An introduction to graph limits

Travis Dillon

Contents

1 Introduction

The question that motivates the entire subject is: How do we understand large networks? Examples of such networks include the social graph, where two people are connected if they know each other; the internet graph, where two devices are adjacent if they are connected through the internet; models of ecological systems; and molecular networks. Each of these networks varies throughout time: Animals are born or die and social connections are severed or formed. As a result, we model these networks using sequences of graphs, representing the passage of time. We can ask questions such as: Does the behavior of this sequence converge in some way? If so, what does that mean? Thus begins the theory of graph limits.

There are two special types of graph sequences. To explain, let (G_n) be a sequence of graphs where $|V(G_n)|$ tends to infinity. (We make this assumption so that asymptotic analysis makes sense.) The sequence (G_n) is *dense* if there is a constant $c > 0$ so that G_n has asymptotically at least $c|V(G_n)|^2$ edges. (In other words, G_n contains a positive proportion of its possible connections.) The sequence is *sparse* if there is a constant $D \geq 0$ so that $\deg(v) \leq D$ for all $v \in V(G_n)$ and $n \in \mathbb{N}$.

A dense graph sequence has $\Omega(|V(G_n)|^2)$ edges, while a sparse sequence has $O(|V(G_n)|)$ edges. Two parallel but separate theories have been developed for graph limits of these classes, and little is known about graphs in between, for example graph sequences with $\Theta(|V(G_n)|^{3/2})$ edges. The limit theory of sparse graphs is a bit more technical, so these notes discuss limits of dense graph sequences. As a side note, these ideas have been used to create limit theories for a variety of other combinatorial objects. Naturally, there is now a theory of hypergraph convergence, but there are also limit theories for partially-ordered sets and permutations.

As for these notes, Sections 2–4 introduce the idea of convergence for graph sequences and the limits of convergent sequences. Sections 5–8 develop the theory of graph limits, and Sections 9 and 10 provide applications of graph limits to extremal graph theory. Throughout, there are a variety of connections to material throughout mathematics, including measure theory, probability, and Szemerédi's Regularity Lemma from graph theory.

The first paper on graph limits [1] was published in 2008 (although of course steps toward this theory had been made before then). Consequently, the only textbook on the subject is the one by Lovász [2], which covers far more material than these notes. In particular, it addresses limits of sparse graph sequences. The graph limits portion of Zhao's notes [3] is also a nice resource.

2 Graph homomorphisms

All of the graphs will be finite, undirected, and simple (no multiedges and no loops). We denote by $\mathsf{v}(G)$ and $\mathsf{e}(G)$ the number of vertices and edges of G, respectively. We define $[n] = \{1, 2, \ldots, n\}$.

One of the ways to understand large networks is to ascertain their local structure. Practically speaking, the way we do this is by sampling. Large networks are too, well, *large* to computationally deal with the entire structure. Instead, we can choose some relatively small number of vertices and query the database about the connections between these specific vertices. The hope is that by repeating this multiple times and choosing these vertices in a clever way, we'll be able to learn something about the network. In fact, we won't be clever at all—we'll choose the sampled vertices randomly. It turns out that this procedure actually yields useful information for dense graphs. (For sparse graphs, virtually every random sample is the empty graph, so there's no hope of obtaining information by sampling this way.) So how do we formulate this as math? Instead of focusing on sampling *per se*[1], we'll use a related graph theory notion: homomorphism.

Definition 2.1. A *homomorphism* between two graphs F and G is a map $\phi\colon V(F) \to V(G)$ such that $\phi(i)\phi(j) \in E(G)$ for all $ij \in E(F)$.

A homomorphism need not preserve non-adjacency. We denote by $\mathsf{hom}(F, G)$ the number of homomorphisms from F to G and by $\mathsf{inj}(F, G)$ the number of injective homomorphisms from F to G. In terms of sampling and local structure, $\mathsf{inj}(F, G)$ is the number of (labelled) copies of F that reside in G. An *induced homomorphism* is an injective homomorphism whose image is an induced subgraph; that is, an induced homomorphism preserves non-adjacency as well. The number of induced homomorphisms from F to G is denoted $\mathsf{ind}(F, G)$. Finally, we denote the number of isomorphisms (bijective induced homomorphisms) from F to G by $\mathsf{iso}(F, G)$ and the number of automorphisms of F (isomorphisms from F to F) by $\mathsf{aut}(F)$.

Example 2.2. We denote the path graph and cycle graph with n vertices by P_n and C_n, respectively.

- $\mathsf{hom}(P_3, P_2) = 2$ and $\mathsf{inj}(P_3, P_2) = \mathsf{ind}(P_3, P_2) = 0$.
- $\mathsf{hom}(P_2, P_3) = \mathsf{inj}(P_2, P_3) = \mathsf{ind}(P_2, P_3) = 4$.
- $\mathsf{aut}(C_n) = 2n$.[3]

2.1 Basic relations and properties of homomorphism numbers

We start with two straightforward relations between homomorphism numbers.

Definition 2.3. Let F_1 and F_2 be two graphs. Their *disjoint union* is the graph $F_1 \cup F_2 = F_1F_2$ on vertex set $V(F_1) \sqcup V(F_2)$, where $ij \in E(F_1F_2)$ if and only if $ij \in E(F_1)$ or $ij \in E(F_2)$.[4] We denote the k-fold disjoint union of F by kF.

Proposition 2.4. $\mathsf{hom}(F_1F_2, G) = \mathsf{hom}(F_1, G)\mathsf{hom}(F_2, G)$ *for all graphs* F_1, F_2 *and* G.

Proof. A homomorphism from F_1F_2 to G is an independent pair of homomorphisms $F_1 \to G$ and $F_2 \to G$. □

[1] Amn't[2] I fancy.

[2] Apparently a contraction used in Ireland and Scotland.

[3] Never one to shy away from stirring up controversy with correct notation, the author notes that this set of automorphisms is the dihedral group D_n.

[4] You can think of this as placing the graphs beside each other and then considering the two of them together to be a single graph.

Proposition 2.5. *If $F_1 \subseteq F_2$, then $\mathsf{hom}(F_2, G) \leq \mathsf{hom}(F_1, G)\mathsf{v}(G)^{\mathsf{v}(F_2)-\mathsf{v}(F_1)}$.*

Proof. Any homomorphism $F_2 \to G$ is also a homomorphism from $F_1 \cup (V(F_2) \setminus V(F_1))$, and each of these is distinct. The graph $V(F_2) \setminus V(F_1)$ is just $\mathsf{v}(F_2) - \mathsf{v}(F_1)$ isolated vertices and $\mathsf{hom}(K_1, G) = \mathsf{v}(G)$, so we apply Proposition 2.4 to finish the proof. □

We now state and prove some relations between the different types of homomorphism numbers. These relations are used only infrequently, so the rest of this section can be skimmed or postponed.

To state the relations, we introduce some notions. Fix a graph F. A set $S \subseteq V(F)$ is *independent* if no two vertices in S are adjacent. Let P be a partition of $V(F)$ into independent sets. The graph F/P has vertex set P and the edge XY if and only if there are $x \in X$ and $y \in Y$ with $xy \in E(F)$. That is, F/P is formed by gluing together the vertices in each independent set of P and removing multiedges.

Proposition 2.6. *For any simple graphs F and G, the following relations hold:*

$$\mathsf{inj}(F, G) = \sum_{\substack{F' \supseteq F \\ \mathsf{v}(F')=\mathsf{v}(F)}} \mathsf{ind}(F, G) \tag{2.1}$$

$$\mathsf{hom}(F, G) = \sum_{P} \mathsf{inj}(F/P, G) \tag{2.2}$$

$$\mathsf{ind}(F, G) = \sum_{\substack{F' \supseteq F \\ \mathsf{v}(F')=\mathsf{v}(F)}} (-1)^{\mathsf{e}(F')-\mathsf{e}(F)} \mathsf{inj}(F, G) \tag{2.3}$$

$$\mathsf{inj}(F, G) = \sum_{P} \mu_P \mathsf{hom}(F/P, G), \tag{2.4}$$

where the μ_P are integer coefficients depending only on F.

To prove this, we'll use some lemmas.

Lemma 2.7. *For any two finite sets $A \subseteq C$, we have*

$$\sum_{A \subseteq B \subseteq C} (-1)^{|C|-|B|} = \begin{cases} 1 & \text{if } A = C \\ 0 & \text{otherwise.} \end{cases}$$

Proof. There are $\binom{|C|-|A|}{k}$ sets B of cardinality $|A|+k$ so that $A \subseteq B \subseteq C$. The sum is therefore

$$\sum_{k=0}^{|C|-|A|} (-1)^k \binom{|C| - |A|}{k},$$

which is 0 by the binomial theorem whenever $A \neq C$. If $A = C$ the lemma is obvious. □

Lemma 2.8 (Möbius inversion)**.** *Let X be a finite set and $f, g \colon \mathcal{P}(X) \to \mathbb{C}$ with $f(A) = \sum_{B \subseteq A} g(B)$ for all $A \subseteq X$. Then $g(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|} f(B)$.*

Proof. We can prove this by directly calculating from the definition:

$$\sum_{B \subseteq A} (-1)^{|A|-|B|} f(B) = \sum_{B \subseteq A} (-1)^{|A|-|B|} \sum_{C \subseteq B} g(C).$$

The two sums index over all sets B and C with $C \subseteq B \subseteq A$, so we can rewrite the sum as

$$\sum_{C \subseteq A} g(C) \sum_{C \subseteq B \subseteq A} (-1)^{|A|-|B|}.$$

Lemma 2.7 shows that every term except $C = A$ vanishes, so the expression collapses to the single term $g(A)$. □

We can also present the proof using matrices: Consider f and g as column vectors indexed by subsets of X. If we define the $2^{|X|} \times 2^{|X|}$ matrix M by

$$M(S,T) = \begin{cases} 1 & \text{if } S \supseteq T \\ 0 & \text{otherwise,} \end{cases}$$

then $f = Mg$. If we order the subsets of X according to increasing size and then use that ordering for the rows and columns of M, we find that M is an upper-triangular matrix where every diagonal entry is 1. So $\det(M) = 1$ and M has an inverse. Indeed, by defining

$$M'(S,T) = \begin{cases} (-1)^{|S|-|T|} & \text{if } S \supseteq T \\ 0 & \text{otherwise,} \end{cases}$$

Lemma 2.7 gives that $M'M = I$. Then $g = M'f$, which shows that

$$g(S) = \sum_{T \subseteq X} M'(S,T) f(T) = \sum_{T \subseteq S} (-1)^{|S|-|T|} f(T).$$

Definition 2.9. Let A be an $n \times n$ matrix and $A'(i,j)$ be the matrix obtained by deleting the ith row and jth column of A. The *cofactor matrix* of a square matrix A is the $n \times n$ matrix $C_A = (c_{i,j})$ whose entries are the cofactors of A, that is, $c_{i,j} = (-1)^{i+j} \det(A'(i,j))$. The *adjugate matrix* of A is the transpose of the cofactor matrix: $\operatorname{adj}(A) = C_A^T$.

Proposition 2.10. *If A is an invertible matrix, then $A^{-1} = \frac{1}{\det A} \operatorname{adj}(A)$.*

Proof. Let A be an $n \times n$ invertible matrix and set $B = \operatorname{adj}(A)$. We calculate $(AB)_{i,j}$. Note that $B_{i,j} = (C_A)_{j,i}$. If $i = j$, then we have

$$(AB)_{i,i} = \sum_{k=1}^{n} A_{i,k} (C_A)_{i,k} = \det(A)$$

by the definition of cofactor expansion. On the other hand, if $i \neq j$, then

$$(AB)_{i,j} = \sum_{k=1}^{n} A_{i,k} (C_A)_{j,k}$$

is the cofactor expansion of a matrix whose ith and jth rows are equal, so $(AB)_{i,j} = 0$. Therefore $AB = \det(A) I_n$, and the conclusion follows. □

Corollary 2.11. *If A is a invertible matrix of integers and $|\det A| = 1$, then A^{-1} also has integer entries.*

With all this preparation, we can prove the proposition.

Proof of Proposition 2.6. For (2.1), note that each injective homomorphism is an induced homomorphism for a graph with more edges, and vice versa. To get (2.2), we see that any homomorphism is an injective homomorphism from the quotient graph F/Q, where Q is the collection of preimages of vertices of G (and Q is an independent set because the map is a homomorphism, so adjacent vertices in F cannot be mapped to the same vertex in G). Further, any injective homomorphism from a quotient graph gives rise to a homomorphism of F. Equation (2.3) follows from Lemma 2.8 applied to (2.1).

Finally, we prove (2.4). Let f be the vector whose entries are indexed by all quotient graphs of F, with entries $f(F/P) = \mathsf{hom}(F/P, G)$. Similarly, let g be the whose entries are $g(F/P) = \mathsf{inj}(F/P, G)$. We define a square matrix M indexed by these quotient graphs, where $M(F/P, F/Q)$ is 1 if F/Q is a quotient of F/P and 0 otherwise. Then $f = Mg$ by (2.2).

If we order the quotient graphs by number of vertices, then M is a triangular matrix with every diagonal entry equal to 1. So $|\det M| = 1$ and by Corollary 2.11, M^{-1} is an integer-valued matrix. Then $g = M^{-1}f$ and

$$\mathsf{inj}(F, G) = g(F) = \sum_P M^{-1}(F, F/P)\mathsf{hom}(F/P, G).$$

Setting $\mu_P = M^{-1}(F, F/P)$ finishes the proof. □

Exercise 2.12. Work out similar relations for the homomorphism numbers as functions of the second argument.

Here is a nice application of these relations: The homomorphism numbers of a graph completely determine the graph itself.

Theorem 2.13. *Two graphs G_1 and G_2 are isomorphic if and only if $\mathsf{hom}(F, G_1) = \mathsf{hom}(F, G_2)$ for every finite simple graph F.*

Proof. The forward direction is immediate. For the reverse direction, identity (2.4) shows that $\mathsf{inj}(F, G_1) = \mathsf{inj}(F, G_2)$ for all F. Similarly, identity (2.3) shows that the number of induced homomorphisms must be the same. Taking $F = G_1$ shows that $\mathsf{ind}(G_1, G_2) = \mathsf{ind}(G_2, G_2) > 0$, so there is an induced homomorphism from G_1 to G_2. It follows that $\mathsf{v}(G_1) \leq \mathsf{v}(G_2)$. But taking $F = G_2$ shows the opposite, so $\mathsf{v}(G_1) = \mathsf{v}(G_2)$, and any induced homomorphism is an isomorphism. □

As a remark, the reconstruction conjecture posits that we need slightly less. The conjecture is notoriously difficult and is still unsolved.

Conjecture 2.14 (Reconstruction conjecture, 1963)**.** *Let G_1 and G_2 be two graphs with at least 3 vertices. If $\mathsf{hom}(F, G_1) = \mathsf{hom}(F, G_2)$ for all graphs F with $\mathsf{v}(F) < \mathsf{v}(G_1)$, then $G_1 \cong G_2$.*

2.2 Homomorphism numbers at work

Many quantities can be expressed in terms of homomorphism numbers: Denoting by P_k the path with k vertices, the number of walks of k vertices in G is $\mathsf{hom}(P_k, G)$. The number of closed walks of length k is $\mathsf{hom}(C_k, G)$. Let S_k denote the star on k vertices. A homomorphism from S_k to G consists of mapping the central vertex of S_k and then mapping the remaining vertices to its neighbors, so $\mathsf{hom}(S_k, G) = \sum_{v \in G} \deg(v)^{k-1}$. The number of proper q-colorings of G is $\mathsf{hom}(G, K_q)$, and the chromatic number is the least integer n so that $\mathsf{hom}(G, K_n) > 0$. The clique and independence numbers of G are the largest n so that $\mathsf{hom}(K_n, G) > 0$ and $\mathsf{hom}(K_n, \overline{G}) > 0$, respectively. ($\overline{G}$ is the complement of G.)

2.3 Homomorphism densities

For large graphs, the homomorphism numbers are impractical—they're just really big numbers. So we normalize them as follows.

Definition 2.15. The *homomorphism density* of F in G is

$$t(F,G) = \frac{\mathsf{hom}(F,G)}{\mathsf{v}(G)^{\mathsf{v}(F)}}.$$

In other words, $t(F,G)$ is the proportion of random mappings $V(F) \to V(G)$ that are homomorphisms, or the probability that a uniform random such mapping is a homomorphism. We set $n^{\underline{k}} = n(n-1)\cdots(n-k+1)$, the number of injective maps from $[k]$ to $[n]$. The *injective homomorphism density* of F in G is

$$t_{\mathrm{inj}}(F,G) = \frac{\mathsf{inj}(F,G)}{\mathsf{v}(G)^{\underline{\mathsf{v}(F)}}} = \frac{\mathsf{inj}(F,G)}{\mathsf{v}(G)(\mathsf{v}(G)-1)\cdots(\mathsf{v}(G)-\mathsf{v}(F)+1)},$$

the proportion of injective mappings $V(F) \to V(G)$ that are homomorphisms.

If G is large compared to F, then a random vertex mapping is almost certainly injective. This observation can be formalized to show that $t(F,G)$ and $t_{\mathrm{inj}}(F,G)$ are close when G is large.

Proposition 2.16. *Let F and G be graphs on k and n vertices, respectively. Then*

$$|t(F,G) - t_{\mathrm{inj}}(F,G)| \le \frac{1}{n}\binom{k}{2}.$$

Proof. Provisionally, let $\mathsf{ninj}(F,G)$ denote the number of non-injective homomorphisms from F to G, so that $\mathsf{hom}(F,G) = \mathsf{inj}(F,G) + \mathsf{ninj}(F,G)$. Then

$$t(F,G) - t_{\mathrm{inj}}(F,G) = \frac{\mathsf{ninj}(F,G)}{n^k} - \mathsf{inj}(F,G)\left(\frac{1}{n^{\underline{k}}} - \frac{1}{n^k}\right).$$

We bound each term individually. Any non-injective homomorphism is also a non-injective mapping of vertices. To form such mappings, we can choose two vertices of F to be identified and then map the $k-1$ vertices arbitrarily, so $\mathsf{ninj}(F,G) \le \binom{k}{2}n^{k-1}$. It follows that

$$0 \le \frac{\mathsf{ninj}(F,G)}{n^k} \le \frac{1}{n}\binom{k}{2}.$$

For the other term, we have $\mathsf{inj}(F,G) \le n^{\underline{k}}$, so

$$\mathsf{inj}(F,G)\left(\frac{1}{n^{\underline{k}}} - \frac{1}{n^k}\right) \le 1 - \frac{n^{\underline{k}}}{n^k} = \frac{n^k - n^{\underline{k}}}{n^k}.$$

Since $n^k - n^{\underline{k}}$ is the number of non-injective mappings from a k-set into an n-set, we have $n^k - n^{\underline{k}} \le n^{k-1}\binom{k}{2}$, which shows that

$$0 \le \mathsf{inj}(F,G)\left(\frac{1}{n^{\underline{k}}} - \frac{1}{n^k}\right) \le \frac{1}{n}\binom{k}{2}.$$

Combining, we have

$$-\frac{1}{n}\binom{k}{2} \le t(F,G) - t_{\mathrm{inj}}(F,G) \le \frac{1}{n}\binom{k}{2}. \qquad \square$$

This is useful when considering sequences of graphs (G_n) where $\mathsf{v}(G_n) \to \infty$. Some calculations are easier using injective homomorphisms, and some are easier using all homomorphisms; Proposition 2.16 allows us to switch between the two by picking up an error term of $O(1/n)$. We'll see some of this in Section 3.2.

3 Convergence of graph sequences

3.1 Definition and first examples

Let's get to the heart of the matter. What does it mean for a graph sequence to be convergent?[5]

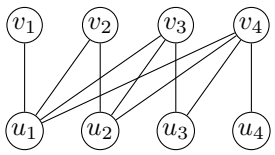
Definition 3.1. A sequence (G_n) of finite simple graphs with $\mathsf{v}(G) \to \infty$ *converges* if $t(F, G_n)$ converges for each finite simple graph F. In this case, we write $t(F, G_n) \to t_F$ for each F.

So a graph sequence is convergent (for us) if its local behavior "settles down". Since we're interested in properties that we can estimate via sampling, this seems like a sensible definition of convergence. What does such a sequence converge to? Let's ignore that in favor of examples. Yay, examples!

Example 3.2. $G_n = K_n$. Any injective mapping $V(F) \to V(K_n)$ is a homomorphism, so $t_{\text{inj}}(F, K_n) = 1$, and $t(F, K_n) \to 1$ for all F by Proposition 2.16. Therefore (K_n) is convergent and $t_F = 1$ for all F.

Example 3.3. $G_n = K_{n,n}$. If F is not bipartite, then $t(F, K_{n,n}) = 0$ for all n. If F has bipartition $A \sqcup B$, then we choose which bipartition of $K_{n,n}$ to map A into, and then there are n choices for each vertex of F, so $\mathsf{hom}(F, K_{n,n}) = 2n^{\mathsf{v}(F)}$. It follows that $t(F, K_{n,n}) = 2^{1-\mathsf{v}(F)}$ for any bipartite F. Since the homomorphism densities are constant, $(K_{n,n})$ is convergent.

Let H_n denote the bipartite graph with vertex set $\{u_1, \dots, u_n\} \sqcup \{v_1, \dots, v_n\}$ and the edge $u_i v_j$ if and only if $i \le j$, called the *half-graph* on $2n$ vertices. For example, H_4 is shown below.



It's possible to prove that (H_n) converges using what we've proven so far, though it's a challenging exercise. It will also follow from more powerful results once we actually dive into graph limits.

Example 3.4. Sparse graphs. Suppose that $\mathsf{e}(G_n)/\mathsf{v}(G)^2 \to 0$, that is, that (G_n) is sparse. If F has at least one edge, we may overcount $\mathsf{hom}(F, G_n)$ by mapping that edge first and the remaining vertices arbitrarily; thus $\mathsf{hom}(F, G_n) \le 2\mathsf{e}(G_n)\mathsf{v}(G)^{\mathsf{v}(F)-2}$. Then

$$t(F, G_n) = \frac{\mathsf{hom}(F, G_n)}{\mathsf{v}(G)^{\mathsf{v}(F)}} \le 2\frac{\mathsf{e}(G_n)}{\mathsf{v}(G)^2} \longrightarrow 0.$$

Since every map from a graph kP_1 with no edges is a homomorphism, the homomorphism density $t(kP_1, G_n)$ is always 1. Therefore any non-dense graph sequence always converges, but always to the same homomorphism densities. So the theory we're going to develop is only useful for dense graph sequences.

[5] Actually there are multiple definitions of graph convergence that are not equivalent. Here's one used in spectral graph theory for sparse graphs: Let $\mathcal{G}_D$ denote the set of simple rooted graphs (finite and infinite) with maximal degree at most D. (Each graph has one specified "'root" vertex.) We let $B_k(G)$ denote the ball of radius k in G centered at the root. A sequence $(G_n) \subseteq \mathcal{G}_D$ is *convergent* if $B_k(G_n)$ is eventually constant for each k. So if n is large enough, the part of G_n near the root is fixed.

3.2 An extended example: Erdős-Rényi random graphs

3.2.1 Expected value

Definition 3.5. The *Erdős-Rényi random graph* $G(n,p)$ is a random variable on the set of labelled graphs with vertex set $\{1,2,\dots,n\}$. The graph $G(n,p)$ contains each edge (not loops) independently with probability p.

The goal of this section is to show that for a fixed $p \in [0,1]$, the sequence $(G(n,p))$ converges with probability 1.[6] Since p is fixed, we will denote the random variable $G(n,p)$ by G_n.

Fix a graph F. We first find the expected value of $t(F,G_n)$. With an eye on Propostion 2.16, we'll evaluate $t_{\text{inj}}(F,G_n)$.[7] We have

$$\mathbb{E}_{G_n}[t_{\text{inj}}(F,G_n)] = \mathbb{E}_{G_n}\left[\mathbb{E}_\phi[\delta(\phi \text{ is a hom.})]\right],$$

where the second expected value is over all injective maps $\phi\colon V(F) \to V(G_n)$. The vertex set of G_n is the same no matter its value, so $\mathbb{E}_\phi$ is a finite sum whose range is independent of G_n. We may therefore switch the order to get

$$\mathbb{E}_{G_n}[t_{\text{inj}}(F,G_n)] = \mathbb{E}_\phi\left[\mathbb{E}_{G_n}[\delta(\phi \text{ is a hom.})]\right] = \mathbb{E}_\phi p^{\mathbf{e}(F)} = p^{\mathbf{e}(F)}.$$

So the expected value of $t_{\text{inj}}(F,G_n)$ converges to $p^{e(F)}$, so the expected value of $t(F,G_n)$ does, as well. To show that $t(F,G_n)$ itself converges to the expected value with probability 1, however, we need to get a handle on how far it can stray from its expected value. So let's do that.

3.2.2 A probability interregnum

Here's a fundamental result from probability theory.

Theorem 3.6 (Markov's inequality)**.** *Let X be a random variable on the positive reals. For every $a > 0$, we have $\mathbb{P}(X \geq a) \leq \frac{1}{a}\mathbb{E}[X]$.*

Proof. Let f be the probability density function of X, so that $\mathbb{E}[X] = \sum_{x=0}^{\infty} x f(x)$. Then

$$\mathbb{E}[X] = \sum_{x=0}^{a} x f(x) + \sum_{x=a}^{\infty} x f(x) \geq a \sum_{x=a}^{\infty} f(x) = a\mathbb{P}(X \geq a).$$

The argument if X is continuous is the same—just replace the sums with integrals.[8] □

And now you know all of probability theory.[9]

Definition 3.7. Let X be a random variable with finite expected value μ. The *variance* of X is $\text{Var}(X) = \mathbb{E}[(X-\mu)^2]$, and the *standard deviation* of X is $\sqrt{\text{Var}(X)}$.

A straightforward calculation using linearity of expectation shows that $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Intuitively, variance measures how much the variable X varies from its mean. The following inequality makes this precise.

[6] We won't. But we'll get close!

[7] Using the other eye, of course.

[8] A measure-theoretic view of random variables unites these two arguments. This perspective is presented in Section 7.3.

[9] That's not true. But sometimes it seems like the rest of probability is one long corollary to Markov's inequality.

