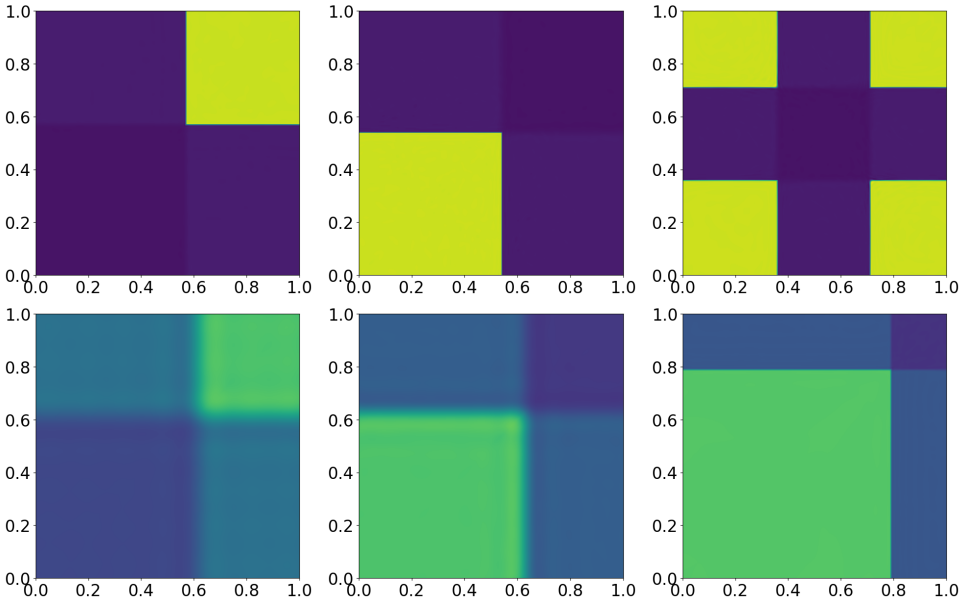


Figure 8: Found graphon for upper tail variational problems for K_3, for $q = 0.05$ (top) and $q = 0.1$ (bottom) and $r = 0.4, 0.5, 0.6$ (left to right).

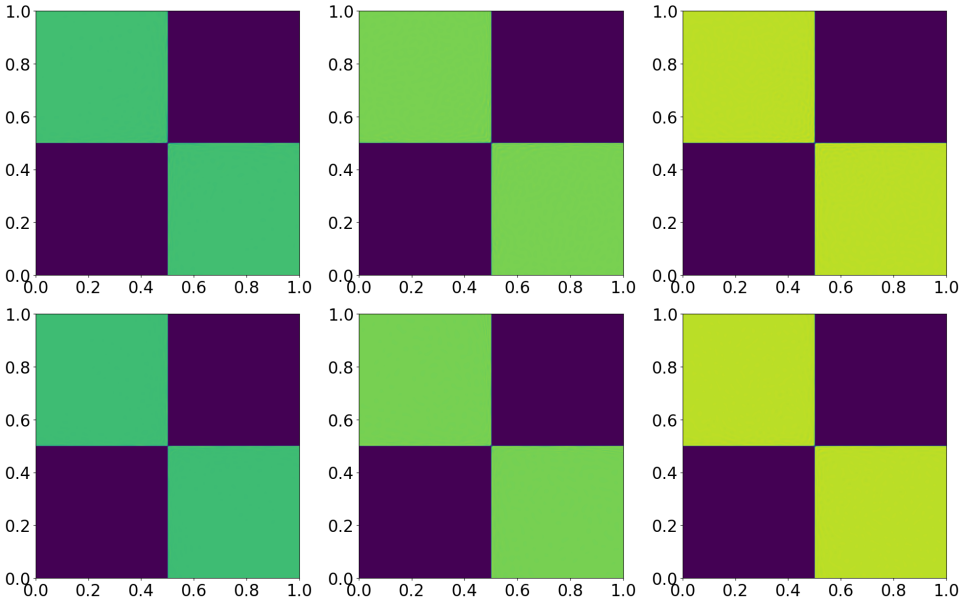


Figure 9: Found graphon for lower tail variational problems for K_3, for $r = 0.05$ (top) and $r = 0.1$ (bottom) and $q = 0.7, 0.8, 0.9$ (left to right).