Theorem 3.8 (Chebyshev's inequality)**.** *Let X be a random variable with finite mean μ and standard deviation σ, and let $\varepsilon > 0$. Then $\mathbb{P}(|X - \mu| \geq \varepsilon) \leq \sigma^2/\varepsilon^2$.*

Proof. We apply Markov's inequality[10] to the new random variable $(X - \mu)^2$:

$$\mathbb{P}\big((X-\mu)^2 \geq \varepsilon^2\big) \leq \frac{1}{\varepsilon^2}\mathbb{E}[(X-\mu)^2] = \frac{\sigma^2}{\varepsilon^2}$$

and note that $(X - \mu)^2 \geq \varepsilon^2$ if and only if $|X - \mu| \geq \varepsilon$. □

We need one more ingredient from probability.

Theorem 3.9 (Borel–Cantelli Lemma)**.** *Let (A_i) be a sequence of events such that $\sum_{i=1}^{\infty} \mathbb{P}(A_i) < \infty$. Then the probability that infinitely many A_i occur is 0.*

Proof. Let p be the probability that infinitely many A_i occur. Any event which causes infinitely many A_i to occur is in $\bigcup_{i=N}^{\infty} A_i$ for all $N \in \mathbb{N}$. Since $\sum \mathbb{P}(A_i)$ converges,

$$p \leq \mathbb{P}\left(\bigcup_{i=N}^{\infty} A_i\right) \leq \sum_{i=N}^{\infty} \mathbb{P}(A_i) \xrightarrow{N\to\infty} 0,$$

so $p = 0$. □

3.2.3 Variance and convergence

Now we return to the sequence $G_n = G(n, p)$. The expected value of $t(F, G_n)$ over G_n is $p^{\mathrm{e}(F)}$. The variance is

$$\mathrm{Var}(t(F, G_n)) = \mathbb{E}[t(F, G)^2] - \mathbb{E}[t(F, G)]^2.$$

Using Propositions 2.4 and 2.16, we have

$$\begin{aligned}\mathrm{Var}(t(F, G_n)) &= \mathbb{E}[t(FF, G)] - \mathbb{E}[t(F, G)]^2 \\ &\leq \mathbb{E}[t_{\mathrm{inj}}(FF, G)] - \mathbb{E}[t_{\mathrm{inj}}(F, G)]^2 + O\left(\frac{1}{n}\right) \\ &= O\left(\frac{1}{n}\right),\end{aligned}$$

since $\mathbb{E}[t_{\mathrm{inj}}(FF, G)] = \mathbb{E}[t_{\mathrm{inj}}(F, G)]^2 = p^{2\mathrm{e}(F)}$.

Chebyshev's inequality therefore shows that $\mathbb{P}\big(|t(F, G_n) - p^{\mathrm{e}(F)}| > \varepsilon\big) = O\left(1/n\varepsilon^2\right)$. Here we encounter the problem foreshadowed in footnote 6: The sum $\sum \frac{1}{n}$ is divergent, so we cannot apply the Borell-Cantelli Lemma. But that's not too much of a problem. We modify G_n by choosing $G_n = G(\lfloor n^r \rfloor, p)$ for any $r > 1$. The analysis conducted before is exactly the same, and $\sum \lfloor n^{-r} \rfloor$ is convergent, so the Borell-Cantelli Lemma implies that $|t(F, G_n) - p^{\mathrm{e}(F)}| > \varepsilon$ for only finitely many n with probability 1. Therefore (G_n) converges with probability 1. To show that $(G(n, p))$ converges with probability 1 requires better estimates than Chebyshev's inequality provides. We do this in Section 5.2.4.

[10] See?

4 Limit objects

So what should the limit object be?

4.1 Product space

From an analytical perspective, we can come up with one potential limit object pretty quickly. Let (G_n) be a convergent graph sequence. The collection $\mathcal{F}$ of all simple graphs is countable (since there are finitely many graphs on any given number of vertices), so let $\mathcal{F} = \{F_1, F_2, \dots\}$. For each F_i, we have the limit $\lim_{n\to\infty} t(F_i, G_n)$ exists; call it t_i. One way to construct a limit object is to just record these limit densities as an element of the set $[0,1]^{\mathcal{F}} = [0,1]^{\mathbb{N}}$, so that $\lim_{n\to\infty}(G_n) = (t_1, t_2, \dots)$. In fact, by associating each graph G_n with the vector $(t(F_1, G_n), t(F_2, G_n), \dots) \in [0,1]^{\mathbb{N}}$, graph convergence simplifies to coordinate convergence in $[0,1]^{\mathbb{N}}$.

This is one possible answer. Here's why it's a bad one: Not every point of $[0,1]^{\mathbb{N}}$ is a limit point of a graph sequence. This is simply because the coordinates of graph sequences are not independent.

Lemma 4.1. *If $F_1 \subseteq F_2$, then $t(F_1, G) \geq t(F_2, G)$.*

Proof. Divide both sides of Proposition 2.5 by $\mathsf{v}(G)^{\mathsf{v}(F_2)}$. □

Thus, any point of $[0,1]^{\mathcal{F}}$ whose F_1-coordinate is less than its F_2-coordinate is not a limit point. There's an even easier way to see that $[0,1]^{\mathbb{N}}$ is too big of a space: Any point of $[0,1]^{\mathcal{F}}$ whose kK_1 coordinate is not 1 is not a limit point.

But one problem with this is that the limit object does not resemble the terms of the sequence. In particular, it is hard to see any combinatorial structure in the elements of $[0,1]^{\mathbb{N}}$. Indeed, we will devise a different limit object later. However, this structure will yield some useful insights, so we'll continue investigating it for now. There are other relations between coordinates, as well:

Lemma 4.2. $t(F_1 \cup F_2, G) = t(F_1, G)t(F_2, G)$.

Proof. Divide both sides of 2.4 by $\mathsf{v}(G)^{\mathsf{v}(F_1)+\mathsf{v}(F_2)}$. □

So most points of $[0,1]^{\mathbb{N}}$ are not limits points. However, this space does have one redeeming aspect. We can equip $[0,1]^{\mathbb{N}}$ with the distance function $d(x,y) = \sum_{i=1}^{\infty} 2^{-i}|x_i - y_i|$, which turns it into a compact (and therefore complete) metric space. (You can either trust me or look in the appendix for a proof.) Compactness is man's best friend and a stalwart companion in developing limit theories. Whenever we want to prove something about generic graph sequences, we can pass to a convergent subsequence, which is much easier to deal with. We'll see this technique once we develop a better limit object.

4.2 Ultraproducts

We can also take a set-theoretic tack. The method presented in this section is actually quite general, and it can be used to form a limit object for sequences from *any* space. However, it has some drawbacks. It's not constructive (it uses the Axiom of Choice), and it is very abstract. This subsection can be skimmed or even skipped; the next one discusses the limit objects we'll actually end up using.[11] But ultrafilters are fun, and the ultraproduct method is very general. Let's dive in.

[11] Yep, spoiler: This one won't work either. Well, not without significantly more effort.

Definition 4.3. A *filter* $\mathcal{F}$ on a set X is a nonempty collection of subsets of X such that
1. if $A \in \mathcal{F}$ and $B \supseteq A$, then $B \in \mathcal{F}$,
2. if $A, B \in \mathcal{F}$, then $A \cap B \in \mathcal{F}$, and
3. $\emptyset \notin \mathcal{F}$.

A filter $\mathcal{U}$ is called an *ultrafilter* if, moreover, for every set $A \subseteq X$, either $A \in \mathcal{U}$ or $X \setminus A \in \mathcal{U}$.

Any filter cannot contain both A and $X \setminus A$, and any filter on X contains X as an element. Therefore every ultrafilter is contains exactly one of A and $X \setminus A$ for each $A \subseteq X$. Because of this, you can think of an ultrafilter as a perfect voting system when a set X of voters is choosing between two alternatives. Some set A votes for one, and the set $X \setminus A$ votes for the other; the one that's in the ultrafilter wins.

There's one easy way to construct an ultrafilter. Choose any element $a \in X$. The collection of subsets of X that contain a is an ultrafilter, called the *principal ultrafilter generated by* a. A filter is principal if and only if it contains a finite set.

So . . . are there any nonprincipal ultrafilters? Yes. In fact, every infinite set X admits a nonprincipal ultrafilter, but you'll never see it—it hides behind Zorn's Lemma. Fix an infinite set X. We start with the cofinite filter $\mathcal{F} = \{A \subseteq X : X \setminus A \text{ is finite}\}$, which is a filter. Using Zorn's Lemma, we can obtain a maximal filter ω that contains $\mathcal{F}$ and then prove that such a filter is an ultrafilter. Since it contains no finite sets, ω is not principal. If you've seen Zorn's Lemma before, then the argument is rather routine. If you haven't, then it's not. The details aren't so important.

Now we can use ultrafilters to define limits of sequences.

Definition 4.4. Let X be a set and ω a nonprincipal ultrafilter on $\mathbb{N}$. Two sequences $x, y \in X^{\mathbb{N}}$ are ω-equivalent, denoted $x \sim_\omega y$, if $\{i \in \mathbb{N} : x_i = y_i\} \in \omega$. This defines an equivalence relation on $X^{\mathbb{N}}$. The *ω-ultrapoduct* of X is $\prod_\omega X := X^{\mathbb{N}}/\sim_\omega$. That is, the elements of $\prod_\omega X$ are ω-equivalence classes of elements of $X^{\mathbb{N}}$.

If we set X to be the space of finite simple graphs and fix an ultrafilter ω on $\mathbb{N}$, we can define the limit $\lim_{n\to\infty}(G_n)$ simply as the equivalence class of (G_n) in $X^{\mathbb{N}}$.

See how general this is? We need nothing about the structure of the things we're taking limits of. In fact, we could be even more wild and take some set X_α for each α in some infinite index set I, fix a nonprincipal ultrafilter ω on I, and define $\prod_\omega X_i = (\prod_{\alpha\in I} X_\alpha)/\sim_\omega$. Also, it assigns a limit object to *every* sequence, not just convergent ones.

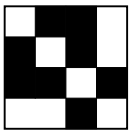
One note: Why do we want ω to be nonprincipal? If ω is the principal ultrafilter generated by a, then $\prod_\omega X$ essentially just selects the coordinate with index a, so it's not good as a limit definition. A nonprincipal ultrafilter takes the whole sequence into account.

Of course, this definition is very abstract and hard to work with, which is why we won't use it. There are other issues, too. For one, the limit depends on the choice of ultrafilter. Let $G_n = K_n$ if n is odd and $G_n = nK_1$ if n is even. The "limit" $\prod_\omega G_n$ is essentially the empty graph if $2\mathbb{N} \in \omega$ and the complete graph if $2\mathbb{N} \notin \omega$. There's a way to remove this dependence on the specific choice of ω, but it's technical and no fun, so we skip it.

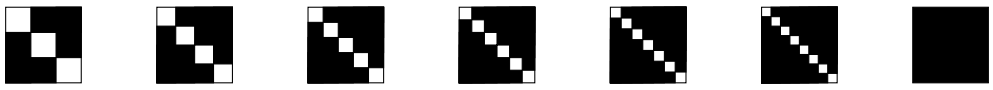
One cool application of this type of construction is ultralimits. Again fix an ultralimit ω on $\mathbb{N}$. We say that a is the ω-ultralimit of $(a_n) \subseteq \mathbb{R}$ if $\{n \in \mathbb{N} : |a_n - a| \leq \varepsilon\} \in \omega$ for every $\varepsilon > 0$. This might seem an odd definition, but replacing ω with the cofinite filter recovers the usual notion of convergence. By upgrading to an ultrafilter, it turns out that *every* bounded sequence has an ultralimit. This has a lot of cool applications which we won't go into here.

4.3 Graphons

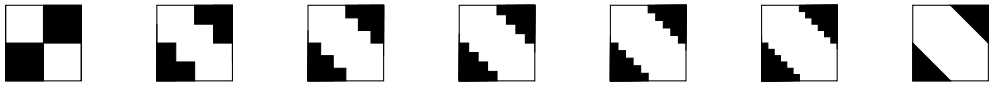
Let's try again. You know what they say: Third time yields a satisfactory limit object. This time the starting point is adjacency matrices. One way to visualize an adjacency matrix is to replace each 0 or 1 with a white or black square, respectively.

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \longrightarrow$$ 

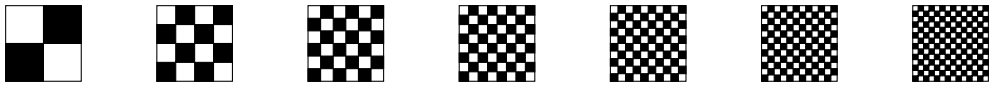
We'll call the picture on the right the *wallpaper* of the graph because it doesn't matter what we call it.[12] If we look at the wallpapers for the sequence (K_n), they seem to converge to the Café noir[13] wallpaper:



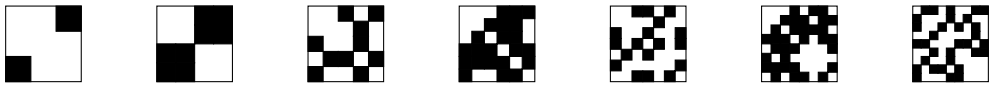
The picture for half-graphs is similar. The wallpapers for (H_n) are shown below, along with their "limit" wallpaper.



Based on these examples, it seems that the adjacency matrices of convergent graph sequences themselves exhibit convergent behavior. But we have to be a bit careful: Adjacency matrices depend on an ordering of the vertices. If we order the vertices in $K_{n,n}$ with $\{1, 2, \dots, n\}$ in one bipartition and $\{n+1, n+2, \dots, 2n\}$ in the other, then the wallpaper looks like ◪ for every n. But if we label one bipartition with even numbers and the other with odd, then the wallpapers look like



This doesn't even look like it converges—the white and black mesh just gets finer, but most points will alternate between black and white infinitely often. The situation is even worse for random graphs. Wallpapers for $(G(n, 1/2))$ look like



We know that $(G(n, 1/2))$ converges, but the corresponding wallpapers don't look like they have any pattern at all. Let's take a step back.

To find a satisfactory limit object, we have to consider a wider class of objects than just graphs. Since convergence is defined in terms of homomorphism densities of graphs, we'll start to do this by generalizing homomorphism density. And this time, we'll be successful!

Given a statement P, the *delta function* $\delta(P)$ evaluates as 1 if P is true and 0 if P is false. We can rewrite the homomorphism density as

$$t(F, G) = \frac{1}{\mathsf{v}(G)^{\mathsf{v}(F)}} \sum_{\phi\colon V(F)\to V(G)} \delta(\phi \text{ is a homomorphism}),$$

[12] This is not a standard term.

[13] An actual shade of black. See wikipedia.org/wiki/Shades_of_black for other good choices.

since the sum counts the number of vertex maps that are homomorphisms. A vertex map is a homomorphism if and only if it preserves the adjacency of each edge in F, so given some $\phi\colon V(F) \to V(G)$,

$$\delta(\phi \text{ is a homomorphism}) = \prod_{ij \in E(F)} \delta\big(\phi(i)\phi(j) \in E(G)\big).$$

In all, then

$$t(F,G) = \frac{1}{\mathsf{v}(G)^{\mathsf{v}(F)}} \sum_{\phi\colon V(F)\to V(G)} \prod_{ij \in E(F)} \delta\big(\phi(i)\phi(j) \in E(G)\big).$$

This all seems fairly abstruse, but the benefit is that this formulation of homomorphism density suggests a generalization. We can interpret the expression

$$\frac{1}{\mathsf{v}(G)^{\mathsf{v}(F)}} \sum_{\phi\colon V(F)\to V(G)}$$

as a Riemann approximation of an integral over a volume 1 set. The expression $\delta\big(\phi(i)\phi(j) \in E(G)\big)$ is just a function of two variables. These observations suggest that we might generalize homomorphism densities in the following way.

Definition 4.5. A *kernel* is a bounded, symmetric, measurable function $W\colon [0,1]^2 \to \mathbb{R}$. The homomorphism density of a finite simple graph F in a kernel W is

$$t(F,W) := \int_{[0,1]^{\mathsf{v}(F)}} \prod_{ij \in E(F)} W(x_i, x_j)\, dx_1 \cdots dx_{\mathsf{v}(F)}.$$

The expression for homomorphism density might look a bit daunting, but here's an interpretation. We think of W as the adjacency matrix for some edge-weighted complete graph on $[0,1]$. Each map $V(F) \to [0,1]$ is weighted by the product of the weights of the edges in its image, and the homomorphism density of F in W is the average weight of all maps $V(F) \to [0,1]$.

So how does this generalize homomorphism density for graphs? Choose any graph G on the vertex set $[n]$. We define the kernel

$$W_G(x,y) = \delta\big(\phi(\lceil nx \rceil)\phi(\lceil ny \rceil) \in E(G)\big).$$

The kernel W_G is a scaled version of the delta function testing $\phi(i)\phi(j) \in E(G)$, where the interval $(\frac{i-1}{n}, \frac{i}{n}]$ corresponds to vertex i. Then

$$\begin{aligned} t(F, W_G) &= \int_{(0,1]^{\mathsf{v}(F)}} \prod_{ij \in E(F)} W_G(x_i, x_j)\, dx_1\, dx_2 \cdots dx_{\mathsf{v}(F)} \\ &= \frac{1}{n^{\mathsf{v}(F)}} \sum_{\phi\colon V(F)\to V(G)} \prod_{ij \in E(F)} \delta\big(\phi(i)\phi(j) \in E(G)\big) \\ &= t(F,G). \end{aligned}$$

The kernel W_G is related to the wallpapers we discussed earlier. If we picture the unit square with the y-axis pointed downward and color the points (x,y) with $W_G(x,y) = 1$ black, then this image is exactly the wallpaper derived from the adjacency matrix of G. The connection between kernels and adjacency matrices provides a combinatorial perspective to this limit object which was missing in the previous two attempts to construct one.

Definition 4.6. A *graphon*[14] is a symmetric measurable function $W\colon [0,1]^2 \to [0,1]$. We denote by $\mathcal{W}$ the set of kernels and by $\mathcal{W}_0$ the set of graphons. Sometimes we'll to consider those kernels that map $[0,1]^2 \to [-1,1]$; we'll denote this set by $\mathcal{W}_1$.

The next theorem shows that graphons are "just right" as limit objects.

Theorem 4.7. *Given any convergent graph sequence (G_n), there is a graphon W so that $t(F,G_n) \to t(F,W)$ for all $F \in \mathcal{F}$. Moreover, every graphon is a limit point in this sense.*

Definition 4.8. We say that a graph sequence (G_n) *converges to the graphon W* if $t(F,G_n) \to t(F,W)$ for every finite simple graph F. More generally, a graphon sequence (W_n) converges to W if $t(F,W_n) \to t(F,W)$ for every finite simple graph F.

In the next section we look at distance notions of graphon convergence, which will remedy the ambiguities present in wallpaper convergence.

We can already determine the limit graphon for several graph sequences.

Example 4.9. From Example 3.2, we have $t(F,K_n) \to 1$ for all $F \in \mathcal{F}$. Since $t(F,\mathbf{1}) = 1$ for all $F \in \mathcal{F}$, the sequence (K_n) converges to the constant-1 graphon. Similarly, Example 3.4 shows that any sparse graph sequence converges to the identically 0 graphon.

Example 4.10. The graph sequence (G_n) converges to the graphon W if and only if (W_{G_n}) does, since the homomorphism densities for the two sequences are the same. Define

$$W(x,y) = \begin{cases} 1 & \max\{x,y\} > \frac{1}{2} \text{ and } \min\{x,y\} < \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Then $W_{K_{n,n}} = W$ for all $n \in \mathbb{N}$ (up to a set of measure 0). This is constant, so $(K_{n,n})$ converges to W.

Example 4.11. Let $G_n = G(n^2,p)$ for some fixed $p \in (0,1)$. From Section 3.2, we have $t(F,G_n) \to p^{\mathsf{e}(F)}$ for all $F \in \mathcal{F}$ with probability 1. Thus (G_n) converges to the graphon $W(x,y) = p$ almost surely.

It is important that the limit graphon is not unique. If $G_n \to W$ and U is a graphon that differs from W on a set of measure 0, then G_n also converges to U. This is not too unusual: In measure theory, we often regard almost-everywhere-equal functions as the same. There are more delicate problems however. For example, let $\{x\}$ denote the fractional part of x. Given a kernel W, we define

$$W_{(n)}(x,y) = W(\{nx\},\{ny\}).$$

Then $t(F,W) = t(F,W_{(n)})$ for every $n \in \mathbb{N}$. We'll be able to address this after we find a good metric for graphons in the next section.

Exercise 4.12. Given a kernel W and a simple graph F, the induced homomorphism density of F in W is

$$t_{\mathrm{ind}}(F,W) = \int_{[0,1]^{\mathsf{v}(F)}} \prod_{ij \in E(F)} W(x_i,x_j) \prod_{ij \in E(\overline{F})} (1 - W(x_i,x_j))\, dx_1 \cdots dx_{\mathsf{v}(F)},$$

where $\overline{F}$ is the complement of F.[15] Show that $t_{\mathrm{ind}}(F,G) \neq t_{\mathrm{ind}}(F,W_G)$ in general, and convince yourself that $t_{\mathrm{ind}}(F,W_G)$ is the proportion of not necessarily injective maps $V(F) \to V(G)$ that preserve both adjacency and non-adjacency.

[14] An elision of "graph function".

[15] That is, $V(F) = V(\overline{F})$ and $ij \in E(\overline{F})$ if and only if $i \neq j$ and $ij \notin E(F)$.

Exercise 4.13. Show that the following formula holds:

$$t_{\mathrm{ind}}(F,W) = \sum_{\substack{F' \supseteq F \\ V(F')=V(F)}} (-1)^{\mathsf{e}(F')-\mathsf{e}(F)} t(F',W).$$

(This just uses the definition of t_{ind}.) Then use Möbius inversion to conclude that

$$t(F,W) = \sum_{\substack{F' \subseteq F \\ V(F')=V(F)}} t_{\mathrm{ind}}(F,W).$$

5 Graph distances

Now that we have a space of limit objects, our next goal is to find a metric that reflects our notion of convergence. Unfortunately, there are lots of plausible metrics on graphs, and most of them aren't good.[16]

5.1 Some first attempts

We could just use the metric from before:

$$d(W_1, W_2) = \sum_{k=1}^{\infty} \frac{|t(F_k, W_1) - t(F_k, W_2)|}{2^k},$$

but there are multiple problems with this. Not least, it depends heavily on the ordering of $\mathcal{F}$. Also, it's awkward to work with.

The set of kernels is also the L^∞ space on $[0,1]^2$ restricted to symmetric functions, so it makes sense to consider using an L^p norm. Here's why that won't work. If a sequence (W_n) converges in L^p for $p > 1$, then it also converges in L^1. From Example 4.11, the sequence $(G(n^2, \frac{1}{2}))$ converges to the constant $1/2$ graphon almost surely. On the other hand, $W_n := W_{G(n^2,1/2)}$ takes values only in $\{0,1\}$, so

$$\left\| W_n - \frac{1}{2} \right\|_1 = \int\limits_{[0,1]^2} \frac{1}{2}\, dx = \frac{1}{2}$$

for all $n \in \mathbb{N}$, which means $\|W_n - 1/2\|_1 \not\to 0$ (which is bad). So L^p norms are out.

In some sense, the L^p norms, which compare two kernels pointwise, are too strict; we need to allow some kind of blurring. One very common way to do this (for example, in probability theory) is called weak convergence.

Definition 5.1. A sequence of measures (μ_n) on a metric space Ω *weakly converges* to the measure μ if for all continuous bounded functions $f\colon \Omega \to \mathbb{R}$, we have $\int_\Omega f\, d\mu_n \to \int_\Omega f\, d\mu$.

Example 5.2. Let $\Omega = [0,1]$ with the standard metric. Set $Q_n = \{\frac{k}{n} \in \mathbb{Q} : 1 \le k < n\}$ and define $\mu_n(A) = \frac{1}{n}|A \cap Q_n|$. Then $\int_\Omega f\, d\mu_n$ is the usual Riemann integral approximation, which converges to $\int_0^1 f\, dx$. Thus (μ_n) converges to the standard measure on $[0,1]$.

Weak convergence does a good job at blurring: Each of the measures μ_n in the previous example is discrete, but the limit is not. Unfortunately, it's too good. Revisiting a problem we had with wallpapers, order the vertices of $K_{n,n}$ so that one bipartition contains all the even labels and the other contains all the odd labels. Then $W_{K_{n,n}}$ is a chessboard, and it weakly converges to the constant $\frac{1}{2}$ graphon! This is of course a problem, since we don't want relabelling to affect convergence, so let's abandon this notion, too.[17]

A good metric will instead come from the graph side of things. One of the most common metrics on graphs is called the *edit distance*. If G_1 and G_2 are two graphs with the same vertex set, then their edit distance is

$$d_1(G_1, G_2) = \frac{|E(G_1) \triangle E(G_2)}{\mathsf{v}(G)^2},$$

[16] In this setting.
[17] Szia!

the number of edges we need to add or delete to turn G_1 into G_2. (Here $\triangle$ denotes the symmetric difference of two sets.) This definition can be extended to two arbitrary graphs, but it suffers from the same problem that the L^p metrics do—it doesn't allow enough blurring.

The good stuff, it turns out, is something called the cut distance, which measures, in some sense, the global similarity between two graphs. We turn to that next.

5.2 The cut distance and cut norm

5.2.1 Graphs

We build up to the cut distance in steps.

Definition 5.3. Let G_1 and G_2 be two graphs on the same vertex set V of size n. For each $S, T \subseteq V$, we define $e_G(S,T) = |\{ij \in E(G) : i \in S \text{ and } j \in T\}|$, the number of edges with one endpoint in S and the other in T. (Here, order of endpoints matters, so $e_G(S,S)$ is twice the number of edges in $G[S]$.) The *labelled cut distance* between G_1 and G_2 is

$$d_\square(G_1, G_2) = \max_{S,T \subseteq V} \frac{|e_{G_1}(S,T) - e_{G_2}(S,T)|}{n^2}.$$

It might seem odd to divide by n^2 instead of $|S||T|$. But this would in fact result in the discrete metric: If $G_1 \neq G_2$, then setting $S = \{u\}$ and $T = \{v\}$ where $uv \in E(G_1)$ and $uv \notin E(G_2)$ gives $d_\square(G_1, G_2) = 1$ (since $d_\square(G_1, G_2) \leq 1$ for any two graphs).

Next we compare two unlabelled graphs, which we do by finding the "optimal overlay".

Definition 5.4. Let G_1 and G_2 be two graphs with n vertices, but with possibly different vertex sets. For a bijection $\phi\colon V(G_1) \to V(G_2)$, we let G_1^ϕ denote the result of relabelling G_1 according to ϕ. (That is, G_1^ϕ has vertex set $V(G_2)$ and $\phi(i)\phi(j) \in E(G_1^\phi)$ if and only if $ij \in E(G_1)$.) We define

$$\hat{\delta}_\square(G_1, G_2) = \min_\phi d_\square(G_1^\phi, G_2),$$

where the minimum ranges over all bijections $\phi\colon V(G_1) \to V(G_2)$.

Now we need to find a way to compare two graphs on a different number of vertices.

Definition 5.5. Let G be a graph and $n \in \mathbb{N}$. The *blowup* of G by n is the graph $G(n)$ with vertex set $V(G) \times [n]$ and edge set $E(G(n)) = \{(i,a)(j,b) : ij \in E(G)\}$.

In other words, we replace each vertex of G by an independent set of size n and then connect the independent sets as in G. We use this to give the definitive definition of cut distance.

Definition 5.6. Let G_1 and G_2 be graphs with n_1 and n_2 vertices, respectively. The *cut distance* between G_1 and G_2 is

$$\delta_\square(G_1, G_2) = \inf_{n_1 N_1 = n_2 N_2} \hat{\delta}_\square(G_1(N_1), G_2(N_2)).$$

If G_1 and G_2 have the same number of vertices, it is not clear (and in fact not true) that $\hat{\delta}_\square(G_1, G_2) = \delta_\square(G_1, G_2)$. It more obvious that for two graphs G_1 and G_2 on the same vertex set $d_\square(G_1, G_2)$ is not necessarily equal to $\hat{\delta}_\square(G_1, G_2)$. In general, we have

$$\delta_\square(G_1, G_2) \leq \hat{\delta}_\square(G_1, G_2) \leq d_\square(G_1, G_2)$$

whenever the individual distances make sense.

Reverse inequalities between $\delta_\square$ and $\hat{\delta}_\square$ are limited. It seems like the two should not be that different, which leads to the following conjecture.

Conjecture 5.7 (Lovász)**.** *There exists an absolute constant $C > 0$ so that $\hat{\delta}_\square(G_1, G_2) \leq C\delta_\square(G_1, G_2)$ for all graphs G_1 and G_2.*

The best known result is far from this: In Theorem 2.3 of [1], the authors proved that

$$\hat{\delta}_\square(G_1, G_2) \leq 32\delta_\square(G_1, G_2)^{1/67}.$$

(Recall that these distances are in the set $[0, 1]$, so $\delta_\square(G_1, G_2)^{1/67}$ is quite large.)

There is in fact another problem: Although $d_\square$ and $\hat{\delta}_\square$ are metrics, $\delta_\square$ is not. If, for example, $G_2 = G_1(2)$, then $\delta_\square(G_1, G_2) = 0$, but clearly $G_1 \ncong G_2$. In fact, this is essentially the example that shows limit graphons are not unique: If $W = W_G$, then $W_{(n)} = W_{G(n)}$ (if the vertices in $G(n)$ are ordered correctly).

Exercise 5.8. Show that $\delta_\square$ satisfies the triangle inequality.

So: Why is the cut distance good? Because convergence with respect to $\delta_\square$ is exactly convergence in homomorphism density.

Theorem 5.9. *The graph sequence (G_n) is convergent if and only if it is Cauchy with respect to $\delta_\square$.*

5.2.2 Norms

The cut distance can be recontextualized in terms of matrix norms.

Definition 5.10. Let A be an $n \times n$ matrix. The *ℓ_1-norm* of A is

$$\|A\|_1 = \frac{1}{n^2} \sum_{i,j=1}^{n} |a_{i,j}|.$$

The *cut norm* of A is

$$\|A\|_\square = \frac{1}{n^2} \max_{S,T \subseteq [n]} \left| \sum_{\substack{i \in S \\ j \in T}} a_{i,j} \right|.$$

It is straightforward to check that both $\|\cdot\|_1$ and $\|\cdot\|_\square$ are, in fact, norms. If we set A_1 and A_2 as the adjacency matrices of two graphs G_1 and G_2 on the vertex set $[n]$, then the edit distance is

$$d_1(G_1, G_2) = \|A_1 - A_2\|_1.$$

Similarly, the labelled cut distance is

$$d_\square(G_1, G_2) = \|A_1 - A_2\|_\square,$$

and the cut distance $\delta_\square$ can be recovered through bijections of $[n]$ and blowups.

Exercise 5.11. Show the following relations between $\|A\|_1$ and $\|A\|_\square$ for any $n \times n$ matrix A. (Find a simpler argument for (2) than (3).)

1. $\|A\|_\square \leq \|A\|_1$
2. $\|A\|_1 \leq n^2\|A\|_\square$
3. $\|A\|_1 \leq 2n\|A\|_\square$

To simplify notation, we denote $\sum_{\substack{i \in S \\ j \in T}} a_{i,j}$ by $A(S, T)$. By restricting the possible sets S and T, we can estimate the cut norm to within a constant factor.

Exercise 5.12. Let A be a symmetric $n \times n$ matrix. For each $\mathcal{S} \subseteq \mathcal{P}([n])^2$, we define $\|A\|_{\mathcal{S}} = \frac{1}{n^2} \max_{(S,T)\in\mathcal{S}} |A(S,T)|$. (Then $\|A\|_\square = \|A\|_{\mathcal{S}}$ when $\mathcal{S} = \mathcal{P}([n])^2$.) We have $\|A\|_{\mathcal{S}} \le \|A\|_\square$ for all $\mathcal{S}$. Prove that

1. $\|A\|_\square \le 4\|A\|_{\mathcal{S}}$ when $\mathcal{S} = \{(S,T) : |S|, |T| \le \lceil n/2 \rceil\}$.
2. $\|A\|_\square \le 2\|A\|_{\mathcal{S}}$ when $\mathcal{S} = \{(S,S) : S \subseteq [n]\}$.
3. $\|A\|_\square \le 4\|A\|_{\mathcal{S}}$ when $\mathcal{S} = \{(S,T) : S \cap T = \emptyset\}$.

5.2.3 Kernels

We extend the cut norm to kernels as follows.

Definition 5.13. The *cut norm* of $W \in \mathcal{W}$ is

$$\|W\|_\square = \sup_{S,T\subseteq[0,1]} \left| \int_{S\times T} W(x,y)\,dxdy \right|,$$

where the supremum is taken over all measurable sets S and T.

The cut norm satisfies the triangle inequality, dilation by real numbers, and is nonnegative. If $\|W\|_\square = 0$, then W is zero almost everywhere; following the conventions of measure theory, we consider two almost-everywhere-equal functions equal. So the cut norm is indeed a norm.

For every measurable $S, T \subseteq [0,1]$, we define $W(S,T) = \int_{S\times T} W$, so $\|W\|_\square = \sup_{S,T} |W(S,T)|$. The *characteristic function* of a set A is $\chi_A(x) = \delta(x \in A)$. We denote the set of kernels $W\colon [0,1]^2 \to [-1,1]$ by $\mathcal{W}_1$. A kernel W is in $\mathcal{W}_1$ if and only if there exist two graphons $W_1, W_2 \in \mathcal{W}_0$ so that $W = W_1 - W_2$.

Definition 5.14. For each $p \ge 1$, the L^p norm of a kernel W is

$$\|W\|_p = \left(\int_{[0,1]^2} |W|^p \right)^{1/p}.$$

Exercise 5.15. Show that $\|W\|_\square \le \|W\|_1 \le \|W\|_2$ for all kernels W. (The second inequality is just Cauchy-Schwarz.)

Definition 5.16. Let A be a symmetric $n \times n$ matrix. The kernel W_A is defined by

$$W_A(x,y) = a_{\lceil nx\rceil,\lceil ny\rceil}.$$

In particular, if A is the adjacency matrix of G, then $W_A = W_G$.

Exercise 5.17. Let $\mathcal{P} = \{V_1, \dots, V_k\}$ be a partition of $[0,1]$ into finitely many measurable sets. Show that for any graphon W that is constant on the sets $V_i \times V_j$,

$$\|W\|_\square = \max\left\{ \left| \int_{S\times T} W(x,y)\,dxdy \right| : S, T \text{ are unions of elements of } \mathcal{P} \right\}.$$

Use this to show that if A is a symmetric square matrix, then $\|A\|_\square = \|W_A\|_\square$.

To get the cut distance from the cut norm, we need to "unlabel" the graphons, which we do with bijections of the "vertex set" $[0,1]$.

Definition 5.18. The *cut distance* between two kernels W_1 and W_2 is

$$\delta_\square(W_1, W_2) = \inf_{\phi\colon [0,1]\to[0,1]} \|W_1^\phi - W_2\|_\square,$$

where $W^\phi(x,y) := W(\phi(x),\phi(y))$ and the infimum is taken over all measure-preserving bijections ϕ. We define the cut distance between a graph G and a kernel W as $\delta_\square(G,W) = \delta_\square(W_G, W)$.

It's important that the cut distances for graphs and their corresponding kernels agree.

Proposition 5.19. *$\delta_\square(G,H) = \delta_\square(W_G, W_H)$ for all finite graphs G and H.*

The proof is rather technical and really no fun, but an outline of it appears in Section A.2 if you really must see it. The following theorem explains the key relationship between cut distance and convergence in homomorphism density in the space of kernels.

Theorem 5.20. *The sequence $(W_n) \subseteq \mathcal{W}$ converges to the kernel $W \in \mathcal{W}$ (that is, $t(F,W_n) \to t(F,W)$ for every graph $F \in \mathcal{F}$) if and only if $\delta_\square(W_n, W) \to 0$.*

As with the cut distance for graphs, the cut distance for kernels is only a pseudometric. This time, we get around it by cheating.

Definition 5.21. Two kernels are *weakly isomorphic* if their cut distance is 0. An equivalence class of weakly isomorphic kernels is called an *unlabelled kernel*, and the set of all unlabelled kernels is denoted $\widetilde{\mathcal{W}}$. The sets $\widetilde{\mathcal{W}_0}$ and $\widetilde{\mathcal{W}_1}$ are obtained from $\mathcal{W}_0$ and $\mathcal{W}_1$ in the same manner.

Voilà! The cut distance is a metric on the space $\widetilde{\mathcal{W}}$. We can alternatively define two kernels U and W to be weakly isomorphic if $t(F,U) = t(F,W)$ for every finite graph F. This is equivalent by Theorem 5.20.

Add note about measure-preserving bijections definition of weak convergence

Exercise 5.22. Show that the kernel cut distance satisfies the triangle inequality.

Exercise 5.23. Let T_n denote the *threshold graph* on the vertex set $[n]$, with $ij \in E(T_n)$ if and only if $i + j \le n$. Define the *threshold graphon* by $W(x,y) = \delta(x + y \le 1)$. Show that $\delta_\square(T_n, W) \to 0$ as $n \to \infty$.

Exercise 5.24. Let H_n denote the half-graph on n vertices. Find a graphon W so that $\delta_\square(H_n, W) \to 0$ as $n \to \infty$.

5.2.4 Cut norm of Erdős-Rényi random graphs

The goal of this section is to show that $(G(n,p))$ converges to the constant p graphon *in the cut distance* with probability 1. To do that, we'll need some more probability.

Proposition 5.25 (Chernoff's bound)**.** *For any real random variable X and $\lambda \in \mathbb{R}$,*

$$\mathbb{P}(X \ge \alpha) \le e^{-\lambda\alpha}\mathbb{E}[e^{\lambda X}].$$

Proof. The exponential function is monotonic, so $\mathbb{P}(X \ge \alpha) = \mathbb{P}(e^{\lambda X} \ge e^{\lambda\alpha})$. Now apply the probabilist's favorite tool—Markov's inequality. □

We first prove the special case $p = 1/2$. The following proof contains the key ideas for the more general theorems later in this section.

Theorem 5.26. *Let $G_n = G(n, 1/2)$ and W be the constant $1/2$-graphon. With probability 1, $\|W_{G_n} - W\|_\square \to 0$ as $n \to \infty$.*

Proof. For each $n \in \mathbb{N}$, let $\{X_{i,j}^{(n)}\}_{1 \le i < j \le n}$ be a collection of independent identically distributed random variables that are 0 or 1 with equal probability. We let A_n be the adjacency matrix of G_n, so that

$$(A_n)_{i,j} = \begin{cases} X_{i,j}^{(n)} & \text{if } i < j \\ 0 & \text{if } i = j \\ X_{j,i}^{(n)} & \text{if } i > j. \end{cases}$$

We will suppress the superscripts on $X_{i,j}$. From Exercise 5.17, we can write $\|W_n - W\|_\square$ as the matrix cut norm $\|A_n - \frac{1}{2}\|_\square$. Let $B_n = A_n - \frac{1}{2}$ and $Y_{i,j} = X_{i,j} - \frac{1}{2}$. We want to show that $\|B_n\|_\square$ is small with high probability. To do that, we need to show that $|B_n(S,S)|$ is small with high probability for each $S \subseteq [n]$. (We can assume the sets are the same by Exercise 5.11.) Fix some $S \subseteq [n]$. Since $\mathbb{E}[X_{i,j}] = 1/2$, we have $\mathbb{E}[B_n(S,S)] = 0$. From the definition,

$$B_n(S,S) = 2 \sum_{\substack{i,j \in S \\ i<j}} Y_{i,j}.$$

Applying Chernoff's bound shows

$$\mathbb{P}(B_n(S,S) \ge \alpha) = \mathbb{P}\left(\sum_{\substack{i,j \in S \\ i<j}} Y_{i,j} \ge \frac{\alpha}{2} \right) \le e^{-\lambda\alpha/2}\, \mathbb{E}\left[\exp\left(\lambda \sum_{\substack{i,j \in S \\ i<j}} Y_{i,j} \right) \right],$$

and since the $Y_{i,j}$ are independent and identically distributed, this is equal to

$$e^{-\lambda\alpha/2} \prod_{\substack{i,j \in S \\ i<j}} \mathbb{E}[\exp(\lambda Y_{i,j})] = e^{-\lambda\alpha/2}\, \mathbb{E}[\exp(\lambda Y_{i,j})]^{\binom{|S|}{2}}.$$

We can explicitly calculate the expected value:

$$\mathbb{E}[\exp(\lambda Y_{i,j})] = \frac{1}{2}(e^{-\lambda/2} + e^{\lambda/2}).$$

For any $t \in \mathbb{R}$,

$$\frac{1}{2}(e^{-t} + e^{t}) = \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!} \le \sum_{k=0}^{\infty} \frac{t^{2k}}{2^k k!} = e^{t^2/2},$$

so in particular $\mathbb{E}[\exp(\lambda Y_{i,j})] \le \exp(\lambda^2/8)$. In all, then,

$$\mathbb{P}\big(B_n(S,S) \ge \alpha\big) \le \exp\left(-\frac{\lambda}{2}\alpha + \frac{\lambda^2}{8}\binom{|S|}{2} \right).$$

Since $\mathbb{P}\big(B_n(S,S) \le -\alpha\big) = \mathbb{P}\big(-B_n(S,S) \ge \alpha\big)$, and $B_n(S,S)$ is symmetric about 0, we have $\mathbb{P}\big(B_n(S,S) \le -\alpha\big) = \mathbb{P}\big(B_n(S,S) \ge \alpha\big)$. In other words,

$$\mathbb{P}\big(|B_n(S,S)| \ge \alpha\big) \le 2\exp\left(-\frac{\lambda}{2}\alpha + \frac{\lambda^2}{8}\binom{|S|}{2} \right).$$

The inequality holds for all $\lambda \in \mathbb{R}$, so we choose λ to minimize the right-hand side. The expression in the exponent is a quadratic, which is minimized at $\lambda = 2\alpha/\binom{|S|}{2}$. Together with the inequality $\binom{|S|}{2} \leq n^2/2$, this yields

$$\mathbb{P}\big(|B_n(S,S)| \geq \alpha\big) \leq 2\exp\left(-\frac{\alpha^2}{2\binom{|S|}{2}}\right) \leq 2\exp\left(-\frac{\alpha^2}{n^2}\right).$$

From part (2) of Exercise 5.12, we have

$$\mathbb{P}\big(\|B_n\|_\square \geq 2n^{-1/2}\big) \leq \mathbb{P}\left(|B_n(S,S)| \geq n^{3/2} \text{ for some } S \subseteq [n]\right),$$

so setting $\alpha = n^{3/2}$ gives

$$\mathbb{P}\big(\|B_n\|_\square \geq 2n^{-1/2}\big) \leq \sum_{S\subseteq[n]} 2e^{-n} = 2\left(\frac{2}{e}\right)^n.$$

Applying the Borel-Cantelli lemma (Theorem 3.9) shows that $\|B_n\|_\square \leq 2n^{-1/2}$ for all but finitely many $n \in \mathbb{N}$ with probability 1. So $\|B_n\|_\square \to 0$ almost surely. □

As a corollary, we can derive that the L^1-norm for kernels cannot be upper-bounded by the cut norm, even for kernels with bounded range.

Corollary 5.27. *There is no absolute constant $C > 0$ so that $\|W\|_1 \leq C\|W\|_\square$ for all kernels $W \in \mathcal{W}_1$.*

Proof. Fix a graph sequence (G_n) with $\|W_{G_n} - \frac{1}{2}\|_\square \to 0$ (which exists by the previous theorem). We set $W_n = W_{G_n} - \frac{1}{2}$. Then $\|W_n\|_1 = \frac{1}{2}$ for all $n \in \mathbb{N}$, but $\|W_n\|_\square \to 0$ as $n \to \infty$, so for any $C > 0$, some W_n with large enough n provides a counterexample. □

In fact, the proof shows that any graph sequence that converges to a graphon that has values other than $\{0, 1\}$ on a set of positive measure provides a counterexample.

The argument used in the previous proof can be generalized.

Theorem 5.28 (Hoeffding's inequality)**.** *Assume that $\{X_1, \ldots, X_k\}$ is an collection of independent random variables such that $a_i \leq X_i \leq b_i$ for each $1 \leq i \leq r$. If $Z = \sum_{i=1}^k X_i$, then for any*

$$\mathbb{P}(|Z - \mathbb{E}[Z]| > \alpha) \leq 2\exp\left(\frac{-\alpha^2}{2\sum_{i=1}^k (b_i - a_i)^2}\right).$$

Proof. We set $Y_i = X_i - \mathbb{E}[X_i]$, so $Z - \mathbb{E}[Z] = \sum_{i=1}^k Y_i$. As in the previous proof, Chernoff's bound and the fact that the Y_i are i.i.d. yields

$$\mathbb{P}\left(\sum_{i=1}^k Y_i \geq \alpha\right) \leq e^{-\lambda\alpha} \prod_{i=1}^k \mathbb{E}[e^{\lambda Y_i}]. \tag{5.1}$$

The main new step is to bound $\mathbb{E}[e^{\lambda Y_i}]$. We do this with a trick. Let $d_i = b_i - a_i$ and $t = (1 + Y_i/d_i)/2$. We can write λY_i as the convex combination $td_i + (1-t)(-d_i)$; since $e^{\lambda x}$ is convex, this gives

$$e^{\lambda Y_i} \leq te^{\lambda d_i} + (1-t)e^{-\lambda d_i} = \frac{1}{2}(e^{\lambda d_i} + e^{-\lambda d_i}) + \frac{Y_i}{2d_i}(e^{\lambda d_i} - e^{-\lambda d_i}).$$

Taking expectations on both sides and noting that $\mathbb{E}[Y_i] = 0$ gives

$$\mathbb{E}[e^{\lambda Y_i}] \leq \frac{1}{2}(e^{\lambda d_i} + e^{-\lambda d_i}) \leq e^{\lambda^2 d_i^2/2}.$$

We can therefore upper bound Equation (5.1) as

$$\mathbb{P}\left(\sum_{i=1}^{k} Y_i \geq \alpha\right) \leq \exp\left(-\lambda\alpha + \frac{\lambda^2}{2}\sum_{i=1}^{k} d_i^2\right).$$

Choosing $\lambda = \alpha / \sum_{i=1}^{k} d_i^2$ gives

$$\mathbb{P}\left(\sum_{i=1}^{k} Y_i \geq \alpha\right) \leq \exp\left(\frac{-\alpha^2}{2\sum_{i=1}^{k} d_i^2}\right).$$

The same argument applies to $\mathbb{P}\left(\sum_{i=1}^{k} Y_i \leq -\alpha\right)$, and adding the two yields the theorem. □

With some more involved calculations, it's possible to improve the bound slightly to

$$\mathbb{P}(|Z - \mathbb{E}[Z]| > \alpha) \leq 2\exp\left(\frac{-2\alpha^2}{\sum_{i=1}^{k}(b_i - a_i)^2}\right).$$

This is the version that is most often cited as Hoeffding's inequality.

Theorem 5.29. *Fix $p \in (0,1)$. Let $G_n = G(n,p)$ and W be the constant p graphon. With probability 1, $\|W_{G_n} - W\|_\square \to 0$ as $n \to \infty$.*

Proof. For $n \in \mathbb{N}$ and $1 \leq i < j \leq n$, we let $X_{i,j}^{(n)}$ be the random variable that is 1 with probability p and 0 with probability $1-p$. We let A_n be the $n \times n$ matrix as defined in the proof of Theorem 5.26 and set $B_n = A_n - p$. Hoeffding's inequality applied to the variables $X_{i,j}^{(n)}$ for $i, j \in S \subseteq [n]$ yields

$$\mathbb{P}(|B_n(S,S)| \leq 2\exp\left(-\frac{\alpha^2}{2\binom{|S|}{2}}\right).$$

From here, the proof is exactly the same as in Theorem 5.26. □

Exercise 5.30. For $p \in (0,1)$ and $\alpha \in \mathbb{R}$, we define $B(n,\alpha,p)$ to be the random bipartite graph with vertex set $\{x_1, \ldots, x_n, y_1, \ldots, y_{\lfloor \alpha n \rfloor}\}$ where each edge $x_i y_j$ appears independently with probability p. Show that for fixed α and p, the sequence $(B(n,\alpha,p))$ converges to the graphon

$$W(x,y) = \begin{cases} p & \max\{x,y\} > \frac{1}{2} \text{ and } \min\{x,y\} < \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

in cut distance with probability 1.

5.2.5 All graphons are limit points

This section outlines a proof that every graphon is the limit of some graph sequence (with respect to the cut distance).

Definition 5.31. A kernel $W\colon [0,1]^2 \to \mathbb{R}$ is a called a *step function* with k steps if there exists a finite partition $\{V_1, \dots, V_k\}$ of $[0,1]$ so that W is constant on each rectangle $V_i \times V_j$. A function $s\colon [0,1]^2 \to \mathbb{R}$ is called *simple* if it can be expressed as a linear combination of characteristic functions of measurable sets, i.e. as $\sum_{i=1}^n a_i \chi_{A_i}$.

Exercise 5.32. Prove the following statements.

1. For every $W \in \mathcal{W}$, there is a sequence (s_n) of simple functions so that $\int |W - s_n|\, dm \to 0$.
2. For every measurable set $A \subseteq [0,1]^2$ and $\varepsilon > 0$, there is a finite collection $R_1, \dots, R_m$ of axis-parallel squares so that $m(A \triangle \bigcup_{i=1}^m R_m) < \varepsilon$. (Hint: It may be helpful to use that $m(A) = \inf\{m(U) : U \supseteq A \text{ and } U \text{ is open}\}$ for every measurable set $A \subseteq [0,1]^2$.)
3. For every symmetric characteristic function χ_A, there is a sequence (W_n) of step functions so that $\int |\chi_A - W_n|\, dm \to 0$.
4. For every symmetric simple function s, there is a sequence (W_n) of step functions so that $\int |s - W_n|\, dm \to 0$.
5. Step functions are dense in $\mathcal{W}$ with respect to the L^1 norm.
6. For every partition $\{V_1, \dots, V_k\}$ of $[0,1]$, there is a measure-preserving bijection $\phi\colon [0,1] \to [0,1]$ (up to a set of measure 0) so that $\phi(V_j)$ is an interval for each $1 \le j \le k$.
7. Let $W\colon [0,1]^2 \to [0,1]$ be a step function with the partition $\{V_1, \dots, V_k\}$ and ϕ be as in part (6). There is a graph sequence (G_n) such that $\|W_{G_n} - W^\phi\|_\square \to 0$. (Use Theorem 5.29 and Exercise 5.30.)
8. For every graphon W there is a graph sequence (G_n) so that $\delta_\square(G_n, W) \to 0$.

We can replace steps (1) – (5) with a quicker proof of (5) that uses Lusin's theorem.

Exercise 5.33. (A special case of) Lusin's theorem states that for any Lebesgue-measurable function $f\colon [0,1]^2 \to \mathbb{R}$ and $\varepsilon > 0$, there exists a continuous function $g\colon [0,1]^2 \to \mathbb{R}$ so that $m(\{x \in [0,1]^2 : f(x) \neq g(x)\}) < \varepsilon$ and $\sup|g(x)| \le \sup|f(x)|$, where m denotes the Lebesgue measure. Use Lusin's theorem to show that for any kernel W, there is a sequence of step functions that converges to W in the L^1 norm. (Recall that every kernel is bounded.)

Alternatively, the fact that step functions are dense in $\mathcal{W}$ with respect to the cut norm follows from the Weak Regularity Lemma (Theorem 7.1), which is a sufficient replacement for step (5).

6 Cut distance and homomorphism density

Our goal for the next two sections is to elucidate the connections between graphs, graphons, homomorphism density, and cut distance. In Section 5.2.5, we showed that every graphon is the limit of some graph sequence. In this section, we connect cut distance to homomorphism density by proving that a graphon sequence (W_n) converges to a graphon W in homomorphism density (that is, $t(F, W_n) \to t(F, W)$ for every finite simple graph F) if and only if it converges in cut distance ($\delta_\square(W_n, W) \to 0$). In Section 7, we prove that every convergent graph sequence has a limit graphon by showing that the space of graphons is compact.

Together, the Counting and Inverse Counting Lemmas prove Theorem 5.20.

6.1 Counting Lemma

The counting lemma gives a bound on the difference of homomorphism densities of two kernels based on their cut distance. To prove it, we need a reformulation of the cut norm.