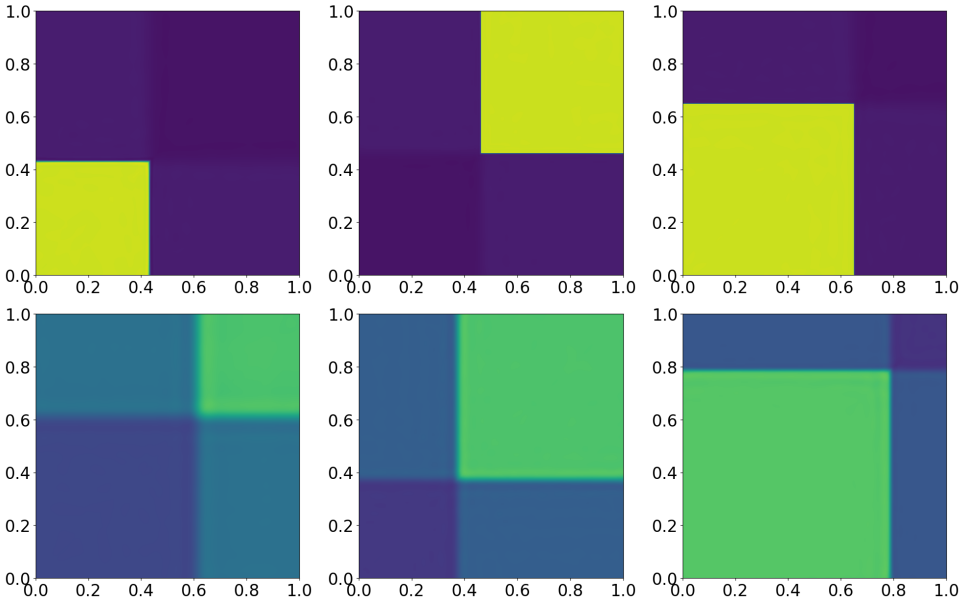


Figure 10: Found graphon for upper tail variational problems for C_4, for $q = 0.05$ (top) and $q = 0.1$ (bottom) and $r = 0.4, 0.5, 0.6$ (left to right).

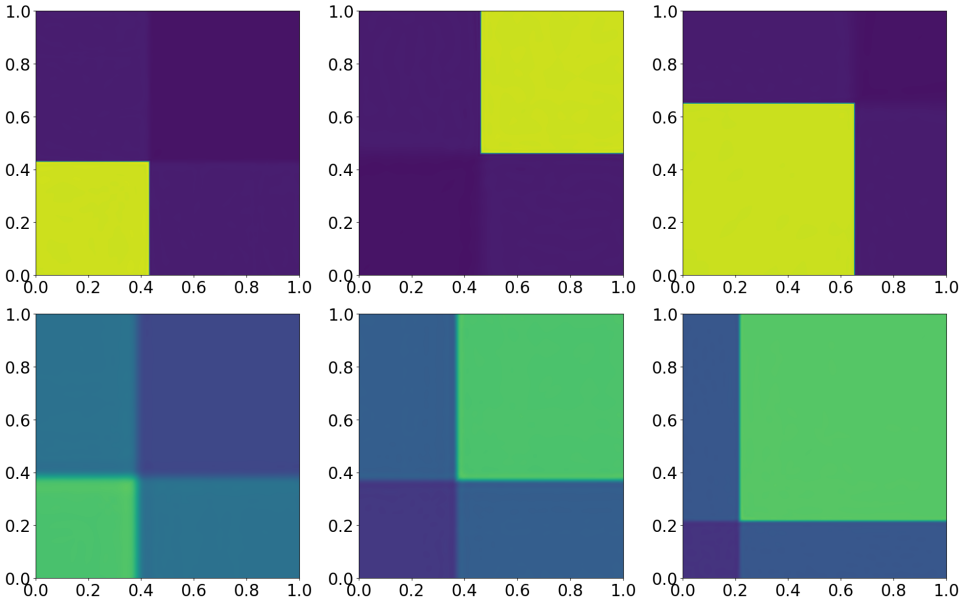


Figure 11: Found graphon for upper tail variational problems for C_5, for $q = 0.05$ (top) and $q = 0.1$ (bottom) and $r = 0.4, 0.5, 0.6$ (left to right).