Lemma 6.1. *For any kernel $W \in \mathcal{W}$,*

$$\|W\|_\square = \sup_{S,T\subseteq[0,1]} \left| \int_{S\times T} W(x,y) \right| = \sup_{f,g\colon [0,1]\to[0,1]} \left| \int_{[0,1]^2} f(x)W(x,y)g(y) \right|.$$

Proof. For every pair $S, T \subseteq [0,1]$, we can set $f = \chi_S$ and $g = \chi_T$, which shows that the left integral is at most the right. Let $H(f,g) = |\int_{[0,1]^2} f(x)W(x,y)g(y)|$. To prove the reverse inequality, choose any $\varepsilon > 0$ and pick a pair of functions $f, g\colon [0,1] \to [0,1]$ so that $H(f,g) > \sup H - \varepsilon$. We define new functions $f_1, f_2\colon [0,1] \to [0,1]$ by

$$f_1(x) = \begin{cases} 1 & \text{if } f(x) > \frac{1}{2} \\ 0 & \text{if } f(x) \le \frac{1}{2} \end{cases} \quad \text{and} \quad f_2(x) = \begin{cases} 2f(x) - 1 & \text{if } f(x) > \frac{1}{2} \\ 2f(x) & \text{if } f(x) \le \frac{1}{2}, \end{cases}$$

and g_1, g_2 similarly. Thus $f = \frac{1}{2}(f_1 + f_2)$ and $g = \frac{1}{2}(g_1 + g_2)$. By the triangle inequality,

$$\sup H - \varepsilon < H(f,g) \le \frac{1}{4}\big(H(f_1,g_1) + H(f_1,g_2) + H(f_2,g_1) + H(f_2,g_2)\big).$$

Each term on the right is at most $\sup H$, so $H(f_1,g_1) > \sup H - 4\varepsilon$. Setting $S = f_1^{-1}(1)$ and $T = g_1^{-1}(1)$, we have

$$\left| \int_{S\times T} W(x,y) \right| = H(f_1,g_1) > \sup H - 4\varepsilon.$$

Because ε was arbitrary, this implies that

$$\left| \int_{S\times T} W(x,y) \right| \ge \sup_{f,g\colon [0,1]\to[0,1]} \left| \int_{[0,1]^2} f(x)W(x,y)g(y) \right|. \qquad \square$$

In fact, both suprema are attained (they are actually maximums). To prove this requires some functional analysis that has been banished to the appendix (see Section A.3).

Theorem 6.2 (Counting Lemma)**.** *Let F be a finite simple graph. For any two graphons $U, W \in \mathcal{W}_0$,*

$$|t(F,U) - t(F,W)| \le \mathsf{e}(F)\delta_\square(U,W).$$

Proof. It suffices to prove that $|t(F,U)-t(F,W)| \le \mathsf{e}(F)\|U-W\|_\square$. Since $t(F,W) = t(F,W^\phi)$ for any measure-preserving bijection $\phi\colon [0,1] \to [0,1]$, replacing W by W^ϕ and taking the infimum over ϕ yields the desired inequality.

The left-hand side expands as

$$\left| \int_{[0,1]^{\mathsf{e}(F)}} \left(\prod_{ij \in E(F)} U(x_i,x_j) - \prod_{ij\in E(F)} W(x_i,x_j) \right) dx \right|. \tag{6.1}$$

To work with this expression, we pull a trick from analysis. In proving that $\lim a_n b_n = ab$ for convergent real sequences $a_n \to a$ and $b_n \to b$, we want to bound the inequality $|a_n b_n - ab|$, which also the absolute value of a difference of products. The solution is to add a "ghost term" $a_n b - a_n b$ and use the triangle inequality to get

$$|a_n b_n - ab| \le a_n |b_n - b| + b|a_n - a|,$$

which we can bound and finish the proof. The same trick works for larger products; we just need to introduce more ghost terms. Order the edges in F as $e_1, \dots, e_{\mathsf{e}(F)}$, where edge e_k has endpoints p_k and q_k. We use this phantasmal[18] trick to rewrite (6.1) as

$$\left| \int_{[0,1]^{\mathsf{e}(F)}} \sum_{k=1}^{\mathsf{e}(F)} \left(\prod_{i=1}^{k} U(x_{p_i},x_{q_i}) \prod_{i=k+1}^{\mathsf{e}(F)} W(x_{p_i},x_{q_i}) - \prod_{i=1}^{k-1} U(x_{p_i},x_{q_i}) \prod_{i=k}^{\mathsf{e}(F)} W(x_{p_i},x_{q_i}) \right) dx \right|.$$

By factoring, using linearity, and applying the triangle inequality, we get

$$|t(F,U)-t(F,W)| \le \sum_{k=1}^{\mathsf{e}(F)} \left| \int_{[0,1]^{\mathsf{e}(F)}} \big(U(x_{p_k},x_{q_k}) - W(x_{p_k},x_{q_k})\big) \prod_{i=1}^{k-1} U(x_{p_i},x_{q_i}) \prod_{i=k+1}^{\mathsf{e}(F)} W(x_{p_i},x_{q_i})\, dx \right|.$$

To finish the proof, we show that each term of this sum is at most $\|U-W\|_\square$. At this point, the main hurdle is notation. To clean it up a bit, for each k we define

$$W_i^{(k)} = \begin{cases} U & \text{if } i \le k-1 \\ W & \text{if } i \ge k+1, \end{cases}$$

so that

$$|t(F,U) - t(F,W)| \le \sum_{k=1}^{\mathsf{e}(F)} \left| \int_{[0,1]^{\mathsf{e}(F)}} \big(U(x_{p_k},x_{q_k}) - W(x_{p_k},x_{q_k})\big) \prod_{i\ne k} W_i^{(k)}(x_{p_i},x_{q_i})\, dx \right|. \tag{6.2}$$

For each $v \in V(F)$, we let $\nabla(v)$ denote the set of indices of edges adjacent to v. We define

$$f_k(x) = \prod_{j \in \nabla(p_k)\setminus\{k\}} W_j^{(k)}(x_{p_j},x_{q_j}) \quad \text{and} \quad g_k(x) = \prod_{j \in E(F)\setminus\nabla(p_k)} W_j^{(k)}(x_{p_j},x_{q_j}).$$

[18] phantastic?

The functions f_k and g_k partition the product $\prod_{i\neq k} W_i^{(k)}(x_{p_i}, x_{q_i})$, so the kth term in the sum of (6.2) becomes

$$\left| \int_{[0,1]^{\mathsf{e}(F)}} f_k(x)\big(U(x_{p_k}, x_{q_k}) - W(x_{p_k}, x_{q_k})\big)g_k(x)\,dx \right|. \tag{6.3}$$

Temporarily fix all variables except x_{p_k} and x_{q_k}. Then f_k is a function of only x_{p_k} and g_k is a function of only x_{q_k}. The integral (6.3) is bounded above by

$$\int_{[0,1]^{\mathsf{e}(F)-2}} \left| \int_{[0,1]^2} f_k(x)\big(U(x_{p_k}, x_{q_k}) - W(x_{p_k}, x_{q_k})\big)g_k(x)\,dx_{p_k}dx_{q_k} \right| dx.$$

The inner integral is bounded above by $\|U - W\|_\square$ by Lemma 6.1. Integrating with respect to the remaining variables shows that each term of (6.2) is at most $\|U - W\|_\square$. □

Corollary 6.3. *If $(W_n) \subseteq \mathcal{W}_0$ converges to a graphon $W \in \mathcal{W}_0$ with respect to $\delta_\square$, then $t(F, W_n) \to t(F, W)$ for every finite graph F.*

In fact, this statement holds for any convergent uniformly bounded sequence of kernels; this is the content of Exercise 6.4.

We proved in Theorem 5.29 that $\delta_\square(G(n,p), p) \to 0$ with probability 1, so the Counting Lemma (finally!) shows that $(G(n,p))$ is almost surely a convergent graph sequence, in the sense that $t(F, G(n,p)) \to p^{\mathsf{e}(F)}$ for every finite graph F with probability 1. Together with Exercises 5.23 and 5.24, the Counting Lemma also shows that the threshold graphs and half-graphs are convergent graph sequences.

Exercise 6.4. Prove that

$$\|W\|_\square \leq \sup_{f,g\colon [0,1]\to[-1,1]} \left| \int_{[0,1]^2} f(x)W(x,y)g(y) \right| \leq 4\|W\|_\square.$$

Use this to show that $|t(F,U) - t(F,W)| \leq 4\mathsf{e}(F)\delta_\square(U,W)$ for every graph F and kernels $U, W \in \mathcal{W}_1$. What if $U, W\colon [0,1]^2 \to [-K, K]$?

Exercise 6.5. A $\mathcal{W}_0$-decorated graph (F, w) is a simple graph F with a map w that assigns a graphon W_e to each edge e. The homomorphism density of a $\mathcal{W}_0$-decorated graph (F, w) is

$$t(F, w) = \int_{[0,1]^{\mathsf{v}(F)}} \prod_{ij\in E(F)} W_{ij}(x_i, x_j)\,dx_1 \cdots dx_{\mathsf{v}(F)}.$$

For example, if $W_e = W$ for every $e \in E(F)$, then $t(F, w) = t(F, W)$. Show that if (F, u) and (F, w) are two $\mathcal{W}_0$-decorated graphs (with the same underlying graph) that

$$|t(F, u) - t(F, w)| \leq \mathsf{e}(F) \sum_{e\in E(F)} \|U_e - W_e\|_\square.$$

(This only requires a small modification to the proof of the Counting Lemma.)

Exercise 6.6. Prove that for every pair $U, W \in \mathcal{W}_0$,

$$|t_{\mathrm{ind}}(F, U) - t_{\mathrm{ind}}(F, W)| \leq \binom{\mathsf{v}(F)}{2} \|U - W\|_\square.$$

(See Exercise 4.12.)

6.2 Inverse Counting Lemma

Now we want to show that two graphons with similar homomorphism densities are close in the cut distance.

Theorem 6.7 (Inverse Counting Lemma)**.** *Assume $U, W \in \mathcal{W}_0$. If $|t(F,U) - t(F,W)| \le 2^{-k^2}$ for every simple graph F with k vertices, then*

$$\delta_\square(U,W) \le \frac{50}{\sqrt{\log k}}.$$

The proof of this relies on the Second Sampling Lemma in Section 8 and is a bit technical, so we omit it. The key consequence of the Inverse Counting Lemma is the following.

Corollary 6.8. *Let $(W_n) \subseteq \mathcal{W}_0$ and $W \in \mathcal{W}_0$. If $t(F,W_n) \to t(F,W)$ for every finite simple graph F, then $\delta_\square(W_n, W) \to 0$.*

With the (Inverse) Counting Lemmas in hand, we can freely switch between convergence with respect to $\delta_\square$ and with respect to homomorphism density.

Problem 6.9. Find a proof of the Inverse Counting Lemma that does not rely on the Second Sampling Lemma. (I don't know of any, but I'd love to see one!)

7 Compactness of graphon space

7.1 Regularity partitions

As a consequence of Exercise 5.32, every kernel can be approximated arbitrarily well by step functions. Kernels cannot, however, be *uniformly* approximated by step functions. In fact, even graphons cannot be uniformly approximated. By this, we mean that the error in a step function approximation of a kernel depends not only on the number of steps, but also the kernel itself. In contrast, uniform approximation is possible in the cut norm.

Theorem 7.1 (Weak Regularity Lemma for Kernels)**.** *For every kernel W and $k \in \mathbb{N}$ there is a step function U with k steps so that*

$$\|W - U\|_\square \leq \frac{2}{\sqrt{\log_2 k}} \|W\|_2.$$

Since $\|W\|_2 \leq 1$ when $W \in \mathcal{W}_0$, this means that graphons (or any other subset of $\mathcal{W}$ with bounded range) are uniformly approximable by step functions. We prove the theorem using the following lemma.

Lemma 7.2. *For any $W \in \mathcal{W}$ and $k \in \mathbb{N}$ there exist $2k$ sets $S_1, \dots, S_k, T_1, \dots, T_k \subseteq [0,1]$ and k real numbers $a_1, \dots, a_k$ so that*

$$\left\| W - \sum_{i=1}^{k} a_i \chi_{S_i \times T_i} \right\| \leq \frac{1}{\sqrt{k}} \|W\|_2.$$

To make the proof easier to follow, we will use the result of Proposition A.7. Exercise 7.3 walks through a way to prove a result that's almost as strong (and good enough for our purposes) without resorting to functional analysis mumbo-jumbo.

Proof of Lemma 7.2. Let $S, T \subseteq [0,1]$ so that $\|W\|_\square = |\int_{S\times T} W|$. Since $\|W\|_\square = \|-W\|_\square$, we may assume that $\int_{S\times T} W$ is nonnegative. For every $a \in \mathbb{R}$,

$$\begin{aligned}
\|W - a\chi_{S\times T}\|_2^2 &= \int_{[0,1]^2} (W(x,y) - a\chi_{S\times T}(x,y))^2 \, dxdy \\
&= \|W\|_2^2 - 2b \int_{S\times T} W \, dxdy + b^2 m(S)m(T) \\
&= \|W\|_2^2 - 2b\|W\|_\square + b^2 m(S)m(T),
\end{aligned}$$

where m denotes the Lebesgue measure. This expression is minimized at $b = W(S,T)/m(S)m(T)$, where

$$\|W - b\chi_{S\times T}\|_2^2 = \|W\|_2^2 - \frac{1}{m(S)m(T)} \|W\|_\square^2 \leq \|W\|_2^2 - \|W\|_\square^2.$$

Now set $W_1 = W$ and $b_1 = b$, as well as $S_1 = S$ and $T_1 = T$. We inductively define $W_i = W_{i-1} - b_i \chi_{S_i \times T_i}$, where b_i, S_i, and T_i are chosen so that $\|W_i\|_2^2 \leq \|W_i\|_2^2 - \|W_i\|_\square^2$, which we just showed can always be found. Recursively expanding the inequality for $i = k$ gives

$$\left\| W - \sum_{i=1}^{k} b_i \chi_{S_i \times T_i} \right\|_2^2 \leq \|W\|_2^2 - \sum_{i=1}^{k} \|W_i\|_\square^2.$$

The left-hand side is nonnegative, so $\sum_{i=1}^k \|W_i\|_\square^2 \le \|W\|_2^2$. In particular, there is an $m \in \{1, 2 \dots, k\}$ so that $\|W_m\|_\square^2 \le \frac{1}{k}\|W\|_2^2$. Defining $a_i = b_i$ for $i \le m$ and $a_i = 0$ for $i > m$ finishes the proof, since $W_m = W - \sum_{i=1}^k a_i \chi_{S_i \times T_i}$. □

Now we can prove the theorem.

Proof of Theorem 7.1. For $k = 1$, the theorem is nonsensical; for $2 \le k \le 4$, taking $U = 0$ works. For the rest of the proof we assume that $k \ge 5$. Set $r = \lfloor (\log_2 k)/2 \rfloor$ and apply Lemma 7.2 to obtain sets $S_1, \dots, S_r, T_1, \dots, T_r$ and real numbers $a_1, \dots, a_r$ so that $\|W - \sum_{i=1}^r a_i \chi_{S_i \times T_i}\|_\square \le r^{-1/2}\|W\|_2$. We set $U = \frac{1}{2}\sum_{i=1}^r a_i(\chi_{S_i \times T_i} + \chi_{T_i \times S_i})$, which is a symmetric function that satisfies

$$\|W - U\|_\square \le \frac{1}{2}\left\|W - \sum_{i=1}^r a_i \chi_{S_i \times T_i}\right\|_\square + \frac{1}{2}\left\|W - \sum_{i=1}^r a_i \chi_{T_i \times S_i}\right\|_\square \le \left\|W - \sum_{i=1}^r a_i \chi_{S_i \times T_i}\right\|_\square,$$

since W is symmetric.

The sets $\bigcap_{i=1}^{2r} R_i$ where $R_i \in \{S_i, S_i^c\}$ and $R_{r+i} \in \{T_i, T_i^c\}$ for $1 \le i \le r$ form a partition of $[0,1]$, and U is constant on the product of any two of them. Thus, U is a step function with $2^{2r} \le k$ steps. Since $\lfloor x/2 \rfloor \ge x/4$ for all $x \ge 2$ and $k > 4$, we have

$$\|W - U\|_\square \le \frac{1}{\sqrt{\lfloor (\log_2 k)/2 \rfloor}}\|W\|_2 < \frac{2}{\sqrt{\log_2 k}}\|W\|_2. \qquad \square.$$

It can be shown that

$$\left\lfloor \frac{\log_2 k}{2} \right\rfloor \ge \frac{\log_2 k}{\log_2(15)}$$

for all natural numbers $k \ge 4$. This improves the constant in Theorem 7.1 from 2 to $\sqrt{\log_2 15}$.[19]

Exercise 7.3. This exercise proves a version of Theorem 7.1 for graphons without any functional analysis hocus-pocus.

1. Without assuming that the cut norm is attained by a pair of sets, prove the following approximate version of Lemma 7.2: For every kernel $W \in \mathcal{W}$, natural number $k \in \mathbb{N}$, and $\varepsilon > 0$, there exist $2k$ sets $S_1, \dots, S_k, T_1, \dots, T_k \subseteq [0,1]$ and k real numbers $a_1, \dots, a_k$ so that
$$\left\|W - \sum_{i=1}^k a_i \chi_{S_i \times T_i}\right\|_\square \le \frac{1}{\sqrt{k}}\|W\|_2 + \varepsilon.$$
2. Use part (1) to prove that for every kernel $W \in \mathcal{W}_1$ and $k \in \mathbb{N}$ there exists a step function U with k steps so that $\|W - U\|_\square \le 2(\log_2 k)^{-1/2}$.

One easy way to create a step function from a graphon is to average over the steps.

Definition 7.4. Let $\mathcal{P} = \{V_1, \dots, V_k\}$ be a partition of $[0,1]$ into sets with positive measure. The *stepping* of a kernel W with respect to $\mathcal{P}$ is kernel $W_\mathcal{P}$ defined by

$$W_\mathcal{P}(x,y) = \frac{1}{m(V_i)m(V_j)} \int\limits_{V_i \times V_j} W(x,y)\,dxdy \qquad \text{when } x \in V_i \times V_j.$$

Exercise 7.5. Define $\langle U, W \rangle = \int_{[0,1]^2} U(x,y)W(x,y)\,dxdy$ for all kernels $U, W \in \mathcal{W}$. Show that $\langle U_\mathcal{P}, W \rangle = \langle U_\mathcal{P}, W_\mathcal{P} \rangle = \langle U, W_\mathcal{P} \rangle$.

[19] For reference, $\sqrt{\log_2 15} \approx 1.977$ and $2 \approx 2.000$. The gain is small, but it's helpful.

One way to think about $W_{\mathcal{P}}$ is as the conditional expectation of W relative to $\mathcal{P}$. We can also think of it analytically as follows.

The set of step functions with steps in $\mathcal{P}$ forms a finite-dimensional subspace of $\mathcal{W}$. From Exercise 7.5,

$$\langle W_{\mathcal{P}}, W - W_{\mathcal{P}} \rangle = 0,$$

so $W - W_{\mathcal{P}}$ is orthogonal to W. That is, $W_{\mathcal{P}}$ is the orthogonal projection of W onto the subspace of step functions with steps in $\mathcal{P}$. It follows from the Pythagorean theorem that

$$\|W_{\mathcal{P}}\|_2^2 = \|W\|_2^2 - \|W - W_{\mathcal{P}}\|_2^2 \leq \|W\|_2^2,$$

so $\|W_{\mathcal{P}}\|_2 \leq \|W\|_2$. We say that stepping is *contractive* with respect to the L^2 norm.

Exercise 7.6. Show that stepping is contractive with respect to the L^1 and cut norms, that is, that $\|W_{\mathcal{P}}\|_1 \leq \|W\|_1$ and $\|W_{\mathcal{P}}\|_\square \leq \|W\|_\square$. (Exercise 5.17 may be useful.) Also show that $\|W - W_{\mathcal{P}}\|_\square \leq \|W\|_2$.

So, a question: Is the optimal step function approximation of a kernel always a stepping of that kernel? No, but it's always close.

Exercise 7.7. Find a kernel $W \in \mathcal{W}_1$ and a $k \in \mathbb{N}$ so that the step function with k steps that best approximates W in the cut norm is not a stepping of W.

Proposition 7.8. *If U is a step function with steps in $\mathcal{P}$ and W is any kernel, then*

$$\|W - W_{\mathcal{P}}\|_\square \leq 2\|W - U\|_\square.$$

Proof. Using the triangle inequality and that $U_{\mathcal{P}} = U$, we have

$$\|W - W_{\mathcal{P}}\|_\square \leq \|W - U\|_\square + \|U - W_{\mathcal{P}}\|_\square = \|W - U\|_\square + \|(U - W)_{\mathcal{P}}\|_\square$$

Applying contractivity completes the proof. □

Combining Proposition 7.8 with Theorem 7.1 proves the following result.

Corollary 7.9. *For every kernel W and $k \in \mathbb{N}$, there is a partition $\mathcal{P}$ of $[0,1]$ into at most k sets so that*

$$\|W - W_{\mathcal{P}}\|_\square \leq \frac{4}{\sqrt{\log_2 k}} \|W\|_2.$$

Corollary 7.9 can be strengthened as follows.

Theorem 7.10. *For every partition $\mathcal{Q}$ of $[0,1]$ into m sets and $k \geq m$, there is a partition $\mathcal{P}$ of $[0,1]$ into at most k sets that refines $\mathcal{Q}$ such that*

$$\|W - W_{\mathcal{P}}\|_\square \leq \frac{4}{\sqrt{\log_2(k/m)}} \|W\|_2.$$

Proof. By Theorem 7.1 (and the remark that follows the proof) there is a step function U on a partition $\mathcal{R}$ with $\lfloor k/m \rfloor$ steps that refines $\mathcal{Q}$ such that

$$\|W - U\|_\square \leq \frac{\sqrt{\log_2(15)}}{\sqrt{\log_2(\lfloor k/m \rfloor)}} \|W\|_2.$$

The theorem is trivial if $k/m \le 2^{16}$. If $k/m \ge 2^{16}$, then

$$\log_2(\lfloor k/m \rfloor) \ge \log_2(k/m - 1) \ge \frac{\log_2(2^{16})}{\log_2(2^{16}-1)} \log_2(k/m),$$

from which it follows that $\|W - U\|_\square \le 2(\log_2(k/m))^{-1}\|W\|_2$.

The step function U is also a step function on the common refinement of $\mathcal{Q}$ and $\mathcal{R}$, defined as $\mathcal{P} = \{Q_i \cap R_i : Q_i \in \mathcal{Q} \text{ and } R_i \in \mathcal{R}\}$. It has at most $m\lfloor k/m \rfloor \le k$ sets. By Proposition 7.8,

$$\|W - W_\mathcal{P}\|_\square \le 2\|W - U\|_\square \le \frac{4}{\sqrt{\log_2(k/m)}}\|W\|_2. \qquad \square$$

The astute reader may have noticed the modest number 2^{16} in the proof. Indeed, this theorem (and other versions of the Regularity Lemma) are mainly of theoretical interest: To guarantee a partition $\mathcal{P}$ so that $\|W - W_\mathcal{P}\|_\square < \varepsilon$ for all graphons W, we need to allow $2^{16/\varepsilon^2}$ sets, quite a large number. But fret not—we won't be troubling ourselves with any practical consequences of the Regularity Lemma.

7.2 Szemerédi's Regularity Lemma

The Weak Regularity Lemma has a graph analogue. To state it, we need to introduce weighted graphs.

Definition 7.11. A *weighted graph* on the vertex set V is pair (α, β), where $\alpha = (\alpha_i)_{i\in V}$ is a real vector of *vertex weights* and $\beta = (\beta_{i,j})_{i,j\in V}$ is a real symmetric matrix of *edge weights*.

We think of a weighted graph H as the complete graph with loops at each vertex with a weight α_i assigned to the vertex i and $\beta_{i,j}$ assigned to the edge ij. When $\alpha = \mathbf{1}$, we call H an *edge-weighted graph* and forget about the vertex labels. For edge-weighted graphs, the matrix $(\beta_{i,j})$ is called the *(weighted) adjacency matrix*.

Definition 7.12. The labelled cut distance between two weighted graphs $H = (\alpha, \beta)$ and $H' = (\alpha', \beta')$ on the same vertex set V is

$$d_\square(H, H') = \sum_{i\in V} |\alpha_i - \alpha_i'| + \max_{S,T\subseteq V} \left| \sum_{i,j\in V} (\alpha_i\alpha_j\beta_{i,j} - \alpha_i'\alpha_j'\beta_{i,j}') \right|.$$

This defines a metric on weighted graphs. By considering an ordinary graph G an edge-labelled graph with $\beta_{i,j} = 1$ if $ij \in E(G)$ and $\beta_{i,j} = 0$ otherwise, this metric specializes as the previously-defined $d_\square$ metric for graphs. This metric extends to a $\delta_\square$ metric as before.[20] If A_1 and A_2 are the adjacency matrices of two edge-weighted graphs H_1 and H_2, respectively, then $d_\square(H_1, H_2) = \|A_1 - A_2\|_\square$.

Definition 7.13. Let G be a graph. For any subsets $X, Y \subseteq V(G)$, we let $e_G(X, Y) = |\{(u, v) \in E(G) : u \in X \text{ and } v \in Y\}|$ (order matters: if $u, v \in X \cap Y$, then edge uv is counted twice) and $d_G(X, y) = e_G(X, Y)/|X||Y|$. Given a partition $\mathcal{P} = \{V_1, \dots, V_k\}$ of $V(G)$, we create the edge-weighted graph $G_\mathcal{P}$ on $V(G)$ with $\beta_{u,v} = d_G(V_i, V_j)$ when $u \in V_i$ and $v \in V_j$.

[20] The blowup of a weighted graph is defined by assigning the cloned vertices the same vertex and edge weights as the vertex from which they were cloned.

If A is the adjacency matrix of G, then the adjacency matrix of $G_{\mathcal{P}}$ has entries

$$(A_{\mathcal{P}})_{u,v} = \frac{1}{|V_i||V_j|} \sum_{\substack{x \in V_i \\ y \in V_j}} a_{x,y} \qquad \text{where } u \in V_i \text{ and } v \in V_j.$$

We use this equation as the definition for all matrices, even when A is not an adjacency matrix.

Exercise 7.14. Show that $\|A_{\mathcal{P}}\|_{\square} \leq \|A\|_{\square}$ for every square matrix A.

The graph $G_{\mathcal{P}}$ encodes an average over the partition elements of $\mathcal{P}$. In fact, if we identify the elements of $[n]$ with the intervals of length $1/n$ in $[0,1]$, the partition $\mathcal{P}$ of $[n]$ induces a corresponding $\mathcal{Q}$ of $[0,1]$ into intervals, and $W_{G_{\mathcal{P}}} = W_{\mathcal{Q}}$.

There is a similar but distinct way to construct an "average" weighted graph. For a partition $\mathcal{P} = \{V_1, \ldots, V_k\}$ of $V(G)$, we define the weighted graph $G/\mathcal{P}$ on vertex set $[k]$ with $\alpha_i = \|V_i\|/\mathsf{v}(G)$ and $\beta_{i,j} = d_G(V_i, V_j)$.

Exercise 7.15. Show that $d_{\square}(G, G/\mathcal{P}) \leq d_{\square}(G, G_{\mathcal{P}})$.

The graph $G/\mathcal{P}$ encodes the graph G with some error. The Weak Regularity Lemma for Graphs says that we can always choose a partition $\mathcal{P}$ so that the error is (relatively) small.

Theorem 7.16 (Weak Regularity Lemma for Graphs)**.** *For every graph G and $k \geq 1$, there is a partition $\mathcal{P}$ with k elements so that*

$$d_{\square}(G, G_{\mathcal{P}}) \leq \frac{4}{\sqrt{\log_2 k}}.$$

Exercise 7.17. Derive the Weak Regularity Lemma for Graphs from the Weak Regularity Lemma for Kernels.[21] (Hint: show that you can always choose a step function U whose steps are unions of intervals $[k/\mathsf{v}(G), (k+1)/\mathsf{v}(G)]$.)

Since $d_{\square}(G, G/\mathcal{P}) \leq d_{\square}(G, G_{\mathcal{P}})$, this means that we can approximate G with a weighted graph of bounded size. The benefit is that the approximation is uniform—the size of the approximating graph is independent of the size of G. The trade-off, as before, is that to approximate G with an error of ε, we need $2^{4/\varepsilon^2}$ sets. But pish-posh. Finite is finite.

Why is it called the "Weak" Regularity Lemma? It's not as an insult to its physical prowess. Rather, a different version of the Regularity Lemma was proven first with a stronger bound.

Definition 7.18. A bipartite graph G with bipartitions A and B is *ε-homogeneous* if

$$|e_G(X,Y) - d_G(A,B)|X||Y| \leq \varepsilon|A||B|$$

for all $X \subseteq A$ and $Y \subseteq B$.

So a bipartite graph is ε-homogeneous if its edges are fairly uniformly distributed, meaning that it can be well-approximated by just the average density $d_G(A,B)$.

Theorem 7.19 (Szemerédi's Regularity Lemma)**.** *For every $\varepsilon > 0$ there is a natural number $N(\varepsilon)$ such that every graph G has an equitable partition[22] $\{V_1, \ldots, V_k\}$ of $V(G)$ with $1/\varepsilon \leq k \leq N(\varepsilon)$ such that for all but εk^2 pairs of indices $1 \leq i < j \leq k$, the bipartite graph $G[V_i, V_j]$ is ε-homogeneous.*

[21] The graph version can also be proven by essentially copying the proof for kernels.

[22] Meaning the size of each partition element is either $\lfloor \mathsf{v}(G)/k \rfloor$ or $\lceil \mathsf{v}(G)/k \rceil$.

So both versions of the Regularity Lemma give a uniform approximation for all graphs. Szemerédi's version is more powerful—a cut distance bound can be derived that is much stronger than in the Weak Regularity Lemma. On the other hand, the constant $N(\varepsilon)$ in Szemerédi's lemma is truly astronomical: a power tower $2^{2^{2^{\cdot^{\cdot^{\cdot}}}}}$ of height $1/\varepsilon^2$. It might be difficult not to tremble at its magnitude, but take heart and recall: Pish-posh. Finite is finite.[23]

7.3 Random variables

7.3.1 Martingales

A *martingale* is a safety device used in fencing to prevent a foil from being accidentally thrown during disarmament. It is fashioned as a tape or leather loop attached to the grip of the foil.

Wait, no, um, a *martingale* is a type of tack used by equestrians to control the head carriage of a horse.

Well. This is embarrassing. It seems a martingale is also a concept in probability. To discuss it, we'll need to be a bit more formal with probability theory than we have been. (This section on random variables can be skipped if you're willing to accept Theorem 7.27 as a blackbox.)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. (That is, $\mathcal{F}$ is a σ-algebra on Ω and $\mathbb{P}$ is a measure on $(\Omega, \mathcal{F})$ with $\mathbb{P}(\Omega) = 1$.) A *random variable* on Ω is a measurable function $X \colon \Omega \to \mathbb{R}$. We call X a *finite* random variable if its image is finite.

For any $A \in \mathcal{F}$, we write $X \in A$ to denote the set $X^{-1}(A) = \{x \in \Omega : X(x) \in A\}$ and $X = a$ to denote $X \in \{a\}$. The *expected value* of X is $\mathbb{E}[X] = \int_\Omega X \, d\mathbb{P}$. For measurable sets $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, the *conditional probability* of A with respect to B is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

We think of $\mathbb{P}(A \mid B)$ as the probability of selecting a point in A if we already know that the point is in B. The *conditional expected value* of X with respect to $B \in \mathcal{F}$ is $\mathbb{E}[X \mid B] = (\int_B X \, d\mathbb{P})/\mathbb{P}(B)$. From the definition, both expected value and conditional expected value are linear in the random variable X.

Suppose that X is a finite random variable with range $\{x_1, \ldots, x_k\}$. Then $\{X^{-1}(x_i)\}_{i=1}^k$ is a partition of Ω, and the expected value formula simplifies to the familiar formula

$$\mathbb{E}[X] = \sum_{i=1}^{k} x_i \mathbb{P}(X = x_i).$$

Conditional expected value similarly reduces to

$$\mathbb{E}[X \mid B] = \sum_{i=1}^{k} x_i \mathbb{P}(X = x_i \mid B).$$

Let $\mathcal{G} = \{G_i\}_{i=1}^n$ be a finite partition of Ω into measurable sets. The *conditional expectation of X with respect to $\mathcal{G}$* is the random variable obtained by averaging X over each set G_i:

$$\mathbb{E}[X \mid \mathcal{G}] = \sum_{i=1}^{n} \mathbb{E}[X \mid G_i] \, \chi_{G_i}. \tag{7.1}$$

[23] And anyway, we won't need it. We'll only use the Weak Regularity Lemma.

So $\mathbb{E}[X \mid \mathcal{G}]$ is a function $\Omega \to \mathbb{R}$ that is constant on the generating sets G_i of $\mathcal{G}$, and the value of $\mathbb{E}[X \mid \mathcal{G}]$ on G_i is the average value of X over G_i. If Y is a finite random variable with range $\{y_1, \dots, y_n\}$, then $\mathcal{G}_Y = \{Y^{-1}(y_i)\}_{i=1}^n$ is the partition of Ω *generated* by Y, and we set

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X \mid \mathcal{G}_Y] = \sum_{i=1}^{n} \mathbb{E}[X \mid Y = y_i]\, \chi_{Y^{-1}(y_i)}.$$

For the remainder of the section, $\mathcal{G}$ and $\mathcal{H}$ represent finite partitions of Ω. We say that $\mathcal{G}$ refines $\mathcal{H}$ when each element of $\mathcal{H}$ is a union of elements of $\mathcal{G}$. The *common refinement* of $\mathcal{G}$ and $\mathcal{H}$ is the partition $\mathcal{G} \vee \mathcal{H} = \{G \cap H : G \in \mathcal{G} \text{ and } H \in \mathcal{H}\}$, and it is the partitions with the fewest number of sets that refines both $\mathcal{G}$ and $\mathcal{H}$. The common refinement of any finite number of partitions is defined inductively in the natural way.

Lemma 7.20. *If $\mathcal{G}$ refines $\mathcal{H}$, then $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}] = \mathbb{E}[X \mid \mathcal{H}]$. In particular, $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X]$.*

Proof. Set $Y = \mathbb{E}[X \mid \mathcal{G}]$. It suffices to show that $\mathbb{E}[Y \mid H] = \mathbb{E}[X \mid H]$ for every $H \in \mathcal{H}$. Choose some $H \in \mathcal{H}$; there are disjoint sets $\{G_i\}_{i=1}^k$ in $\mathcal{G}$ so that $H = \bigcup_{i=1}^k G_i$. Since $\mathbb{E}[Y \mid G] = \mathbb{E}[X \mid G]$ for each $G \in \mathcal{G}$, we have

$$\mathbb{P}(H)\, \mathbb{E}[X \mid H] = \int_H X \, d\mathbb{P} = \sum_{i=1}^{k} \int_{G_i} X \, d\mathbb{P} = \sum_{i=1}^{k} \int_{G_i} Y \, d\mathbb{P} = \int_H Y \, d\mathbb{P} = \mathbb{P}(H)\, \mathbb{E}[Y \mid H],$$

which proves the first part of the lemma. Taking $\mathcal{H} = \{\Omega\}$ proves the second part. □

We say that a random variable Z is *constant on* $\mathcal{G}$ if it is constant on every partition element of $\mathcal{G}$.

Lemma 7.21. *If X and Z are random variables and Z is constant on $\mathcal{G}$, then $\mathbb{E}[ZX \mid \mathcal{G}] = Z\mathbb{E}[X \mid \mathcal{G}]$.*

Proof. This is an application of linearity: If $Z(x) = a_i$ for every $x \in G_i$, then $\mathbb{E}[ZX \mid G_i] = a_i \mathbb{E}[X \mid G_i]$. Therefore $\mathbb{E}[ZX \mid G_i]\, \chi_{G_i} = Z\mathbb{E}[X \mid G_i]\, \chi_{G_i}$; then insert this in (7.1) and factor out Z. □

With the preliminaries done, now we can get down to business.

Definition 7.22. A *filtration* on a measurable space $(\Omega, \mathcal{F})$ is a sequence $(\mathcal{G}_i)$ of partitions of Ω into measurable sets such that $\mathcal{G}_{i+1}$ refines $\mathcal{G}_i$ for every $i \in \mathbb{N}$.

And now we can define probability's martingale.[24]

Definition 7.23. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{G}_i)$ be a filtration on Ω. A sequence (X_n) of random variables on Ω is a *martingale* with respect to $(\mathcal{G}_i)$ if, for each $n \in \mathbb{N}$,

1. X_n is constant on $\mathcal{G}_n$ and
2. $\mathbb{E}[X_n \mid \mathcal{G}_{n-1}] = X_{n-1}$.

[24] This definition is somewhat different from the usual definition because we are only considering finite partitions. Doing so makes the presentation easier—in the general case, it's not clear from the definition that the conditional expectation of a random variable necessarily exists! But with a finite partition, these problems don't appear.

Example 7.24. If $(X_n)_{n=0}^{\infty}$ is a collection of independent identically distributed variables with expected value 0, then $Y_n = \sum_{i=0}^{n} X_i$ is a martingale (with respect to the common-refinement filtration.)[25]

We think of the filtration in a martingale representing the information available about the point $x \in \Omega$ at time n; since the information available at time $n+1$ is at least the information available at time n, we want $\mathcal{G}_{n+1}$ to refine $\mathcal{G}_n$.

Since X_n is constant on the finite partition $\mathcal{G}_n$, it is a finite variable. One of the most common choices for $\mathcal{G}_n$ is the common refinement of the partitions generated by $X_1, \ldots, X_n$, so that condition (2) reads as $\mathbb{E}[X_n \mid X_{n-1}, \ldots, X_1] = X_{n-1}$. We can think of a martingale (X_n) as representing the resources of a player in a fair game: Since the game is fair, we expect the total resources at each step, given the previous outcomes, will be the same as the resources in the previous step. Indeed, the following corollary shows that, at least for expected value, the player never prospers.

Corollary 7.25. *If $(X_n)_{n=0}^{\infty}$ is a martingale, then $\mathbb{E}[X_n] = \mathbb{E}[X_0]$ for all $n \in \mathbb{N}$.*

Proof. Applying Lemma 7.20 gives $\mathbb{E}[X_n] = \mathbb{E}[\mathbb{E}[X_n \mid \mathcal{G}_{n-1}]] = \mathbb{E}[X_{n-1}]$. □

We call a sequence of random variables $(C_n)_{n=1}^{\infty}$ on Ω *predictable* (with respect to the filtration $(\mathcal{G}_n)_{n=0}^{\infty}$) if C_n is constant on $\mathcal{G}_{n-1}$ for every $n \geq 1$. Practically speaking, this means that the value of C_n is determined by the information up to time $n-1$.

We can play a game on the martingale (X_n) as follows. At time n, the price X_n of a stock is released. You then decide to bet with or against the stock with weight C_{n+1} (which is positive when betting with, negative when betting against, and 0 if no action is taken), paying $C_{n+1}X_n$ dollars. Then the price of the stock at time $n+1$ is revealed, and you cash out the bet, receiving $C_{n+1}X_{n+1}$ dollars; so you net $C_{n+1}(X_{n+1} - X_n)$ dollars. We do this until time N. Is there a strategy that guarantees a profit?

Proposition 7.26. *Let $(\mathcal{G}_n)$ be a filtration system on $(\Omega, \mathcal{F}, \mathbb{P})$. If (X_n) is a martingale and (C_n) is a predictable sequence, then*

$$M_n = \sum_{i=1}^{n} C_i(X_i - X_{i-1})$$

is a martingale.

Proof. Since $\{C_i\}_{i=1}^{n}$ and $\{X_i\}_{i=0}^{n}$ are all constant on $\mathcal{G}_n$, so is M_n. We use Lemma 7.21 to verify condition (2):

$$\mathbb{E}[M_n - M_{n-1} \mid \mathcal{G}_{n-1}] = \mathbb{E}[C_n(X_n - X_{n-1}) \mid \mathcal{G}_{n-1}] = C_n\mathbb{E}[X_n - X_{n-1} \mid \mathcal{G}_{n-1}] = 0,$$

so $\mathbb{E}[M_n \mid \mathcal{G}_{n-1}] = \mathbb{E}[M_{n-1} \mid \mathcal{G}_{n-1}] = M_{n-1}$. □

Consequently, no strategy can be guaranteed to produce a profit, since your expected winnings are $\mathbb{E}[M_N] = \mathbb{E}[M_0] = 0$.[26]

[25] Here's an explicit construction of such an instance. The set $\Omega = [0,1]$ with the Lebesgue measure is a probability space. Let $(x)_n$ denote the nth digit in the binary expansion of x. We set $X_n(x) := 2(x)_n - 1$, and $\mathcal{G}_n$ is the partition of $[0,1]$ into 2^n intervals of length $1/2^n$.

[26] However, if you have infinite wealth, infinite time, and can choose when to stop playing, you can guarantee a profit by playing the Martingale strategy.

7.3.2 Two convergence results

The following theorem the key result we'll use on martingales.

Theorem 7.27 (Martingale convergence)**.** *If (X_n) is a bounded martingale on $(\Omega, \mathcal{F}, \mathbb{P})$, then the sequence $(X_n(x))_n$ converges for almost every $x \in \Omega$, and the limit function is bounded.*

Proof. If $(X_n(x))_n$ does not converge, then there are rational numbers $a < b$ such that $X_n(x) < a$ for infinitely many n and $X_n(x) > b$ for infinitely many n. Let $E(a,b)$ be the set of points $x \in \Omega$ that satisfy this property. If E is the set of points for which $(X_n(x))_n$ does not converge, then $E = \bigcup_{\substack{a,b\in\mathbb{Q}\\ a<b}} E(a,b)$. To prove that $\mathbb{P}(E) = 0$, it suffices to show that $\mathbb{P}(E(a,b)) = 0$ for arbitrary rational numbers $a < b$.

Fix some rational numbers $a < b$. We consider the gambling game discussed prior to Proposition 7.26. If $x \in E(a,b)$, then it seems we can cook up a good strategy: The first time $X_n < a$, buy a stock, keep holding it until $X_n > b$, at which point sell. There is a guaranteed a profit of at least $b-a$ dollars each time we pull this stunt, so after selling the stock, you repeat, buying it the next time $X_n < a$, selling it once $X_n > b$, and so on. We codify this strategy in the predictable sequence (C_n) by setting $C_1 = \delta(X_0 < a)$ and, for each $n > 1$,

$$C_n = \delta(C_{n-1} = 1)\,\delta(X_{n-1} \le b) + \delta(C_{n-1} = 0)\,\delta(X_{n-1} < a).$$

We denote by $u_N(x)$ the number of times that $(X_n(x))_n$ crosses from below a to above b, called *upcrossings*, in the interval $[0, N]$. More precisely, we set $s_1 = \min\{k : X_k(x) < a\}$, and inductively define $t_i = \min\{k > s_i : X_k(x) > b\}$ and $s_{i+1} = \min\{k > t_i : X_k(x) < a\}$; then $u_N(x) = \max\{m : t_m \le N\}$.

Suppose that $|X_n| \le K$ for all $n \in \mathbb{N}$. The net profit for our betting strategy at time N is given by

$$M_N = \sum_{i=1}^{N} C_i(X_i - X_{i-1}).$$

On the one hand, at time N we have completed $u_N(x)$ successful trades; we may have one unsold stock, which can contribute a loss of at most $2K$, if we bought it at value K and its current value is $-K$. Thus $M_N(x) \ge u_N(x)(b-a) - 2K$, so $\mathbb{E}[M_N] \ge \mathbb{E}[u_N](b-a) - 2K$. On the other hand, (M_N) is a martingale by Proposition 7.26, so $\mathbb{E}[M_N] = 0$ for all $N \in \mathbb{N}$. Combining, we have

$$\mathbb{E}[u_N] \le \frac{2K}{b-a}.$$

We define $u(x) = \lim_{N\to\infty} u_N(x)$, the total number of upcrossings in the sequence $(X_n(x))_n$. (This limit always exists because $(u_N(x))_N$ is an increasing sequence for each $x \in \Omega$.) We apply the Lebesgue Monotone Convergence Theorem to u:

$$\mathbb{E}[u] = \lim_{N\to\infty} \mathbb{E}[u_N] \le \frac{2K}{b-a}.$$

Now $x \in E(a,b)$ if and only if $u(x) = \infty$. But since the expected value of u is finite, $\mathbb{P}(u^{-1}(\infty))$ cannot be strictly positive. Therefore $\mathbb{P}(E(a,b)) = 0$, as desired.

It follows from $|X_n(x)| \le K$ that $|\lim_{n\to\infty} X_n(x)| \le K$, so the limit function is bounded. □

The following is completely unrelated to martingales, but we'll need it.

Proposition 7.28. *Let $(\Omega, \mathcal{F}, \mu)$ be a probability space and (f_n) a uniformly bounded sequence of real functions on Ω. If (f_n) converges pointwise almost everywhere to f, then $\|f - f_n\|_1 \to 0$.*

Proof. Fix some $\varepsilon > 0$. For each $N \in \mathbb{N}$, define $A_N = \{x \in \Omega : |f(x) - f_n(x)| < \varepsilon/2 \text{ for all } n \geq N\}$. This is a measurable set. Since $f_n \to f$ almost everywhere, $\mu(\bigcup_{N=1}^{\infty} A_N = 1$, and since $A_N \subseteq A_{N+1}$ for all $N \in \mathbb{N}$, we have $\mu(A_N) \to 1$. Suppose $|f_n| < K$ for all $n \in \mathbb{N}$. We can choose an $N \in \mathbb{N}$ so that $\mu(A_N) > 1 - \varepsilon/4K$. For every $n \geq N$,

$$\int_\Omega |f - f_n| \, d\mu < \frac{\varepsilon}{2}\mu(A_N) + 2K\mu(\Omega \setminus A_N) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

This shows that $\|f - f_n\|_1 < \varepsilon$ for sufficiently large n, which completes the proof. □

7.4 Proof of compactness

Standing atop our heap of assorted tools, we can now prove the following theorem.

Theorem 7.29. *The metric space $(\widetilde{\mathcal{W}}_0, \delta_\square)$ is compact.*

Proof. We want to prove that every sequence of graphons has a convergent subsequence, so let (W_n) be a sequence of graphons. Before we begin the proof, let's outline the argument with a diagram.

$$\begin{array}{ccccc} W_{1,1} & W_{1,2} & W_{1,3} & \cdots & W_1 \\ W_{2,1} & W_{2,2} & W_{2,3} & \cdots & W_2 \\ W_{3,1} & W_{3,2} & W_{3,3} & \cdots & W_3 \\ \vdots & \vdots & \vdots & & ? \\ U_1 & U_2 & U_3 & \to & W \end{array}$$

Instead of directly taking a subsequence of (W_n), we first approximate each graphon with a sequence $(W_{n,k})_k$ of step functions. These are discrete-ish objects, so if we do it carefully, we can select a subsequence of $(W_{n,k})_k$ that converges to some graphon U_k. The sequence (U_k) will converge to some graphon W, and then we show that W is the limit of a subsequence of (W_n).

With the sketch in mind, we begin by choosing a partition $\mathcal{P}_{n,k}$ for each $n, k \in \mathbb{N}$ so that

1. $\|W_n - (W_n)_{\mathcal{P}_{n,k}}\|_\square < 1/k$,
2. $|\mathcal{P}_{n,k}| = m_k$, independent of n, and
3. $\mathcal{P}_{n,k+1}$ refines $\mathcal{P}_{n,k}$.

The Weak Regularity Lemma (specifically Theorem 7.10) guarantees the existence of a collection of partitions with this property (so long as we allow empty sets in the partition).

For each $n, k \in \mathbb{N}$, there exists a measure-preserving bijection $\phi_{k,n} \colon [0,1] \to [0,1]$ that maps the partition classes of $\mathcal{P}_{n,k}$ to intervals[27]; let $\mathcal{Q}_{n,k}$ denote the resulting partition. We can choose the maps (inductively) so that $\mathcal{Q}_{n,k+1}$ refines $\mathcal{Q}_{n,k}$ (relying on condition (3)). We define $W_{n,k} = (W_n^{\phi_{n,k}})_{\mathcal{Q}_{n,k}}$.

Condition (1) guarantees that

$$\delta_\square(W_n, W_{n,k}) \leq \|W_n^{\phi_{n,k}} - W_{n,k}\|_\square = \|(W_n - (W_n)_{\mathcal{P}_{n,k}})^{\phi_{n,k}}\|_\square < \frac{1}{k}$$

for every $n \in \mathbb{N}$, so $(W_{n,k})$ converges to (W_n) uniformly.

Now we want to get the U_k. Each $W_{n,k}$ is a step function on m_k intervals, ordered by left endpoint as $V_1, \ldots, V_{m_k}$. We assign to it the vector $\alpha_{n,k} \in [0,1]^{m_k}$, whose ith entry is the length of V_i, and the matrix $\beta_{m,k} \in [0,1]^{m_k \times m_k}$, whose (i,j)th entry is the value of $W_{n,k}$ on $V_i \times V_j$ (and 0 if either V_i or V_j is the empty set). Since $[0,1]$ is compact, so too is the set

[27] Construct one!

$[0,1]^{m_k^2+m_k}$, so for each $k \in \mathbb{N}$ there is a subsequence of indices (n_r) such that $(\alpha_{n_r,k}, \beta_{n_r,k})_r$ converges in each element to some (α_k, β_k). This vector uniquely describes a graphon, and this graphon is the pointwise (almost everywhere) limit of $(W_{n_r,k})_r$.

We want to choose one subsequence that works for all k simultaneously. We do that by first choosing a subsequence $(n_{r,1})$ that works for $k = 1$ and from that choosing a subsequence $(n_{r,2})$ that works for $k = 2$, then a further subsequence $(n_{r,3})$ that works for $k = 3$, and so on, with $(n_{r,k})_r$ a subsequence of $(n_{r,k-1})_r$ such that $(W_{n_{r,k},k})_r$ is convergent pointwise in the cut norm. The trick is to combine the kth terms from $(n_{r,k})_r$ into a new sequence: $(W_{n_{r,r},k})_r$ is also convergent for every $k \in \mathbb{N}$, and we call the limit graphon U_k. We denote by $\mathcal{Q}_k$ the partition associated to U_k (that is, the limit of the partitions $Q_{n_{r,r},k}$ as $r \to \infty$).

At this point, we're swimming in subindices. To ease eye strain, let's pass to the subsequence $(n_{r,r})$ and relabel the indices just using r, so that $W_{n_{r,r},k}$ becomes $W_{r,k}$.

The partition $\mathcal{Q}_{n,k+1}$ is a refinement of $\mathcal{Q}_{n,k}$, so $(W_{n,k+1})_{\mathcal{Q}n,k} = W_{n,k}$ by Lemma 7.20. Taking the limit as $n \to \infty$, we get $(U_{k+1})_{Q_k} = U_k$.[28] Now we prepare for magic. The unit square $[0,1]^2$ with the Lebesgue measure is a probability space, and (U_k) is a sequence of bounded random variables on that space (since $0 \le U_k \le 1$). Refinement is preserved in the limit, so $(\mathcal{Q}_k)$ is a filtration, and the relationship $(U_{k+1})_{\mathcal{Q}_k} = U_k$ means that (U_k) is a martingale. Abracadabra! The sequence (U_k) converges pointwise almost everywhere to a limit kernel W. (This bit of legerdemain brought to you by Theorem 7.27.)

All that remains is a standard analysis argument. To make it less standard, let $\varepsilon > 0$. By Proposition 7.28, there is a $\delta > 3/\varepsilon$ so that $\|W - U_\delta\|_1 < \varepsilon/3$, and there is a corresponding $N \in \mathbb{N}$ so that $\|W_{n,\delta} - U_\delta\|_1 < \varepsilon/3$ for all $n > N$. Thus, for all $n > N$,

$$\begin{aligned}\delta_\square(W_n, W) &\le \delta_\square(W_n, W_{n,\delta}) + \delta_\square(W_{n,\delta}, U_\delta) + \delta_\square(U_\delta, W)\\ &\le \delta_\square(W_n, W_{n,\delta}) + \|W_{n,\delta} - U_\delta\|_1 + \|U_\delta - W\|_1 < \varepsilon.\end{aligned}$$

Therefore $W_n \to W$. □

This proof works equally well for any closed subset of kernels with uniformly bounded range. For example, $(\widetilde{\mathcal{W}}_1, \delta_\square)$ is compact.

There are two immediate corollaries.