Table 9: Relative-entropy $h_q(W)$ at our learnt graphon (W_{ours}) for **(P2)** with $H = K_3$ on $(q, r) \in \{0.7, 0.8, 0.9\} \times \{0.05, 0.1\}$. Lower is better.

	$r = 0.05$			$r = 0.10$		
q	0.7	0.8	0.9	0.7	0.8	0.9
$h_q(W_{\text{ours}})$	0.600882	0.591680	0.804127	0.796391	1.151161	1.144472

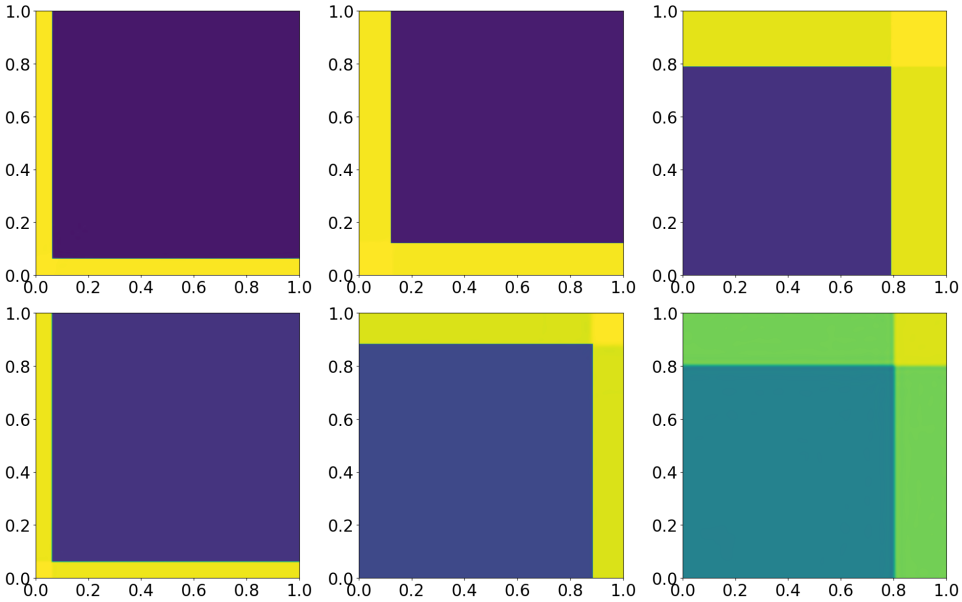


Figure 12: Found graphon for upper tail variational problems for $K_{1,3}$, for $q = 0.05$ (top) and $q = 0.1$ (bottom) and $r = 0.4, 0.5, 0.6$ (left to right).

partitioning $[0, 1]$ into $(k + 1)$ groups, $X_1, \ldots, X_{k+1}$, and defining

$$W_i(x_1, x_2) = \begin{cases} 1 & \text{if } x_1 \in X_m,\ x_2 \in X_n,\ m \neq n \\ 0 & \text{if } x_1 \in X_m,\ x_2 \in X_m. \end{cases}$$

We set $|X_1| = \cdots = |X_k| = x$ and $|X_{k+1}| = 1 - kx$, and choose x to satisfy the edge-density constraint. This amounts to solving the quadratic equation

$$k(k-1)x^2 + 2kx(1 - kx) = p.$$

This quadratic has two solutions: one with $x > 1 - kx$ (denoted W_1) and one with $x < 1 - kx$ (denoted W_2). The first corresponds to the optimum for K_3 density minimisation.

The next too instance is constructed by deleting edges from the complete $(k + 1)$-partite graphon. We partition the interval $[0, 1]$ into $k + 1$ groups $X_1, \ldots, X_k, X_{k+1}$ with $|X_1| = \cdots = |X_k| = |X_{k+1}| = 1/(k + 1)$, and define

$$W_3(x_1, x_2) = \begin{cases} \frac{(k+1)p}{k} & \text{if } x_1 \in X_m,\ x_2 \in X_n,\ m \neq n \\ 0 & \text{if } x_1 \in X_m,\ x_2 \in X_m, \end{cases}$$

$$W_4(x_1, x_2) = \begin{cases} \frac{p(k+1)^2 - k^2 - k + 2}{2} & \text{if } x_1 \in X_1,\ x_2 \in X_2 \text{ or vice versa} \\ 0 & \text{if } x_1 \in X_m, x_2 \in X_m, \\ 1 & \text{otherwise.} \end{cases}$$

For $1 - 1/k < p < 1 - 1/(k + 1)$, both nontrivial edge probabilities above lie in $[0, 1]$: their non-negativity follows from the lower bound on p, while $(k+1)p/k < 1$ and $\big(p(k+1)^2 - k^2 - k + 2\big)/2 < 1$ follow from the upper bound.