Corollary 7.30. *If a graph sequence (G_n) is convergent in the homomorphism density sense, then there is a graphon W so that $t(F, G_n) \to t(F, W)$ for every finite graph F.*

Proof. Let $W_n = W_{G_n}$ and take a convergent subsequence of $W_{n_k} \to W$ with respect to $\delta_\square$. Then $t(F, G_{n_k}) \to t(F, W)$ for every finite graph F by the Counting Lemma. Since $(t(F, G_n))$ converges, it also tends to $t(F, W)$. □

Corollary 7.31. *For every $\varepsilon > 0$, there exists a natural number $k(\varepsilon)$ such that for any graphon W, there is a $k(\varepsilon)$-vertex graph G so that $\delta_\square(G, W) < \varepsilon$.*

Proof. Let $B_\varepsilon(W) = \{U \in \mathcal{W} : \delta_\square(U, W) < \varepsilon\}$. The collection $\{B_\varepsilon(G)\}_{G \in \mathcal{F}}$ is an open cover of $\mathcal{W}_0$, so there is a finite collection of graphs $\{G_i\}_{i=1}^n$ such that $\{B_\varepsilon(G_i)\}_{i=1}^n$ covers $\mathcal{W}_0$. Let $k(\varepsilon)$ be the least common multiple of $\mathsf{v}(G_1), \dots, \mathsf{v}(G_n)$. For each G_i there is a blowup G_i' with $k(\varepsilon)$ vertices. Since $\delta_\square(G_i, G_i') = 0$, the set $\{B_\varepsilon(G_i')\}_{i=1}^n$ covers $\mathcal{W}_0$, which proves the claim. □

[28] This can be verified by first using the triangle inequality:

$$\|(W_{n,k+1})_{\mathcal{Q}_{n,k}} - (U_{k+1})_{\mathcal{Q}_k}\|_1 \le \|(W_{n,k+1})_{\mathcal{Q}_{n,k}} - (W_{n,k+1})_{\mathcal{Q}_k}\|_1 + \|(W_{n,k+1})_{\mathcal{Q}_k} - (U_{k+1})_{\mathcal{Q}_k}\|_1.$$

The second term on the right-hand side is as most $\|W_{n,k+1} - U_{k+1}\|_1$ by Exercise 7.6, and which goes to 0 by Proposition 7.28. With a little work, the left-hand side also goes to zero.

8 Sampling

Up to this point, sampling has been implicit through the homomorphism densities. In this section, we discuss some of the more explicit connections between sampling and the limit theory.

The proofs are in general tedious and quite involved, so they don't appear here.[29] Instead, we'll give the broad strokes, statements of the main theorems, some consequences, and a few examples.

8.1 The Sampling Lemmas

For a subset $S \subseteq V(G)$, we let $G[S]$ denote the subgraph of G induced by S. The first theorem says that the labelled cut distance of two large graphs can be estimated via their samples.

Theorem 8.1 (First Sampling Lemma for Graphs)**.** *Let G and H be two graphs on the same vertex set. Let $k \leq \mathsf{v}(G)$ and S be a subset of $V(G)$ chosen uniformly at random from all subsets of size k. With probability at least $1 - 4e^{-\sqrt{k}/10}$,*

$$|d_\square(G[S], H[S]) - d_\square(G, H)| \leq \frac{8}{k^{1/4}}.$$

To put it another way, suppose we want to estimate the $d_\square$ distance between two large graphs G and H with an error of at most ε. The First Sampling Lemma says that we can do this by picking two random samples on $k \geq 4096/\varepsilon^4$ vertices, and that this strategy almost never fails (in this case at least it fails with probability at most $4e^{-6.4/\varepsilon^2}$, although this gets better with larger k).

There is a similar result for graphons. If U is a kernel and $S = (s_1, \ldots, s_k)$ is an ordered k-tuple of points in $[0,1]$, we let $U[X]$ denote the $k \times k$ matrix with $(U[S])_{i,j} = U(s_i, s_j)$.

Theorem 8.2 (First Sampling Lemma for Kernels)**.** *Let $U \in \mathcal{W}_1$ and $S \in [0,1]^k$ be chosen uniformly at random. With probability at least $1 - 4^{-\sqrt{k}/10}$,*

$$-\frac{3}{k} \leq \|U[S]\|_\square - \|U\|_\square \leq \frac{8}{k^{1/4}}.$$

The interesting part about this result is that the upper and lower bounds are not the same! It might be tempting to say that the lemma for graphs follows from the lemma for kernels by setting $U = W_G - W_H$, but this isn't true: The distribution of $W_G[S]$ is the same as sampling from G allowing repeated vertices, while $G[S]$ does not.

The Second Sampling Lemmas address the distance of a sample from the original graph or kernel.

Theorem 8.3 (Second Sampling Lemma for Graphs)**.** *Let $k \geq 1$ and G be a simple graph with at least k vertices. If S is chosen uniformly at random from the subsets of $V(G)$ of size k, then with probability at least $1 - \exp(-k/(2\log_2 k))$,*

$$\delta_\square(G[S], G) \leq \frac{20}{\sqrt{\log_2 k}}.$$

[29] If you simply must see them, the details are in [2]. The Sampling Lemmas are discussed in Sections 10.1–10.4 and the proof of Theorem 8.6 appears in Section 11.4.

To state the corresponding theorem for kernels, we need to change how we sample from kernels. Let W be a kernel and fix some $S \in [0,1]^k$. The edge-weighted graph $\mathbb{H}(S,W)$ has vertex set $[k]$ and weight $\beta_{i,j} = W(s_i, s_j)$ for each $i, j \in [k]$ with $i \neq j$ and $\beta_{i,i} = 0$ (so $\mathbb{H}(S,W)$ has no loops). The random weighted graph $\mathbb{H}(k,W)$ obtained from W is defined by the distribution of $\mathbb{H}(S,W)$ where $S \in [0,1]^k$ is chosen uniformly at random. So, to get $\mathbb{H}(k,W)$, we choose k random points in $[0,1]$ and then weight the edges with the values given by W.

We can form a simple graph from each edge-weighted graph H as follows. The random simple graph $\mathbb{G}(H)$ has vertex set $V(H)$, and each edge ij appears independently with probability $\beta_{i,j}$ (for every $i \neq j$; we exclude loops). For each $S \in [0,1]^k$, we set $\mathbb{G}(S,W) = \mathbb{G}(\mathbb{H}(S,W)$, and for each $k \geq 1$ we define $\mathbb{G}(k,W) = \mathbb{G}(\mathbb{H}(k,W))$. That is, to sample a simple graph from W, we form a "template" edge-weighted graph by sampling $\mathbb{H}(k,W)$, and then sample that graph. If $W = p$, then $\mathbb{G}(n,W)$ is the Erdős-Rényi random graph. If you wish, you can show that $d_\square(\mathbb{G}(H), H) < \varepsilon$ with high probability for every edge-weighted graph H with weights in $[0,1]$; this is an application of Hoeffding's inequality.

Theorem 8.4 (Second Sampling Lemma for Graphons)**.** *Let $k \geq 1$ and $W \in \mathcal{W}_0$. With probability at least $1 - \exp(-k/(2\log_2 k))$,*

$$\delta_\square(\mathbb{H}(k,W), W) \leq \frac{20}{\sqrt{\log_2 k}} \quad \text{and} \quad \delta_\square(\mathbb{G}(k,W), W) \leq \frac{22}{\sqrt{\log_2 k}}.$$

The Second Sampling Lemma provides another proof that simple graphs are dense in the space of graphons.

Proposition 8.5. *If W is a graphon, then $\mathbb{G}(n,W) \to W$ with probability 1.*

Proof. From Theorem 8.4, we know that $\delta_\square(\mathbb{G}(n,W), W) > 22/\sqrt{\log_2 n}$ with probability at most $\exp(-n/(2\log_2 n)) < e^{-\sqrt{n}}$. The series $\sum_{n=1}^{\infty} e^{-\sqrt{n}}$ is convergent (for example, by the integral test). By Borel-Cantelli (Theorem 3.9), the sequence $(\mathbb{G}(n,W))$ satisfies $\delta_\square(\mathbb{G}(n,W), W) \leq 22/\sqrt{\log_2 n}$ for all but finitely many n with probability 1, so it converges almost surely. □

We conclude with a new technique for showing that a sequence of graphons converges. We call a sequence (S_n) of finite subsets of $[0,1]$ *well-distributed* if $|S_n \cap I|/|S_n| \to m(I)$ for every interval $I \subseteq [0,1]$. Equivalently, (S_n) is well-distributed if the uniform measure on S_n weakly converges to the uniform (i.e. Lebesgue) measure on $[0,1]$. (See Definition 5.1.)

Theorem 8.6. *If $W \in \mathcal{W}_0$ is continuous almost everywhere and (S_n) is a well-distributed set, then $\mathbb{G}(S_n, W)$ converges to W with probability 1.*

The proof is mainly a (not terribly interesting) measure theory argument, so we omit it.

8.2 Examples of convergent randomly-generated graph sequences

We have seen, in detail with Erdős-Rényi random graphs and at a high level with graphon sampling, that a randomized procedure can result in an almost certainly convergent graph sequence. In this section, we present a few more examples of this.

Example 8.7 (Cloning 1)**.** Let G_0 be a finite simple graph. We build a sequence (G_n) randomly, beginning with G_0, by choosing a vertex v from G_0 uniformly at random at each step and cloning it. For example, if $G_0 = K_3$, the sequence may look like this:

This very similar to taking a sample of n points from W_{G_0}. There are two differences: First, we start with G_0, not the empty graph. Second, the graphs are constructed inductively, so they aren't independent. So while we can't simply appeal to the Second Sampling Lemma to say that this converges, it nevertheless converges to W_{G_0} with probability 1. (This can be show using Theorem 8.6 and the Strong Law of Large Numbers, for example.)

Example 8.8 (Cloning 2)**.** Suppose we modify the construction in the previous example: To form G_n, instead of choosing a vertex from G_0 to clone, we choose a vertex uniformly from G_{n-1}. It happens that this graph sequence also converges with probability 1, but its limit is not determined—if you run this sequence multiple times, then it will converge each time, but always to a different limit! (See Example 11.43 of [2] for slightly more detail.)

Example 8.9 (Uniform attachment)**.** We grow another graph sequence, this time starting with a single vertex. Let G_1 be an isolated vertex labelled 0. At time n, we introduce a new isolated vertex n to G_{n-1} and independently connect every pair of nonadjacent vertices with probability $1/n$. These graphs are called *uniform-attachment graphs*.

What's the probability that two vertices are connected in G_n? if $0 \leq i < j \leq n-1$, then i and j are not adjacent in step n if and only if they weren't connected in any of the steps $j, j+1, \ldots, n$. this occurs with probability $\frac{j}{j+1}\frac{j+1}{j+2}\cdots\frac{n-1}{n} = \frac{j}{n}$. This is independent for each pair ij by assumption. Thinking in terms of a graphon, the probability that i and j are connected is $1 - \max\{i/n, j/n\}$.

Set $S = \{0, 1/n, \ldots, (n-1)/n\}$. Since the edges are independent, this means that G_n has the same distribution as $\mathbb{G}(S_n, W)$, where $W(x, y) = 1 - \max\{x, y\}$. Since W is continuous almost everywhere and S_n is well-distributed, Theorem 8.6 implies that the graph sequence (G_n) converges with probability 1.

Further examples appear in Section 11.4 of [2].

Exercise 8.10. Let (G_n) be a sequence of uniform attachment graphs. What is the expected average degree of G_n? What is the expected number of edges?

9 Homomorphism inequalities

9.1 Apéritif

How many edges can we stuff into a graph without creating a triangle? With this question began the field of extremal graph theory, often[30] referred to as Xtreme Graph Theory. In general, questions in extremal graph theory ask how certain graph quantities are related (via inequality) and which graphs maximize or minimize this relationship. It's the search for graphs that live on the edge.

But wait, you say: Extremal graph theory is Section 10, and this is only Section 9. Good eye. In this section, we'll only look at the first part of this question: How are different homomorphism densities related? For example, the edge-stuffing question can be rephrased to ask: If $t(K_3, G) = 0$, how large can $t(K_2, G)$ be? Mantel first answered this question in 1907.

Theorem 9.1 (Mantel)**.** *If G contains no triangle, then $\mathsf{e}(G) \le \mathsf{v}(G)^2/4$.*

Proof. Suppose G has no triangle. Choose any two adjacent vertices x and y. Any other vertex z is adjacent to at most one of x and y, so $\deg(x) + \deg(y) \le \mathsf{v}(G)$. Therefore

$$\sum_{xy \in E(G)} (\deg(x) + \deg(y)) \le \mathsf{e}(G)\mathsf{v}(G).$$

Imagine placing, for each $x \in V(G)$, the degree of x at each of its incidences. Adding up the two labels on each edge and then summing over edges gives the left-hand sum above. On the other hand, adding up the labels at each vertex and then summing over vertices gives $\sum_{x \in V(G)} \deg(x)^2$. From Cauchy-Schwarz,

$$(2\mathsf{e}(G))^2 = \left(\sum_{x \in V(G)} \deg(x) \right)^2 \le \left(\sum_{x \in V(G)} 1^2 \right) \left(\sum_{x \in V(G)} \deg(x)^2 \right) \le \mathsf{e}(G)\mathsf{v}(G)^2.$$

Dividing both sides by $4\mathsf{e}(G)$ finishes the proof. □

This is sharp: $K_{n,n}$ is the extremal graph. The interpretation above was in "building up": how many edges can we put into the empty graph while keeping it triangle free? We can turn this on its head and ask: What is the minimum number of edges we must remove from K_n to eliminate all triangles? Mantel's Theorem says, basically, half of them. It turns out that this is not peculiar to the complete graphs.

Exercise 9.2. Show that it is always possible to remove at most $\mathsf{e}(G)/2$ edges from a graph G to form a bipartite graph.

What of the other direction? How does the number of edges limit the number of triangles?

Theorem 9.3 (Kruskal–Katona)**.** *For every graph G,*

$$t(K_3, G) \le t(K_2, G)^{3/2}.$$

Although there is no graph where equality holds, the inequality is asymptotically tight.

Exercise 9.4. Let G_n be the graph $K_n\overline{K_n}$, that is, the union of a clique of size n with n isolated vertices. Show that $\lim_n t(K_3, G_n) = \lim_n t(K_2, G)^{3/2}$.

[30] Actually never.

We give two proofs of Theorem 9.3. The first has a spectral flavor. It relies on the following observation.

Lemma 9.5. *If $a_1, \dots, a_n \geq 0$ and $t \geq 1$, then $\sum a_i^t \leq (\sum a_i)^t$.*

Proof. If $\sum a_i = 0$ then all a_i are 0 and the result is trivial. Otherwise scale the coefficients by setting $b_i = a_i(\sum a_i)^{-t}$. The inequality is true with $b_1, \dots, b_n$ if and only if it is for $a_1, \dots, a_n$, and the nice part is that $b_i \in [0,1]$ and $\sum b_i = 1$. So the right-hand side of the inequality is 1, and $\sum b_i^t \leq \sum b_i \leq 1$. □

Proof 1. Let G be a simple graph on n vertices and A be its adjacency matrix. Since A is real and symmetric, it has a full spectrum by the Spectral Theorem. Let $👻_1, \dots, 👻_n$ be the eigenvalues of A.[31] The quantities $n^2 t(K_2, G)$ and $n^3 t(K_3, G)$ (that is, $\mathsf{hom}(K_2, G)$ and $\mathsf{hom}(K_3, G)$) are the number of closed walks of length 2 and 3, respectively, in G. Thus

$$n^2 t(K_2, G) = \sum_{i=1}^{n} (A^2)_{i,i} = \operatorname{tr}(A) = \sum_{i=1}^{n} 👻_i^2.$$

Similarly $n^3 t(K_3, G) = \sum_{i=1}^{n} 👻_i^3$. Using the lemma with $a_i = 👻_i^2$ and $t = 3/2$ gives

$$n^3 t(K_3, G) = \sum_{i=1}^{n} 👻_i^3 \leq \left(\sum_{i=1}^{n} 👻_i^2 \right)^{3/2} = (n^2 t(K_2, G))^{3/2}.$$

Then divide by n^3. □

With not even a pause for a breath, let's start proof two. This one is a tricky application of Cauchy-Schwarz.

Proof 2. The usual: G is a graph with n vertices and A is its adjacency matrix. For clarity, we denote the (i,j) entry of A by $a(i,j)$. As mentioned before, $n^3 t(K_3, G)$ is the number of closed walks of length 3 in G, so

$$(n^3 t(K_3, G))^2 = \Big(\sum_{x,y,z \in V(G)} a(x,y)a(y,z)a(z,x) \Big)^2 = \Big(\sum_{y,z \in V(G)} a(y,z) \Big(\sum_{x \in V(G)} a(x,y)a(z,x) \Big) \Big)^2.$$

Applying Cauchy-Schwarz to the last expression provides an upper bound of

$$\Big(\sum_{y,z \in V(G)} a(y,z)^2 \Big) \Big(\sum_{y,z \in V(G)} \Big(\sum_{x \in V(G)} a(x,y)a(z,x) \Big)^2 \Big).$$

The left term is $n^2 t(K_2, G)$. Expanding the square in the right term yields the expression

$$\sum_{y,z \in V(G)} \Big(\sum_{x \in V(G)} a(x,y)a(z,x) \Big) \Big(\sum_{w \in V(G)} a(w,y)a(z,w) \Big) = \sum_{x,y,w,z} a(x,y)a(y,w)a(w,z)a(z,x),$$

which can be recognized as $n^4 t(C_4, G)$. We finish by applying Lemma 4.1 with $K_2 K_2 \subseteq C_4$ (recall that $t(K_2, G)^2 = t(K_2 K_2, G)$):

$$t(K_3, G)^2 \leq t(C_4, G) t(K_2, G) \leq t(K_2, G)^3.$$ □

Perchance you judge this a collection of *ad hoc* skullduggery. You would be right—except that some of this can be distilled into extremely effective voodoo that gives very short magical proofs of statements about these and similar inequalities. This journey into the occult, unlike a typical supernatural foray, will be well-grounded, with plenty rigor. Let's begin. ——==👻

[31] Spectral . . . get it?

9.2 Quantum graphs and gluing

The type of relationships that we will focus on are *algebraic inequalities* between homomorphism densities that are valid for all graphs, like Mantel's Theorem and the Kruskal–Katona Theorem. Because homomorphism numbers are nice, it turns out that all algebraic inequalities reduce to linear ones, since products of homomorphism densities collapse into one homomorphism density via Lemma 4.2. For brevity, we denote the disjoint union of F_1 and F_2 by F_1F_2 and the disjoint union of k copies of F by F^k.

Definition 9.6. A *quantum graph* is a finite linear combination of finite multigraphs (over $\mathbb{R}$). The multigraphs in a quantum graph with nonzero coefficients are called its *constituents*.[32]

Wait, hold up. Multigraphs? We haven't talked about multigraphs at all. Nonetheless, they will be relevant here, so let's briefly discuss extensions from simple graphs to multigraphs.

Let F and G be multigraphs (so multiple edges can appear between a given pair of vertices). A *homomorphism* from F to G is a pair of maps $\phi\colon V(F) \to V(G)$ and $\psi\colon E(F) \to E(G)$ such that if e connects $x, y \in V(F)$, then $\psi(e)$ connects $\phi(x)$ and $\phi(y)$ in G. If F and G are simple graphs, this reduces to a regular homomorphism. We denote by $\mathsf{hom}(F, G)$ the number of homomorphisms from F to G and define the homomorphism density of F in G by

$$t(F, G) = \frac{\mathsf{hom}(F, G)}{\mathsf{v}(G)^{\mathsf{v}(F)}}.$$

The interpretation of this quantity is a bit more subtle than before: It is the average number of homomorphisms from F to G that use a fixed vertex map.

Homomorphism density into kernels is not really different: $t(F, W) = \int \prod_{ij \in E(F)} W(x_i, x_j)$; we just think of $E(F)$ as a multiset of edges. Suppose G is a multigraph on n nodes, and let $\mathcal{P} = \{V_1, \dots, V_n\}$ be the partition of $[0, 1]$ into n intervals of equal length. We define W_G as the step function where, if $x \in V_i$ and $y \in V_j$, then $W(x, y)$ is the number of edges between i and j in G. With this definition, $t(F, G) = t(F, W_G)$ for all multigraphs G. (Alternatively, if A is the adjacency matrix of G, where $a_{i,j}$ is the number of edges between i and j, then $W_G := W_A$, as in Definition 5.16.)

Example 9.7. If $F =$  , then $t(F, W) = \int W^2\, dxdy$.

The multigraph content is mostly preparation for the (near) future, but quantum graphs are not. The set of quantum graphs is the real vector space with the set of finite multigraphs as a basis.

Definition 9.8. We extend the homomorphism density operator linearly to the set of all quantum graphs. That is, if  $= \sum_{i=1}^{n} a_i F_i$ is a quantum graph, we define $t($$, W) = \sum_{i=1}^{n} a_i t(F_i, W)$. We write ≥ 0 if $t($$, W) \geq 0$ for every graphon W, and we call such a quantum graph *nonnegative*. Naturally, we write $\geq$ if $-$ ≥ 0.[33]

Since algebraic statements about homomorphisms can be reduced to linear ones, every algebraic inequality can be expressed as ≥ 0 for some quantum graph . The type of result we're interested in here can be concisely summed up as:

Problem. Find nonnegative quantum graphs.

Since graphs are dense in the space of graphons, ≥ 0 if and only if $t($$, G) \geq 0$ for every finite simple graph G. Thus, from the Kruskal–Katona Theorem, we know that $K_2^3 - K_3^2 \geq 0$.

[32] This definition has absolutely nothing to do with the *other* definition of quantum graph, which involves labelling each edge with a differential equation.

[33] Atom icon by Fengquan Li from the Noun Project.

Exercise 9.9. Convince yourself: Show that $t(\bowtie, W) \geq 0$ for every graphon W if and only if $t(\bowtie, G) \geq 0$ for every graph G.

Now we have to introduce something weird—a product on graphs (not just the disjoint union). The best motivation is that it drastically reduces the difficulty of many homomorphism density inequalities, but unfortunately this is only possible to show after it's introduced. You can gain some perspective, perhaps, by glancing ahead at the shortened proof of Kruskal–Katona (Theorem 9.29), as well as Proposition 9.30 and Proposition 9.31.

Back? Good.

Definition 9.10. A k-labelled graph is a multigraph where each of the elements of $\{1, 2, \ldots, k\}$ is assigned to a distinct vertex. A *partially labelled graph* is a k-labelled graph for some $k \in \mathbb{N}$.

In other words, a k-labelled graph G is an injective function $[k] \to G$.

Definition 9.11. Let F_1 and F_2 be two k-labelled graphs. The *gluing product* of F_1 and F_2 is the graph obtained by identifying vertices with the same labels in the disjoint union of F_1 and F_2.

Here are two examples of gluing products.

Exercise 9.12. The gluing product is associative and commutative. Among the set of k-labelled graphs, what is the identity element?

A *partially labelled quantum graph* is a linear combination of partially labelled graphs. We extend the gluing product linearly the entire space of partially labelled quantum graphs.

Definition 9.13. The *unlabelling operator* $[\![\cdot]\!]$ removes the labels from a partially labelled graph.

9.3 Graph parameters and connection matrices

Definition 9.14. A *graph parameter* is a real function on the isomorphism classes of multigraphs; a *simple graph parameter* is a real function on the isomorphism classes of simple graphs.

For example, number of vertices, chromatic number, independence number, and girth are all graph parameters. If we fix a graph G and a kernel W, then $\mathsf{hom}(\,\cdot\,, G)$ and $t(\,\cdot\,, W)$ are graph parameters as well. Any graph parameter can be extended linearly to all quantum graphs, as we did in the last section for homomorphism density. Every graph parameter has a simple counterpart, by reducing any multigraph to the underlying simple graph[34] before applying the parameter. In the other direction, we can extend any simple graph parameter to all multigraphs by defining the value of a multigraph to be the value of the underlying simple graph. For brevity, we define some notation: Let F be a multigraph and F^* be the underlying simple graph of F. We denote by t^* the induced simple graph parameter of t, that is, $t^*(F, W) = t(F^*, W)$.

Definition 9.15. The *connection matrix* of a graph parameter f is the countably infinite matrix indexed by partially labelled graphs whose (F_i, F_j) entry is $f([\![F_i F_j]\!])$.

[34]The underlying simple graph is what you get by replacing multiple edges between a pair of vertices by a single edge.

So we glue two graphs together, turn it back into your standard, everyday multigraph, and then apply the graph parameter. There's a lot that can go on with connection matrices; we'll restrict ourselves to the cases where f is homomorphism density or the corresponding induced simple graph parameter.[35]

How does the gluing algebra interact with homomorphism densities?

Definition 9.16. Suppose that F is a k-labelled graph. We suppose that $V(F) = [n]$ and that the labels of F correspond to the vertices $1, 2, \dots, k$. For $x \in \mathbb{R}^k$ and any kernel W, we define

$$t_x(F, W) = \int_{[0,1]^{[n]\setminus[k]}} \prod_{ij \in E(F)} W(x_i, x_j)\, dx_{k+1} \cdots dx_n.$$

In other words, we integrate the homomorphism density integrand with respect to all the unlabelled vertices of F, and $t_x(F, W)$ is a function of k variables (each corresponding to a labelled vertex of F). If F_1 and F_2 are k- and r-labelled graphs, respectively, with $k < r$, let $\tilde{F}_1$ denote the disjoint union of F_1 with $r-k$ isolated nodes labelled $k+1, \dots, r$. Then $F_1F_2 = \tilde{F}_1F_2$ and $t_x(F_1, W) = t_x(\tilde{F}_1, W)$, so we may always assume in a product of two partially labelled graphs that the label sets are the same.

Integrating with respect to the remaining variables is what glues two graphs together.

Exercise 9.17. Let F_1 and F_2 be two k-labelled graphs. Show that for every kernel W,

$$t(\llbracket F_1F_2 \rrbracket, W) = \int_{[0,1]^k} t_x(F_1, W) t_x(F_2, W)\, dx. \tag{9.1}$$

The main theoretical result can be phrased as follows.

Theorem 9.18. *Any finite submatrix of the connection matrix for $t(\,\cdot\,, W)$ is positive semidefinite for every $W \in \mathcal{W}$, and the same holds for $t^*(\,\cdot\,, W)$ for every $W \in \mathcal{W}_0$.*

This might seem like gobbledygook, and not very useful gobbledygook at that. Nonetheless, it is important—it's the heart of the voodoo we're trying to create. To make the most of it, we'll need to review positive semidefinite matrices.

9.4 Positive semidefinite matrices

Definition 9.19. A symmetric $n \times n$ real matrix A is called *positive definite* if $v^\top Av$ is strictly positive for every nonzero vector $v \in \mathbb{R}^n$. The matrix A is called *positive semidefinite* if $v^\top Av$ is nonnegative for every $v \in \mathbb{R}^n$.

By the Spectral Theorem, every positive semidefinite matrix has an orthonormal basis of eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ and corresponding real eigenvalues $\lambda_1, \dots, \lambda_n$.