Finally, the last graphon is the constant graphon $W(x_1, x_2) = p$.

We visualised $W_1, \ldots, W_4$ for $p = 0.7$ in Fig. 13.

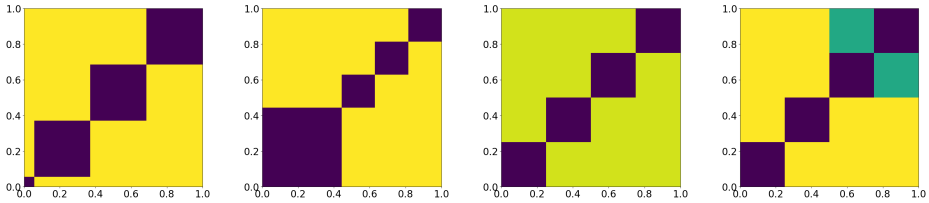


Figure 13: Hand-crafted graphons W_1, W_2, W_3, W_4 (left to right) with edge density $t(K_2, W) = 0.7$.

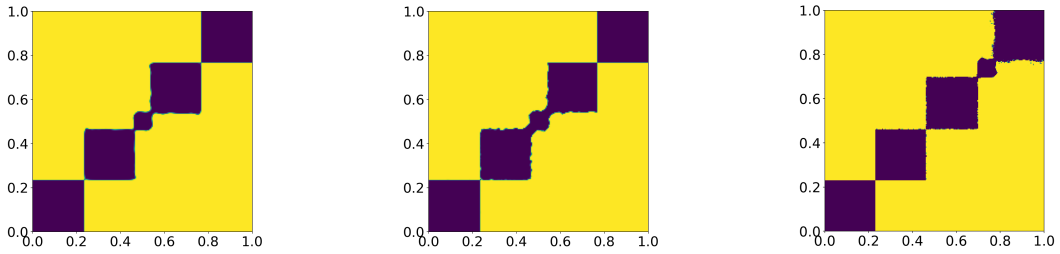
Table 10: K_3 densities in graphons found by our method, ablations, and non-neural baselines for K_3 density minimisation under $t(K_2, W) = 7/9$, with sinusoidal features or with GELU features in the per-layer encoding and activation. The optimum is $98/225 \approx 0.435556$.

	$t(K_3, W)$	
	sin	GELU
Ours	0.435712	0.436066
ResNet	0.435724	0.436405
Constant-scale s	0.435689	0.436347
Regulariser	0.437085	0.437323
Constant LR (low)	0.437544	0.442451
Constant LR (high)	0.436174	0.436743
SBM (256 blocks, fixed sizes)	0.435817	
SBM (5 blocks, trainable sizes)	0.470508	
SBM (64 blocks, trainable sizes)	0.470508	
Decision Tree	0.455558	
Gradient Boosted Decision Tree	0.454485	

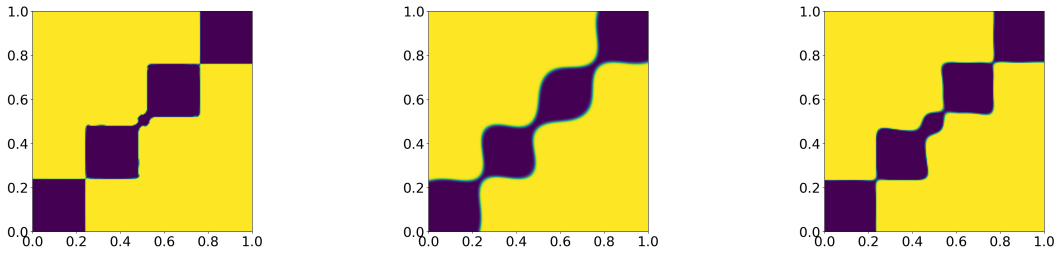
I Ablation study, robustness, non-neural baselines, and multi-task variant

Ablation Study. We perform an ablation study on triangle-density minimisation under the constraint $t(K_2, W) = 7/9$, which is not of the form $1 - 1/k$ and admits multiple types of optima. We consider four variants: (1) replacing our architecture with a ResNet that uses sinusoidal input encoding only in the first layer; (2) using a constant per-layer input encoding scale s; (3) enforcing the constraint via a regulariser; and (4) using a constant learning rate instead of cosine-annealing scheduling. For fair comparison, we performed hyperparameter searches for the constant value of s in (2) and the strength of the regulariser in (3), and report the best run for each variant; for (4), we report the runs with a low and a high learning rate. To separate the effect of the sinusoidal features from that of their per-layer injection, we also repeat our method and every variant with GELU in place of every sine function, keeping the per-layer injection; in the ResNet variant, the sinusoidal encoding of the first layer is kept. As in §G, all values are computed after solving for the scalar bias c on the discretised graphon. The graphons found by these variants are shown in Fig. 14 and quantitative results in Tab. 10. With sinusoidal features, every variant reaches a lower K_3 density than with GELU. Since the GELU column also changes the activation ϕ, we also compare with the ten runs in the robustness study below, which use sinusoidal features with $\phi = \mathrm{GELU}$: their mean 0.435661 is also below the 0.436066 of our method with GELU features. The ResNet variant (1) and the constant-scale variant (2) often fail to recover sharp, straight boundaries, consistent with a bias toward smooth functions in (1) and toward high frequencies in (2). With the penalty-based constraint (3), the learnt graphon often violates fine structural details (e.g., the size of the smallest block). Without cosine-annealing scheduling (4), the learnt graphon either fails to capture the multi-block structure (low learning rates) or exhibits less stable boundaries (high learning rates), highlighting the benefit of alternating between higher and lower learning rates.

Multi-scale encoding. The increasing scale $s(\ell)$ is motivated by the step-like structure of the optima: approximating a step boundary requires increasingly high spatial frequencies. Early layers, whose encodings have low frequencies, provide the coarse block geometry, while later layers, whose encodings have higher frequencies, refine the boundaries through the residual connections, so that earlier layers do not need to carry high-frequency information. The scale enters only through the



(a) Our method. (b) ResNet variant. (c) Constant-scale s.



(d) Regulariser. (e) Const. LR (low). (f) Const. LR (high).

Figure 14: Graphons found by our method and its ablations, for K_3 density minimisation under the constraint $t(K_2, W) = 7/9$.



Figure 15: Graphons obtained by applying the output layer to the hidden state $g_\theta^{(\ell)}$ after layer ℓ of a trained network, for K_3 density minimisation under $t(K_2, W) = 7/9$, with the scalar bias c solved separately for each ℓ so that $t(K_2, W) = 7/9$; $\ell = 1, \ldots, 6$ from left to right, and $\ell = 6$ is the learnt graphon.

initialisation of $U^{(\ell)}$, whose entries are drawn from $\mathrm{Unif}[-s(\ell)/\sqrt{2}, s(\ell)/\sqrt{2}]$, so $s(\ell)$ sets the initial spatial frequency of the encoding $\sin(U^{(\ell)}\mathbf{x})$ in layer ℓ. Since $\sin$ is bounded and the remaining weights are initialised with scale $1/\sqrt{2d}$, increasing $s(\ell)$ gives later layers access to finer scales without increasing the magnitude of the hidden states. The constant-scale variant (2) above isolates the effect of increasing the scale. To see the contribution of each layer, Fig. 15 applies the output layer to the hidden state $g_\theta^{(\ell)}$ after each layer ℓ of a trained network. The first layer gives only a smooth gradient, the next layers form a band along the diagonal and split it into blocks, and the last layers sharpen the block boundaries into the five-block structure of the learnt graphon.

Robustness. Except for the exhaustive study, the results in the main text are the best of eight training runs over the activation and the learning rate (§4). To measure the variability across random initialisations separately, we repeated five instances ten times each under a single fixed configuration (GELU activation, learning rate 10^{-3}), varying only the random initialisation, and report the mean with a 95% confidence interval:

- K_3 at $p = 7/9$: 0.435661, $[0.435656, 0.435667]$, against the optimum $98/225 \approx 0.435556$.
- C_5 at $p = 3/5$: 0.066731, $[0.066706, 0.066756]$, against the conjectured optimum 0.066616.
- C_7 at $p = 3/5$: 0.026467, $[0.026134, 0.026800]$, against our conjectured optimum 0.026261. One run converged to the constant graphon, with density p^7; the other nine give 0.026297, $[0.026279, 0.026315]$.