Lemma 9.20. *Every eigenvalue of a positive semidefinite matrix is nonnegative.*

Proof. By assumption, $0 \le \mathbf{v}_i^\top A \mathbf{v}_i = \lambda_i(\mathbf{v}_i^\top \mathbf{v}_i)$. Since $\mathbf{v}_i^\top \mathbf{v}_i > 0$, the eigenvalue must be nonnegative. □

The converse of Lemma 9.20 holds, as well.

[35] For more on connection matrices, see Chapter 4 of [2] or Lovász's notes at http://web.cs.elte.hu/~lovasz/welsh.pdf.

Corollary 9.21. *The determinant of any positive semidefinite matrix is nonnegative.*

Proof. The determinant is the product of the eigenvalues. □

Since all the eigenvalues are nonnegative, we'll orient them to face the right half of the real line: $\measuredangle_1, \dots, \measuredangle_n$.

There's actually only one other result that we need. The proof is longer than the previous one, but it's not too bad.

Definition 9.22. Let A and B be two matrices with the same dimensions. Their *Schur product*, denoted $A \circ B$, is a matrix of the same dimensions obtained via entrywise multiplication, that is, $(A \circ B)_{i,j} = a_{i,j} b_{i,j}$.

Theorem 9.23 (Schur product theorem)**.** *If A and B are positive semidefinite, then $A \circ B$ is.*

The proof uses the following facts.

Exercise 9.24. Suppose A is a positive semidefinite matrix with orthonormal eigenbasis $\text{🏃}_1, \dots, \text{🏃}_n$ and corresponding eigenvalues $\measuredangle_1, \dots, \measuredangle_n$. Prove that

1. $A = \sum_{i=1}^n \measuredangle_i \text{🏃}_i \text{🏃}_i^\top$. (It might help to show that two $n \times n$ matrices A and B are equal if and only if $Av = Bv$ for every vector $v \in \mathbb{R}^n$.)
2. $(uu^\top) \circ (vv^\top) = (u \circ v)(u \circ v)^\top$ for every pair of vectors $u, v \in \mathbb{R}^n$.
3. $vv^\top$ is positive semidefinite for every $v \in \mathbb{R}^n$.
4. the linear combination of positive semidefinite matrices with nonnegative coefficients is itself positive semidefinite.

Proof of Schur product theorem. Suppose A has orthonormal eigenbasis $\text{🏃}_1, \dots, \text{🏃}_n$ with corresponding eigenvalues $\measuredangle_1, \dots, \measuredangle_n$ and B has orthonormal eigenbasis $\text{☺}_1, \dots, \text{☺}_n$ with corresponding eigenvalues $\text{🦇}_1, \dots, \text{🦇}_n$. From the exercise,

$$A \circ B = \sum_{i,j=1}^n \measuredangle_i \text{🦇}_j (\text{🏃}_i \text{🏃}_i^\top) \circ (\text{☺}_j \text{☺}_j^\top).$$

Also from the exercises, $(\text{🏃}_i \text{🏃}_i^\top) \circ (\text{☺}_j \text{☺}_j^\top)$ is a positive semidefinite matrix. The matrix $A \circ B$ is a nonnegative linear combination of positive semidefinite matrices and is therefore positive semidefinite. □

9.5 Connection matrices are positive semidefinite

Let's do what the title says.

Proof of Theorem 9.18. We first prove it for $t(\,\cdot\,, W)$ with $W \in \mathcal{W}$. We want to show that for any finite collection of k-labelled multigraphs $F_1, \dots, F_n$ and real numbers $a_1, \dots, a_n$,

$$\sum_{i,j=1}^n a_i a_j t([\![F_i F_j]\!], W) \geq 0. \tag{9.2}$$

This is not so hard: Writing $t([\![F_i F_j]\!], W)$ as an integral as in (9.1) and taking the sum inside gives

$$\int_{[0,1]^k} \sum_{i,j=1}^n a_i a_j t_x(F_i, W) t_x(F_j, W)\, dx = \int_{[0,1]^k} \left(\sum_{i=1}^n a_i t_x(F_i, W) \right)^2 dx \geq 0.$$

So much for the first part. Now suppose that $F_1, \dots, F_n$ are k-labelled simple graphs; we want to show (9.2) for t^*. For any k-labelled graph F, let F^ℓ denote the graph on $[k]$ induced by the labelled vertices of F, and let F^r denote the graph obtained from F by deleting the edges spanned by labelled vertices. Then

$$(F_1F_2)^* = F_1^r F_2^r (F_1^\ell \cup F_2^\ell),$$

where the disjoint union is a graph on $[k]$ edge set $E(F_1^\ell) \cup E(F_2^\ell)$. Each of these is a simple graph, so by Lemma 4.2,

$$t^*(\llbracket F_iF_j \rrbracket, W) = t_x(F_1^r, W)t_x(F_2^r, W)t_x(F_1^\ell \cup F_2^\ell, W).$$

Thus

$$\sum_{i,j=1}^n a_i a_j t^*(\llbracket F_iF_j \rrbracket, W) = \int_{[0,1]^k} \sum_{i,j=1}^n a_i a_j t_x(F_i^r, W)t_x(F_j^r, W)t_x(F_i^\ell \cup F_j^\ell, W)\, dx. \tag{9.3}$$

The argument from Exercise 4.13 shows that

$$t_x(F_i^\ell \cup F_j^\ell, W) = \sum_{\substack{H \supseteq F_i^\ell \cup F_j^\ell \\ V(H)=[k]}} t_{\mathrm{ind},x}(H, W),$$

for every $x \in \mathbb{R}^k$, where $t_{\mathrm{ind},x}$ is defined analogously to t_x. Inserting this into (9.3) and exchanging the order of summation gives

$$\begin{aligned}\int_{[0,1]^k} \sum_{V(H)=[k]} \sum_{F_i^\ell \cup F_j^\ell \subseteq H} a_i a_j t_x(F_i^r, W)t_x(F_j^r, W)t_{\mathrm{ind},x}(H, W)\, dx \\ = \int_{[0,1]^k} \sum_{V(H)=[k]} \left(\sum_{F_i^\ell \subseteq H} a_i t_x(F_i^r, W) \right)^2 t_{\mathrm{ind},x}(H, W)\, dx.\end{aligned} \tag{9.4}$$

Since $0 \le W \le 1$, the induced homomorphism $t_{\mathrm{ind},x}(H, W)$ is nonnegative for every $x \in \mathbb{R}^k$, so the integral is as well. This finishes the proof. □

Great! Now let's reap the rewards.[36]

Corollary 9.25. *If g is a partially labelled quantum graph, then $\llbracket g^2 \rrbracket \ge 0$.*

Proof. Let $g = \sum_{i=1}^n a_i F_i$. Since homomorphism density and unlabelling are linear,

$$t(\llbracket g^2 \rrbracket, W) = t\left(\left\llbracket \sum_{i,j=1}^n a_i a_j F_i F_j \right\rrbracket, W \right) = \sum_{i,j=1}^n a_i a_j t(\llbracket F_iF_j \rrbracket, W) \ge 0.$$ □

Corollary 9.26. *Let $F_1, \dots, F_n$ be k-labelled graphs and W be a kernel. If A is the $n \times n$ matrix with $a_{i,j} = t(\llbracket F_iF_j \rrbracket, W)$, then $\det(A) \ge 0$. The same statement holds if $a_{i,j} = t^*(\llbracket F_iF_j \rrbracket, W)$.*

Proof. This follows from Theorem 9.18 and Corollary 9.21. □

[36] I suppose when you sow gluing algebra and fertilize with positive semidefinite matrices, you reap homomorphism density inequalities.

We can interpret the two parallel results, one with t and one with t^*, as two distinct gluing algebras. The first is what we described above. In the second, we glue the graphs together and then remove multiple edges. Both of these products result in a positive semidefinite connection matrix, so both obey the rules we describe here. For the rest of the notes, we'll only use the "simple" gluing product, but when you venture out to glue in the wild, you should remember both approaches.

Practically speaking, Corollary 9.26 says that if we expand the determinant of a special type of matrix, we get a posititve quantum graph. For example, for any k-labelled graphs F_1, F_2, and F_3, we have

$$\begin{vmatrix} [\![F_1F_1]\!] & [\![F_1F_2]\!] & [\![F_1F_3]\!] \\ [\![F_2F_1]\!] & [\![F_2F_2]\!] & [\![F_2F_3]\!] \\ [\![F_3F_1]\!] & [\![F_3F_2]\!] & [\![F_3F_3]\!] \end{vmatrix} \geq 0,$$

or in other terms,

$$\begin{aligned} [\![F_1F_1]\!][\![F_2F_2]\!][\![F_3F_3]\!] + [\![F_1F_2]\!][\![F_2F_3]\!][\![F_3F_1]\!] + [\![F_1F_3]\!][\![F_2F_1]\!][\![F_3F_2]\!] \\ \geq [\![F_1F_1]\!][\![F_2F_3]\!][\![F_3F_2]\!] + [\![F_1F_2]\!][\![F_2F_1]\!][\![F_3F_3]\!] + [\![F_1F_3]\!][\![F_2F_2]\!][\![F_3F_1]\!]. \end{aligned}$$

Corollary 9.27. *If F_1 and F_2 are two k-labelled graphs, then $[\![F_1F_2]\!]^2 \leq [\![F_1^2]\!][\![F_2^2]\!]$.*

Proof. Take Corollary 9.26 with $k = 2$ and expand the determinant. □

If this looks like Cauchy-Schwarz, that's because it is! It can also be proven directly:

$$\begin{aligned} t([\![F_1F_2]\!], W)^2 &= \left(\int_{[0,1]^k} t_x(F_1, W) t_x(F_2, W)\, dx \right)^2 \\ &\leq \int_{[0,1]^k} t_x(F_1, W)^2\, dx \int_{[0,1]^k} t_x(F_2, W)^2\, dx \\ &= t([\![F_1^2]\!], W)\, t([\![F_2^2]\!], W). \end{aligned}$$

Corollary 9.28. *If A is a positive semidefinite matrix and $F_1, \ldots, F_n$ are k-labelled graphs, then*

$$\sum_{i,j=1}^{n} a_{i,j} [\![F_iF_j]\!] \geq 0.$$

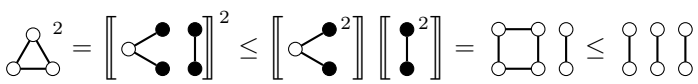
Proof. Apply Theorem 9.23 to Theorem 9.18. □

9.6 Voodoo

To make the pictures less cluttered, we adopt the convention that black nodes represent labelled vertices, and the labels are $1, 2, \ldots$ starting from the lower left corner and progressing counterclockwise. With that convention, here's a quick proof of Kruskal–Katona.[37]

Theorem 9.29 (Kruskal–Katona)**.** $K_3 \leq K_2^{3/2}$.

Proof.



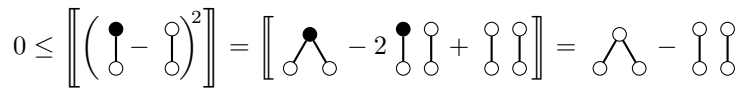
□

Here are two more speedy proofs.

[37] Except this is a trick: It's the same proof as before. Why?

Proposition 9.30. $P_3 \geq K_2^2$.

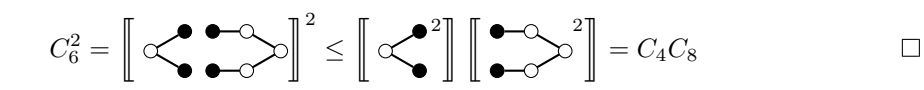
Proof.



□

Proposition 9.31. $C_6^2 \leq C_4 C_8$.

Proof.



□

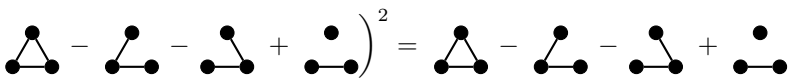
Exercise 9.32. Prove Proposition 9.30 using Cauchy-Schwarz (instead of the fact that squares are nonnegative).

Exercise 9.33. Generalize these results. What can you say about P_n and C_n for general n? Then use these techniques to find other relations.

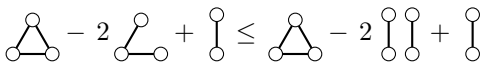
Mantel's Theorem also follows from Goodman's inequality. Here's a brief but tricksy proof.

Theorem 9.34 (Goodman's inequality)**.** $K_3 \geq K_2(2K_2 - 1)$.

Proof. It is straightforward to check that



After unlabelling, we apply Corollary 9.25 and Proposition 9.30 to get



Rearranging finishes the proof. □

Problem 9.35. The squared expression in the previous proof is called *idempotent*, which means that it is equal to its square. Can you find other idempotent expressions? Can you derive other interesting indequalities from them?

Exercise 9.36. The *n-wheel* W_n is the graph obtained by adding a vertex to the n-cycle and connecting it to every other vertex. Find an upper bound for $t(W_n, W)$ in terms of other homomorphism densities.

Exercise 9.37. Let K_n' denote the complete graph with one edge deleted. Prove that

$$t(K_{n+1}', W) \geq \frac{t(K_n, W)^2}{t(K_{n-1}, W)}$$

for every $W \in \mathcal{W}_0$.

Here's one more handy trick we can use. Recall that a convex function is a map $\phi\colon [a,b] \to \mathbb{R}$ such that $\phi(\alpha x + \beta y) \leq \alpha\phi(x) + \beta\phi(y)$ for every $\alpha, \beta \geq 0$ with $\alpha + \beta = 1$ and $x, y \in \mathbb{R}$. For example, $\phi(x) = x^2$ and $\phi(x) = |x|$ are convex functions on $[-1,1]$, whereas $\phi(x) = x^3$ is not. (Though it is convex on $[0,1]$.)

Theorem 9.38 (Jensen's Inequality)**.** *Let (Ω, μ) be a probability space and ϕ a convex function. For any integrable function $f\colon \Omega \to \mathbb{R}$,*

$$\phi\left(\int_\Omega f\, d\mu\right) \leq \int_\Omega \phi(f)\, d\mu.$$

Proof. Set $x_0 = \int_\Omega f\,d\mu$. Since ϕ is convex, there exists a constant m so that $\phi(x) \geq m(x - x_0) + \phi(x_0)$ for all $x \in [a, b]$. (A proof is outlined in Section A.4, but it should be at least believable after drawing a picture.) Then

$$\phi\left(\int_\Omega f\,d\mu\right) = m\left(\int_\Omega f\,d\mu - x_0\right) + \phi(x_0) = \int_\Omega (m(f - x_0) + \phi(x_0))\,d\mu \leq \int_\Omega \phi(f)\,d\mu. \qquad \square$$

Corollary 9.39. *If F is a k-labelled graph and $p \in \mathbb{N}$, then*

$$[\![F^p]\!] \geq [\![F]\!]^p.$$

Proof. Apply Jensen's inequality:

$$t([\![F^p]\!], W) = \int\limits_{[0,1]^k} t_x(F, W)^p\,dx \geq \left(\int\limits_{[0,1]^k} t_x(F, W)\,dx\right)^p = t([\![F]\!], W)^p. \qquad \square$$

With this, we can prove a non-trivial inequality very easily. Recall that $K_{a,b}$ denotes the complete bipartite graph whose bipartitions have size a and b.

Theorem 9.40. $K_{a,b} \geq K_2^{ab}$.

Proof. Let S_n denote the star on $n + 1$ vertices, the graph with one vertex of degree n and n vertices of degree 1, and let $S_n^\bullet$ denote the n-labelled version of S_n where each degree-1 vertex has a label. Then

$$K_{a,b} = [\![(S_a^\bullet)^b]\!] \geq (S_a)^b = [\![(S_1^\bullet)^a]\!]^b \geq K_2^{a,b},$$

since $S_1 = K_2$. $\square$

Since $P_3 = K_{1,2}$, this gives an alternate proof of Proposition 9.30. Also, $C_4 = K_{2,2}$, so this shows that $C_4 \geq K_2^4$, and $S_n \geq K_2^n$.

Exercise 9.41. Find a function f so that $P_n \geq K_2^{f(n)}$ for every $n \geq 1$.

There is a general conjecture underlying these results.

Conjecture 9.42 (Sidorenko)**.** *If F is bipartite, then $F \geq K_2^{\mathsf{e}(F)}$.*

Theorem 9.40 proves Sidorenko's conjecture for complete bipartite graphs. It has also been confirmed for all trees and even cycles, as well as a few other cases, but the general statement is still open. The smallest unconfirmed case has 10 vertices. (Also, the inequality in Sidorenko's conjecture does not hold for non-bipartite graphs, since $t(F, K_2) = 0$ if F is not bipartite, but $t(K_2, K_2)^{\mathsf{e}(F)} = (\frac{1}{2})^{\mathsf{e}(F)}$.)

Exercise 9.43. The case of even cycles can be handled using spectral analysis. Let G be a graph and A its adjacency matrix. Let $\text{ghost}_1 \geq \cdots \geq \text{ghost}_n$ be the (real) eigenvalues of A. Show that

$$\text{ghost}_1 \geq \frac{v^\top A v}{\|v\|^2}$$

for every $v \in \mathbb{R}^n$.[38] Using that $\mathsf{hom}(C_r) = \sum_{i=1}^n \text{ghost}_i^r$ and $\mathsf{hom}(K_2) = \sum_{i=1}^n \text{ghost}_i^2$, prove that $t(C_{2k}, G) \geq t(K_2, G)^{2k}$. (What happens if $v = \mathbf{1}$?)

Exercise 9.44. The proofs in this section can all be phrased directly in terms of the homomorphism density integral, without reference to the gluing algebra. Go back and rewrite some of the proofs in this form. Which do you prefer?

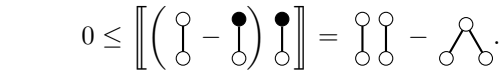
[38] This is called the *Rayleigh quotient*; it can be used to prove the Spectral Theorem and to approximate eigenvalues.

9.7 Advanced Voodoo

We start with a proposition.

Proposition 9.45. $K_2^2 \geq P_3$.

Proof. Clearly[39], $t(P_2, W) \geq 0$ for any graphon W and $P_2 - P_2 = 0 \geq 0$. Therefore their product is nonnegative, and

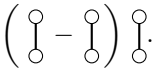


(9.5)

Rearranging finishes the proof. □

Wait, you say. We already proved that $P_3 \geq K_2^2$, so it must be that $t(P_3, W) = t(K_2, W)^2$ for every graphon W. This isn't true (take $W = W_{P_3}$), so the only logical conclusion is that we've discovered a contradiction deep at the heart of mathematics.

Hold on there, bucko. There's another possible resolution, and it's that Proposition 9.45 is false. As with most sticky parts of math, something's hiding underneath the word "clearly". The statements in that sentence are true, and it's also true that the product is nonnegative. But that product is

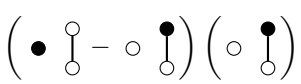


It has no labels, and there's no guarantee that the inequality remains true once we add labels. Indeed, we don't even know what it means for a labelled graph to be nonnegative, since we only defined $\geq$ for unlabelled graphs.

It's instructive to go back to the integral formulas. let F be an unlabelled simple graph and $F^\bullet$ be a k-labelled graph whose underlying graph is F. Then $[\![F^\bullet]\!] = F$, and $F \geq 0$ means that

$$\int_{[0,1]^k} t_x(F^\bullet, W)\, dx \geq 0$$

for every $W \in \mathcal{W}_0$. But $F \geq 0$ does not imply that $t_x(F^\bullet, W) \geq 0$ for all $x \in [0,1]^k$. Here is a more honest way to write the expression in (9.5):



It's now not at all clear that the left term is always nonnegative. Indeed, it isn't: If $K_2^\bullet$ denotes the 1-labelled K_2, then $t_x(K_1^\bullet K_2 - K_1 K_2^\bullet, W) = t(K_2, W) - \int_0^1 W(x,y)\, dy$. If, for example, $W = W_{P_3}$, then this expression is negative if $x \in (1/3, 2/3)$. So we can't just multiply labelled versions of nonnegative quantum graphs willy-nilly and expect the result to be nonnegative. But there's a way to actually get something out of this.

Definition 9.46. Let  be a k-labelled quantum graph (that is, each of its constituents is k-labelled). We write ≥ 0 if $t_x($$, W) \geq 0$ for every $W \in \mathcal{W}_0$.

If  , ≥ 0, then $[\![$$]\!] \geq 0$. So this fixes things. In fact, it not only fixes things but is also useful.

[39] This is when you should realize that something hinky is going on.

Proposition 9.47. *If F_1 and F_2 are simple graphs on the same vertex set, then*

$$(F_1 \cup F_2) + (F_1 \cap F_2) \geq F_1 + F_2.$$

Proof. Let $k = \mathsf{v}(F)$ and choose a k-labelling of $V(F_1) = V(F_2)$. Let $F_1^\bullet$, $F_2^\bullet$, and $(F_1 \cap F_2)^\bullet$ be the k-labelled graphs corresponding to F_1, F_2, and $F_1 \cap F_2)^\bullet$ that have this labelling. Since $F_1 \cap F_2 \subseteq F_1$, we have $t_x(F_1, W) \leq t_x(F_1 \cap F_2, W)$ for all $W \in \mathcal{W}_0$. Therefore $(F_1 \cap F_2)^\bullet - F_1^\bullet \geq 0$, and

$$0 \leq [\![\left((F_1 \cap F_2)^\bullet - F_1^\bullet\right) \left((F_1 \cap F_2)^\bullet - F_2^\bullet\right)]\!] = (F_1 \cap F_2) - F_1 - F_2 + (F_1 \cup F_2). \qquad \square$$

10 Extremal graph theory

In this section, we prove a generalization of Mantel's Theorem first published in 1941. This theorem is commonly regarded as the beginning of extremal graph theory. (Of course, the theorem was phrased without homomorphism densities.)

Theorem 10.1 (Turán)**.** *If G is a simple graph with $t(K_2, G) > 1 - \frac{1}{n-1}$, then $t(K_n, G) > 0$.*

We will prove it as a consequence of the following theorem on homomorphism inequalities.

Theorem 10.2. *Suppose that $\boxtimes$ is a quantum graph, each of whose constituents is a complete graph. Then $\boxtimes \geq 0$ if and only if $t(\boxtimes, K_n) \geq 0$ for every $n \geq 1$.*

To prove this theorem, we need just one preliminary. Let H be a node-weighted graph on n vertices (recall that each node i is assigned a weight $\alpha_i \geq 0$ and $\sum \alpha_i = 1$), and let $\mathcal{P} = \{V_1, \ldots, V_n\}$ be a partition of $[0,1]$ into intervals with $m(V_i) = \alpha_i$. We let W_H denote the $\{0,1\}$-valued step function on $\mathcal{P}$ with $W = 1$ on $V_i \times V_j$ if $ij \in E(H)$ and 0 otherwise. For a simple graph F, we define the homomorphism density of F in H by $t(F, H) = t(F, W_H)$. More explicitly, $t(F, H) = \sum_\phi \prod_{v \in V(F)} \alpha_{\phi(v)}$, where the sum ranges over all homomorphisms ϕ from F to the underlying graph of H.

Proof of Theorem 10.2. One direction is easy: If $\boxtimes \geq 0$, then certainly $t(\boxtimes, K_n) \geq 0$ for every $n \in \mathbb{N}$.

For the other direction, we proceed by contapositive: Supposing that $\boxtimes \not\geq 0$, we find a complete graph K that witnesses this, that is, such that $t(\boxtimes, K) < 0$. To that end, suppose that W is a graphon such that $t(\boxtimes, W) < 0$. Since the set of graphs is dense in graphon space, the set of node-weighted graphs is, as well. There is a node-weighted graph H with $t(\boxtimes, H) < 0$. We choose H to have the minimum number of vertices q among all such node-weighted graphs. Further, since $t(\boxtimes, H)$ is a continuous function of $\alpha_1, \ldots, \alpha_q$, we may choose a graph H on q vertices that minimizes $t(\boxtimes, H)$.

If H had a node with weight 0 and $\tilde{H}$ were the graph obtained by deleting that vertex, then $t(\boxtimes, H) = t(\boxtimes, \tilde{H})$. But $\tilde{H}$ has once fewer vertex than H; since H is minimal, all of its vertices have positive weight.

At this point, it is useful to consider the weights $\alpha_1, \ldots, \alpha_q$ as indeterminates. Assume that $\boxtimes = \sum_{i=1}^n a_i F_i$. Consider expanding the polynomial $t(\boxtimes, H) = \sum_{i=1}^n a_i \sum_{\phi_i} \prod_{v \in V(F_i)} \alpha_{\phi_i(v)}$, where the sum over ϕ_i is over every homomorphism $\phi_i \colon F_i \to H$. Since every F_i is complete, every homomorphism ϕ_i is injective. This means that $t(\boxtimes, H)$ considered as a polynomial in the variables $\alpha_1, \ldots, \alpha_q$ is multilinear; that is, no term has a variable with degree 2 or more.

We now show that H must be a (node-weighted) complete graph. Suppose to the contrary that H did not contain the edge ij. Temporarily fixing all variables other than α_i and α_j, we have $t(\boxtimes, H) = b_1 + b_2\alpha_i + b_3\alpha_j$ for some $b_1, b_2, b_3 \in \mathbb{R}$, since no term contains both α_i and α_j. Either $b_2 \leq b_3$ or $b_3 \leq b_2$; in the former case, consider the new graph H' obtained by deleting vertex α_j and replacing the weight on i by $\alpha_i + \alpha_j$. Then

$$t(\boxtimes, H') = b_1 + b_2(\alpha_i + \alpha_j) \leq b + b_2\alpha_i + b_3\alpha_j = t(\boxtimes, H) < 0.$$

But H' has $q - 1$ vertices, a contradiction to minimality. The other case is similar, so H is a complete graph.

Finally, we show that every vertex of H is weighted $1/q$. This will finish the proof, because then $W_H = W_{K_q}$, and $t(\boxtimes, K_q) < 0$. Because H is complete, the polynomial $\sum_{\phi_i} \prod_{v \in V(F_i)} \alpha_{\phi(v)}$ is symmetric (that is, it remains unchanged under any permutation of the indices $1, \ldots, q$). This is simply because any permutation of the indices corresponds to a

permutation of $V(H)$, and a homomorphism into a complete graph composed with a permutation is still a homomorphsim (and vice versa, every homomorphism is such a composition). It follows that $t(\bowtie, H)$ is symmetric in $\alpha_1, \dots, \alpha_q$, as well.