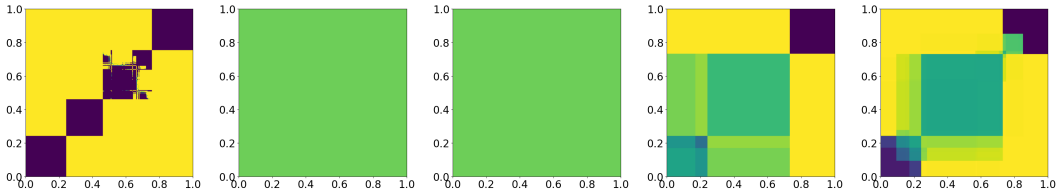


Figure 16: Graphons found by non-neural baselines, fixed-block SBM, trainable-block SBM with 5 / 64 blocks, decision tree, and tree ensemble (left to right), for K_3 density minimisation under the constraint $t(K_2, W) = 7/9$.

- H_6 at $p = 4/5$: 0.129877, $[0.129851, 0.129902]$, against 0.13056 for the k-partite graphon and 0.134218 for the constant graphon.
- The upper-tail problem for K_3 at $(q, r) = (0.05, 0.5)$: 0.727215, $[0.727210, 0.727220]$, against the lower bound 0.725284 (Tab. 6).

The learnt structures are also consistent across seeds. For K_3, seven runs recover the clean five-block partition, and the other three produce five parts whose smallest part is not an exact block, with comparable objective values; since the optimum is not unique, this does not indicate suboptimality. For C_5, all ten runs recover the two-block structure of the bottom row of Fig. 3, and for C_7, the nine nonconstant runs recover that of Fig. 4. For H_6, eight runs recover the pattern of Fig. 5, and the other two show the same band with an oscillating boundary, at objectives 0.04% and 0.10% above the average of the eight. For the upper-tail problem, all ten runs recover the two-block structure of Fig. 8. Across the 702 instances of the exhaustive study in §4, the three seeds have a median relative standard deviation of 1.4×10^{-4} and differ by more than 1% in 5.8% of the instances.

We also checked the sensitivity to the other sources of randomness. For C_5, training with permutation sets $\mathcal{S}$ of size 1, 2, and 12 in (8) gives objectives that agree within 1.6×10^{-4} at each of the six densities of Tab. 3. Finally, for the learnt graphon for K_3 at $p = 7/9$, solving for the scalar bias c on 1000 independent batches of 2^{16} samples, as in training, gives a constraint violation with standard deviation 1.1×10^{-3} and maximum 3.4×10^{-3}; this error does not affect the reported values, which solve for c on the discretised graphon, and refining the discretisation from 256^2 to 512^2 points changes $t(K_3, W)$ by 9.7×10^{-5}.

Non-neural Baselines. On the same triangle-density-minimisation task we evaluated four non-neural baselines tailored to step graphons: (a) a stochastic block model (SBM) with 256 equal-sized blocks and trainable edge densities; (b) SBMs with 5 blocks and 64 blocks, with trainable edge densities and trainable block sizes; (c) a decision tree; and (d) a tree ensemble. Each baseline was run four times from independent random initialisations, and we report the best objective value across runs in Tab. 10. The SBMs with trainable block sizes collapsed to nearly constant graphons, suggesting a difficult optimisation landscape; the fixed-block SBM recovered a noisy multi-block structure; the tree-based baselines produced structured but qualitatively different graphons. See Fig. 16 for visualisations of the found graphons.

These baselines can represent the optimal graphon, at least approximately; the difference lies in optimisation. For the fixed-block SBM and the tree-based baselines, moving a block boundary is hard, since almost every configuration with values in $\{0, 1\}$ is a local optimum, so these methods rarely leave the geometry of their initialisation. Making the block sizes trainable does not help, as these SBMs collapse to nearly constant graphons. In contrast, the neural graphon adjusts the values and the boundaries of its regions jointly and continuously, and it does not require choosing in advance between multipartite, banded, circular-distance, or two-block structures.

Multi-task Variant. The main paper trains one network per (H, p). To probe whether shared training across related tasks is viable, we trained a single conditional network jointly over $p \in [1/2, 2/3]$ for $H = K_3$. Concretely, we condition the network on p using FiLM-style feature-wise affine modulation of hidden activations [Perez et al., 2018]. The resulting conditional network achieves absolute error at most 2.09×10^{-4} and average absolute error 3.12×10^{-5} relative to the known optimum, for p on a uniform grid of 100 points in $[1/2, 2/3]$. See Fig. 17 for the learnt density curve and its error, and Fig. 18 for the learnt graphons. While limited to a single graph, this experiment

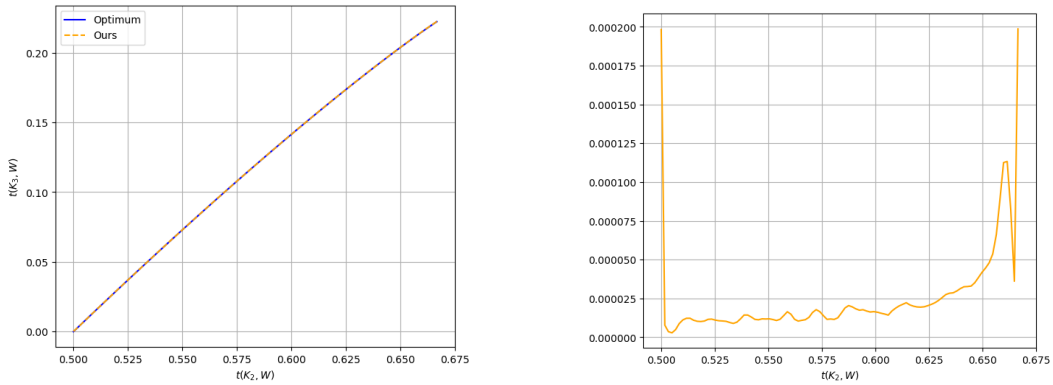


Figure 17: (Left) Triangle density as a function of p for the conditional network, compared to the known optimum. (Right) Absolute error of the conditional network's triangle density relative to the known optimum.

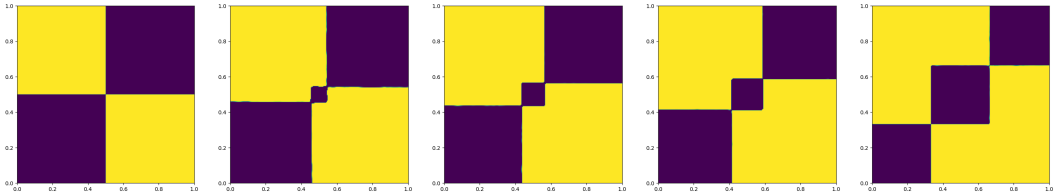


Figure 18: Learnt graphons for the conditional network at $p \in \{1/2, 4/7, 3/5, 5/8, 2/3\}$ (left to right).

indicates that shared training across related tasks can substantially reduce the per-instance burden of our framework, and we view a fuller multi-task / multi-graph treatment as a natural extension.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide quantitative and qualitative results in the main paper that support these claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section B of the supplementary material discusses the limitations of our work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We include a supporting lemma for our methodology (Lemma 3.1) and a complete proof in the supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The supplementary material lists all hyperparameters and training details, and we provide code to reproduce our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide open access to the code needed to reproduce our main experimental results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: §F of the supplementary material describes the hyperparameters and optimiser, and the code reproduces the experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: For experiments that rely on random sampling in evaluation, we report 95% bootstrap confidence intervals. Variability across training runs was measured only for K_3 at $p = 7/9$, C_5 and C_7 at $p = 3/5$, H_6 at $p = 4/5$, and the upper-tail problem for K_3 at $(q, r) = (0.05, 0.5)$, each repeated ten times (§I).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: §F of the supplementary material reports the compute resources used in our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section B of the supplementary material discusses the potential societal impacts of our work.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not foresee misuse risks, so no special safeguards are required.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: We do not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We do not use LLMs in the core method or experiments; we use them for writing/editing and some implementation support.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.