Choose any two vertices $i, j \in V(H)$. Temporarily fix the variables other than α_i and α_j, we have $t(\bowtie, H) = c_1 + c_2\alpha_i + c_3\alpha_j + c_4\alpha_i\alpha_j$ for some $c_1, c_2, c_3, c_4 \in \mathbb{R}$. But since $t(\bowtie, H)$ is symmetric, $c_2 = c_3$, so $t(\bowtie, H) = c_1 + c_2(\alpha_i + \alpha_j) + c_4\alpha_i\alpha_j$. Since we fixed all the other variables, $\alpha_i + \alpha_j = \sum_{r \neq i,j} \alpha_r$ is constant. So setting $c = c_1 + c_2(\alpha_i + \alpha_j)$ gives that $t(\bowtie, H) = c + c_4\alpha_i\alpha_j$. We chose $t(\bowtie, H)$ as a minimum, so α_i and α_j must be the values that minimize $c + c_4\alpha_i\alpha_j$. If $c_4 > 0$, this is minimized when $\alpha_i = 0$ or $\alpha_j = 0$, which contradicts the minimality of q. Otherwise $c_4 \leq 0$, in which case the expression is minimized when $\alpha_i = \alpha_j$. But we chose i and j arbitarily, so all node weights are equal. In particular, $\alpha_i = 1/q$ for every $i \in V(H)$. □

Since $t(K_k, K_m) = n^{\underline{k}}/n^k$, an equivalent formulation of Theorem 10.2 is that $\sum_{k=1}^n a_k K_k \geq 0$ if and only if $\sum_{k=1}^n a_k m^{\underline{k}}/m^k \geq 0$ for every $m \geq 1$. Turán's Theorem follows with a clever application of this restatement.

Proof of Theorem 10.1. We prove the inequality

$$n^n t(K_n, W) - (n-1)t(K_2, W) + (n-2) \geq 0 \tag{10.1}$$

for every graphon W. Setting $\bowtie = n^n K_n - (n-1)K_2 + (n-2)K_1$, we recognize that Theorem 10.2 can be applied. So to verify the inequality, we need only check that

$$f(m) := n^n \frac{m^{\underline{n}}}{m^n} - (n-1)\frac{m(m-1)}{m^2} + (n-2)$$

is nonnegative for every $n \geq 1$. If $m < n$, then the first term vanishes[40], and we get

$$f(m) = \frac{n-1}{m} - 1 \geq 0.$$

If $m \geq n$, then

$$f(m) > \left(\frac{n(m-n+1)}{m}\right)^n + \frac{n-1}{m} - 1.$$

Since $n(m-n+1) - n = (n-1)(m-n) \geq 0$, the first term is at least 1, so $f(m) > (n-1)/m > 0$. Setting $t(K_n, G) = 0$ in (10.1) shows that $t(K_2, G) \leq 1 - \frac{1}{n-1}$. □

To squeeze a bit more out of Theorem 10.2, we need a bit of convex geometry. A set $A \subseteq \mathbb{R}^n$ is *convex* if the line segment between any two points in A is contained in A. The smallest closed convex set that contains a given set A is called the *convex hull* of A, denoted $\text{conv}(A)$. We write $\text{conv}(x_1, \dots, x_n)$ for $\text{conv}(\{x_1, \dots, x_n\})$. There is an equivalent characterization (which we won't prove is equivalent[41]).

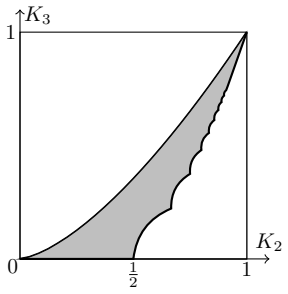
A *hyperplane* in $\mathbb{R}^n$ is an affine subspace of dimension $n-1$ (which means it's the translation of a linear subspace of dimension $n-1$). All hyperplanes can be written as $\{x \in \mathbb{R}^n : \langle x, \alpha \rangle = a\}$ for some $\alpha \in \mathbb{R}^n$ and $a \in \mathbb{R}$, and each such set is a hyperplane. (The vector α is called a *normal vector* to the hyperplane.) A *half-space* is, intuitively, one half of $\mathbb{R}^n$ that is cut off by a hyper plane; algebraically, a half-space is $\{x \in \mathbb{R}^n : \langle x, \alpha \rangle \geq a\}$ for some $\alpha \in \mathbb{R}^n$ and $a \in \mathbb{R}$.

Fact. The convex hull of A is the intersection of all half-spaces containing A.

Now fix some $n \geq 2$, and let $\mathbf{t}_W = (t(K_2, W), \dots, t(K_n, W))$ for every graphon W and set $\mathbf{t}_G = t_{W_G}$ for every graph G. We define $T_n = \{\mathbf{t}_W : W \in \mathcal{W}_0\} \subseteq \mathbb{R}^{n-1}$. For example, the set T_3 looks something like this:

[40] Poof!

[41] If you want to learn more, it is a consequence of the Hyperplane Separation Theorem.



You can see Mantel's Theorem exhibited on the horizontal axis, in that the gray region extends only to 1/2. The upper boundary of T_3 is the curve $K_3 = K_2^{3/2}$ from the Kruskal–Katona Theorem. We'll be able to prove some further results (including a uniqueness statement for Turán's Theorem) by studying this set. In fact, it's enough (for us) to study its convex hull.

Corollary 10.3. $\operatorname{conv}(T_n) = \operatorname{conv}(\mathbf{t}_{K_1}, \mathbf{t}_{K_2}, \mathbf{t}_{K_3}, \dots)$.

Proof. Certainly $\operatorname{conv}(\mathbf{t}_{K_n} : n \in \mathbb{N}) \subseteq \operatorname{conv}(T_n)$, since each $\mathbf{t}_{K_n} \in T_n$.

With every quantum graph $\boxtimes = \sum_{i=1}^n a_i K_i$ we associate the half-space $H_{\boxtimes} = \{x \in \mathbb{R}^{n-1} : \sum_{i=2}^m a_i x_{i-1} \geq -a_1\}$. This definition is cleverly chosen so that $\mathbf{t}_W \in H_{\boxtimes}$ if and only if $t(\boxtimes, W) \geq 0$. That is, $T_n \subseteq H_{\boxtimes}$ if and only if $\boxtimes \geq 0$. It follows from Theorem 10.2 that $T_n \subseteq H_{\boxtimes}$ if and only if $\mathbf{t}_{K_n} \in H_{\boxtimes}$ for every $n \geq 1$. But then the same half-spaces contain T_n as do $\{\mathbf{t}_{K_n} : n \geq 1\}$, so they have the same convex hull. □

From this, we can obtain the following sharpening of Turán's Theorem.

Corollary 10.4. *For every $n \geq 2$,*

$$\max\{t(K_2, W) : W \in \widetilde{\mathcal{W}_0} \text{ and } t(K_n, W) = 0\} = 1 - \frac{1}{n-1}$$

and the unique maximizing graphon is $W = W_{K_{n-1}}$.

Proof. The first statement is just Turán's Theorem rephrased. Setting $\boxtimes = n^n K_n - (n-1)K_2 + (n-2)K_1$, we found in the proof of Turán's Theorem that $t(\boxtimes, K_m) \geq 0$, and re-examining the proof shows that equality holds if and only if $m = n-1$. That is, $\mathbf{t}_{K_{n-1}}$ lies on the hyperplane boundary of $H_{\boxtimes}$. Since $\operatorname{conv}(T_n) = \operatorname{conv}(\mathbf{t}_{K_n} : n \geq 1)$ and the other $\mathbf{t}_{K_m}$ are uniformly bounded away from the boundary of $H_{\boxtimes}$ (which can be shown with a slightly more careful calculation of $f(m)$ when $m \geq n$), the only point $\mathbf{t}_W$ that is on the boundary of $H_{\boxtimes}$ is $W = W_{K_{n-1}}$. Setting $t(K_n, W) = 0$ in (10.1) and accounting for this information shows that $t(K_2, W) < 1 - \frac{1}{n-1}$ for every graphon W except $W_{K_{n-1}}$. □

This geometric way of looking at things even gives a slick proof of Goodman's inequality.

Alternate proof of Theorem 9.34. In T_3, we have

$$\mathbf{t}_{K_n} = \left(\frac{n-1}{n}, \frac{(n-1)(n-2)}{n^2}\right).$$

Each of these points is contained in the parametrization $\big(x, x(2x-1)\big)$. Since $y = x(2x-1)$ is a convex function, $\operatorname{conv}(T_3)$, and thus T_3 itself is contained in the convex hull of this parabola. In other words, if $t(K_3, W) \geq t(K_2, W)(2t(K_2, W) - 1)$ for every graphon W. □

This proof method can be used to extend Goodman's inequality to all complete graphs.

Exercise 10.5. If $d = t(K_2, W)$, then $t(K_m, W) \geq d(2d-1)(3d-2)\cdots\big((m-1)d - (m-2)\big)$.

Appendix: Miscellaneous proofs

A.1 Compactness of $[0,1]^{\mathbb{N}}$

Proposition A.6. *The set $[0,1]^{\mathbb{N}}$ is compact under the metric $d(x,y) = \sum_{i=1}^{\infty} 2^{-i}|x_i - y_i|$.*

Proof. Let $(x^{(n)}) \subset [0,1]^{\mathbb{N}}$. We want to find a convergent subsequence. First consider the sequence $(x_1^{(n)})_{n\in\mathbb{N}}$. Since $[0,1]$ is compact, it contains a convergent subsequence $(x_1^{(n)})_{n\in E_1}$. Similarly, there is a subset $E_2 \subseteq E_1$ so that $(x_2^{(n)})_{n\in E_2}$ converges. Inductively, we choose a set $E_i \subseteq E_{i-1}$ so that $(x_i^{(n)})_{n\in E_i}$ converges for all $i \in \mathbb{N}$. Now choose elements $n_k \in E_k$ so that $n_k > n_{k-1}$ for all $k \in \mathbb{N}$ (choose $n_1 \in E_1$ arbitrarily). For each $i \in \mathbb{N}$, the sequence $(x_i^{(n_k)})_{k\in\mathbb{N}}$ is eventually a subsequence of $(x_i^{(n)})_{n\in E_i}$, so $(x^{(n_k)})$ converges pointwise so some $x \in [0,1]$.

Now we show that $d(x^{(n_k)}, x) \to 0$. For any $\varepsilon > 0$, choose a $K \in \mathbb{N}$ so that $2^{-K} < \varepsilon$. Since $x^{(n_k)} \to x$ pointwise, we consider the first K coordinates to see that

$$d(x^{(n_k)}, x) = \sum_{i=1}^{\infty} 2^{-i}|x_i^{(n_k)} - x_i| \le \sum_{i=1}^{K} 2^{-i}|x_i^{(n_k)} - x_i| + \sum_{i=K+1}^{\infty} 2^{-i} \longrightarrow 2^{-K} < \varepsilon.$$

Since ε was arbitrary, $d(x^{(n_k)}, x) \to 0$. □

A.2 Cut distance of graph kernels

This section outlines a proof of Proposition 5.19.

Let G_1 and G_2 be two graphs on n_1 and n_2 vertices, respectively, and set $W_1 = W_{G_1}$ and $W_2 = W_{G_2}$. One inequality is straightforward. We define the special partition $\mathcal{I}(m) = \{[k/m, (k+1)/m] : 0 \le k < m\}$ of $[0,1]$ into m intervals. Each vertex in $W_{G_1(N)}$ corresponds to an interval in $\mathcal{I}(n_1 N)$. So any bijection $\phi\colon V(G_1(N_1)) \to V(G_2(N_2))$ with $n_1 N_1 = n_2 N_2$ induces a bijection ϕ' of the intervals in $\mathcal{I}(n_1 N_1)$ and therefore a measure-preserving bijection of $[0,1]$. Then

$$d_\square(G_1^\phi, G_2) = \|W_1^{\phi'} - W_2\|_\square \ge \delta_\square(W_1, W_2).$$

Taking the minimal ϕ shows that $\widehat{\delta}_\square(G_1, G_2) \ge \delta_\square(W_1, W_2)$. Since $W_{G(n)} = W_G$ for all n, the inequality holds for all blowups of G_1 and G_2, so $\delta_\square(G_1, G_2) \ge \delta_\square(W_1, W_2)$. In short, each bijection $V(G_1(N_1)) \to V(G_2(N_2))$ is a measure preserving map applied to W_1; since the cut distance for kernels is the infimum over *all* measure-preserving maps, the kernel cut distance is at most the graph cut distance.

The other direction is more technical. Let $\{S_1, \dots, S_{n_1}\}$ and $\{T_1, \dots, T_{n_2}\}$ be the interval steps of W_1 and W_2, respectively. Fix a measure-preserving bijection $\phi\colon [0,1] \to [0,1]$. We show that $\inf_{n_1 N_1 = n_2 N_2} \widehat{\delta}_\square(G_1(N_1), G_2(N_2)) \le \|W_1^\phi - W_2\|_\square$. Taking the infimum over all ϕ then proves that $\delta_\square(G_1, G_2)) \le \delta_\square(W_1, W_2)$.

If υ is a measure-preserving bijection of $[0,1]$ such that $\upsilon(T_j) = T_j$ for all $1 \le j \le n_2$, then $\upsilon \circ \phi$ is a measure-preserving function, and

$$\|W_1^{\upsilon\circ\phi} - W_2\|_\square = \|W_1^{\upsilon\circ\phi} - W_2^\upsilon\|_\square = \|(W_1^\phi - W_2)^\upsilon\|_\square. = \|W_1^\phi - W_2\|_\square.$$

Define $\alpha_{i,j} = m(\phi(S_i) \cap T_j)$. Fix some N divisible by $\operatorname{lcm}(n_1, n_2)$. Carve up T_j into intervals of length $\lfloor \alpha_{i,j} N \rfloor / N$ for each $1 \le i \le n_1$ and one of length $\sum_{i=1}^{n_1} (\alpha_{i,j} - \lfloor \alpha_{i,j} N \rfloor / N)$. We choose a measure-preserving bijection υ_N that sends part of $\phi(S_i) \cap T_j$ to the interval of length $\lfloor \alpha_{i,j} N \rfloor / N$ in T_j, and the rest of all of the sets $\{\phi(S_i) \cap T_j\}_i$ to the remaining

interval of length $\sum_{i=1}^{n_1}(\alpha_{i,j} - \lfloor \alpha_{i,j}N \rfloor/N)$ in T_j.[42] Further, we define a measure-preserving bijection $\phi_N \colon [0,1] \to [0,1]$: Carve up each set S_i into intervals of length $\lfloor \alpha_{i,j}N \rfloor/N$ for each $1 \le j \le n_2$ and one of length $\sum_{j=1}^{n_1}(\alpha_{i,j} - \lfloor \alpha_{i,j}N \rfloor/N)$. The map ϕ_N takes the interval of length $\lfloor \alpha_{i,j}N \rfloor/N$ in S_i to the of the same length in $\upsilon_N(\phi(S_i) \cap T_j)$; the remaining interval of length $\sum_{j=1}^{n_1}(\alpha_{i,j} - \lfloor \alpha_{i,j}N \rfloor/N)$ can be partitioned into intervals of length $1/N$, and those are mapped injectively in such a manner that each of their images is also an interval of length $1/N$.

The kernels $W_1^{\upsilon_N \circ \phi}$ and $W_1^{\phi_N}$ agree on a set of measure at least

$$\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \frac{\lfloor N\alpha_{i,j} \rfloor}{N} \ge \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \left(\alpha_{i,j} - \frac{1}{N}\right) = 1 - \frac{n_1 n_2}{N}.$$

So $W_1^{\phi_N} \to W_1^{\upsilon_N \circ \phi}$ pointwise as $N \to \infty$, which implies that $\|W_1^{\phi_N} - W_1^{\upsilon_N \circ \phi}\|_1 \to 0$. Since ϕ_N is a map on intervals of length $1/N$, so it induces a vertex map $\tilde{\phi} \colon V(G_1(N/n_1)) \to V(G_2(N/n_2))$. Therefore

$$\begin{aligned}
\widehat{\delta}_\square(G_1(N/n_1), G_2(N/n_2)) &\le d_\square(G_1(N/n_1)^{\tilde{\phi}}, G_2(N/n_2)) \\
&= \|W_1^{\phi_N} - W_2\|_\square \\
&\le \|W_1^{\phi_N} - W_1^{\upsilon_N \circ \phi}\|_1 + \|W_1^{\upsilon_N \circ \phi} - W_2\|_\square \\
&= \|W_1^{\phi_N} - W_1^{\upsilon_N \circ \phi}\|_1 + \|W_1^{\phi} - W_2\|_\square.
\end{aligned}$$

Taking the limit as $N \to \infty$ shows that $\delta_\square(G_1, G_2) \le \|W_1^{\phi} - W_2\|_\square$. Then taking the infimum over all ϕ finishes the proof. □

A.3 Maximization of cut norm

Proposition A.7. *For any kernel $W \in \mathcal{W}$, there exist sets $S, T \subseteq [0,1]$ so that*

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

The outline of the proof is that we consider the more general integral

$$H(f,g) = \left| \int_{[0,1]^2} f(x)W(x,y)g(y)\,dx\,dy \right|,$$

where $f, g \colon [0,1] \to [0,1]$ are measurable and show that the set of such integrals attains its supremum. We then obtain sets S and T that maximize $\left|\int_{S\times T} W(x,y)\,dx\,dy\right|$ from a pair of maximizing functions f and g.

The most difficult part of the proof is showing that $H(f,g)$ attains a maximum. To prepare, let's do some fast-paced functional analysis.

A.3.1 Weak topologies

A topological field k is a field equipped with a topology such that addition, multiplication, subtraction, and division are continuous. A topological vector space over k is a k-vector space together with a topology where vector addition and scalar multiplication are continuous.

[42] Such a map exists! Try constructing one.

Definition A.8. Let k be a topological field and X be a topological vector space over k. The *continuous dual space* X^* of X is the k-vector space of continuous linear functionals from X to k.

The continuous dual space is a restriction of the usual dual space of a vector space to continuous functionals.

Definition A.9. The *weak topology* on X is the coarsest topology such that each functional in X^* remains a continuous function. In other words, the weak topology is generated by $\{\phi^{-1}(V) : \phi \in X^* \text{ and } V \text{ is open in } k\}$. We say that a sequence $(x_n) \subseteq X$ *weakly converges* to $x \in X$ if it converges in the weak topology. In this case, we write $x_n \xrightarrow{w} x$.

What exactly *are* the sets in the weak topology, though? Well, the topology generated by a set $S \subseteq X$ can be constructed as follows:

1. Add $\emptyset, X$ to S.
2. Take all finite intersections in $S \cup \{\emptyset, X\}$; call this collection S_I.
3. Take all (not necessarily finite) unions of elements of S_I; call this S_τ.

It's clear that S_τ is a subset of the topology generated by S, and it's fairly straightforward to check that S_τ is closed under finite intersections and arbitrary unions.[43] So S_τ is the topology generated by S.

The strength of the weak topology[44] comes from the following result.

Proposition A.10. *A sequence $(x_n) \subseteq X$ weakly converges to $x \in X$ if and only if $\phi(x_n) \to \phi(x)$ for every $\phi \in X^*$.*

Proof. The forward direction follows directly from the fact that every $\phi \in X^*$ is continuous under the weak topology. For the other direction, assume that x_n does not weakly converge to x, so that there is an open set V in the weak topology such that $\{n \in \mathbb{N} : x_n \notin V\}$ is infinite. Because V is in the weak topology, it can be written as $V = \bigcup_\alpha V_\alpha$, where each V_α is the finite intersection of preimages of open sets in k.

Each set $\{n \in \mathbb{N} : x_n \notin V_\alpha\}$ is infinite, so choose some V_β. It is a finite intersection $V_\beta = \bigcap_{i=1}^m \phi_i^{-1}(U_i)$ where $\phi_i \in X^*$ and the U_i are open in k. This means that at least one set $\{n \in \mathbb{N} : x_n \notin \phi_j^{-1}(U_j)\}$ is infinite. Then $\{\phi_j(x_n) \in k : \phi(x_n) \notin U_j\}$ is infinite, so $\phi_j(x_n) \not\to \phi(x)$. □

There is a natural map from X into its double dual $(X^*)^*$: We send each $x \in X$ to the evaluation map $T_x(\phi) = \phi(x)$. The weak* topology on X^* is the coarsest topology so that each T_x is continuous. That is, it is the weak topology induced by the set $\{T_x : x \in X\}$. If ϕ_n converges to ϕ in the weak* topology, we write $\phi_n \xrightarrow{w^*} \phi$. Adapting Proposition A.10 to the weak* topology yields the following result.

Proposition A.11. *A sequence of continuous linear functionals $(\phi_n) \subseteq X^*$ converges to $\phi \in X^*$ in the weak* topology if and only if $\phi_n(x) \to \phi(x)$ for each $x \in X$.*

So weak* convergence is pointwise convergence.

Now let's turn to a specific application. Let (Ω, μ) be a measure space and $X = L^1(\Omega)$. We have the following theorem.

Fact A.12. *If (Ω, μ) is σ-finite (that is, Ω is the countable union of sets of finite measure), then $L^\infty(\Omega) = (L^1(\Omega))^*$ by associating $h \in L^\infty(\Omega)$ with the linear functional $\tilde{h}(\zeta) = \int_\Omega \zeta h \, d\mu$.*

[43] This is not the case if we reverse steps 2 and 3.

[44] if you will

One part of the fact is easy to check: If $h \in L^\infty(\Omega)$, then $\tilde{h}$ is a linear functional on $L^1(\Omega)$. The main point of the theorem is that we can transfer the weak* topology to $L^\infty(\Omega)$ via this correspondence. A sequence $(h_n) \subseteq L^\infty(\Omega)$ converges in the (induced) weak* topology to $h \in L^\infty(\Omega)$ if and only if $\tilde{h}_n$ converges pointwise, that is, if $\int_\Omega \zeta h_n \, d\mu \to \int_\Omega \zeta h \, d\mu$ for all $\zeta \in L^1(\Omega)$. At this point, we need one more fact.

Fact A.13 (Banach—Alaoglu theorem)**.** *Let X be a topological vector space. The closed unit ball in X^* is compact with respect to the weak* topology.*

Applying this to $X = L^1(\Omega)$, we find that: Any sequence $(h_n) \subseteq L^\infty(\Omega)$ contained in the unit ball of $L^\infty(\Omega)$ contains a subsequence (h_{n_k}) that is weak*-convergent to a function $\hat{h} \in L^\infty(\Omega)$, that is, so that $\int_\Omega \zeta h_{n_k} \, d\mu \to \int_\Omega \zeta \hat{h} \, d\mu$ for all $\zeta \in L^1(\Omega)$. In the proof, we'll take $\Omega = [0,1]$ at some points and $\Omega = [0,1]^2$ at others.

A.3.2 Proof of Proposition A.7

Recall that

$$H(f,g) = \left| \int\limits_{[0,1]^2} f(x) W(x,y) g(y) \, dx \, dy \right|.$$

We may drop the absolute value by changing the sign of W if necessary to make the integral positive. We first show that there are measurable functions $\hat{f}, \hat{g} \colon [0,1] \to [0,1]$ so that $H(\hat{f},\hat{g}) = \sup_{f,g} H(f,g) =: I$, where the supremum is taken over measurable functions $f, g \colon [0,1] \to [0,1]$. Take a sequence of functions $H(f_n, g_n) \to I$. Since (f_n) and (g_n) are in the unit ball of $L^\infty(\Omega)$, we can choose a subsequence $f_{n_k} \xrightarrow{w^*} \hat{f}$ and a further subsequence $g_{n_{k_m}} \xrightarrow{w^*} \hat{g}$. To ease the burden on the subindices, let's assume that the original sequences converge in the weak* topology to $\hat{f}$ and $\hat{g}$.

We know that $\hat{f}$ and $\hat{g}$ are in the closed unit ball of $L^\infty([0,1])$, so $\hat{f}, \hat{g} \colon [0,1] \to [-1,1]$. We want to show that their range is actually in $[0,1]$ (up to a set of measure 0). Suppose for the sake of contradiction that $\hat{f}$ was negative on a set S of positive measure. Since $f_n \geq 0$ and f_n weak*-converges to $\hat{f}$, we have

$$0 > \int\limits_{[0,1]} \chi_S \hat{f} \, dx = \lim_{n\to\infty} \int\limits_{[0,1]} f_n \chi_S \, dx \geq 0,$$

which is a contradiction. Therefore the images of $\hat{f}$ and $\hat{g}$ are in $[0,1]$.

At this point we have functions $\hat{f}, \hat{g} \colon [0,1] \to [0,1]$, so it would be nice to claim that $H(\hat{f},\hat{g})$ are maximal by writing

$$H(\hat{f},\hat{g}) = \int\limits_{[0,1]^2} \hat{f}(x) W(x,y) \hat{g}(y) \, dx \, dy = \lim_{n\to\infty} \int\limits_{[0,1]^2} f_n(x) W(x,y) g_n(y) \, dx \, dy = I. \tag{A.2}$$

However, it's not clear that the tensor product function $f_n \otimes g_n(x,y) = f_n(x) g_n(y)$ weak*-converges to $\hat{f} \otimes \hat{g}$, so the limit step is not justified. However, it is true, so we now finish up the first part of the proof by justifying this step.

We define the function $\eta_n = f_n \otimes g_n$; that is, $\eta_n \colon [0,1]^2 \to [0,1]$ by $\eta_n(x,y) = f_n(x) g_n(y)$. Then (η_n) is in the closed unit ball of $L^\infty([0,1]^2)$, so some subsequence weak*-converges to a function η in the unit ball. We claim that $\eta = \hat{f} \otimes \hat{g}$.

The function η induces a measure on $[0,1]^2$ by $\mu_\eta(S) = \int_S \eta(x,y)\,dx\,dy$. Similarly, $\hat{f} \otimes \hat{g}$ induces the measure $\mu_{\hat{f}\otimes\hat{g}}(S) = \int_S \hat{f}(x)\hat{g}(y)\,dx\,dy$. Showing that $\eta = \hat{f} \otimes \hat{g}$ almost everywhere is equivalent to showing that $\mu_\eta = \mu_{\hat{f}\otimes\hat{g}}$.

Lemma A.14. *Let (X,μ) be a measure space and f and g be measurable functions on X. Define the measure $\mu_f(S) = \int_S f\,d\mu$. Then $f = g$ almost everywhere if and only if $\mu_f = \mu_g$.*

Proof. If $f = g$ almost everywhere, then $\mu_f(S) = \int_S f\,d\mu = \int_S g\,d\mu = \mu_g(S)$ for all measurable sets S. Now suppose that $f \neq g$ on a set S with $\mu(S) > 0$. Then either $\{x \in S : f(x) > g(x)\}$ or $\{x \in S : f(x) < g(x)\}$ has positive measure. The arguments are symmetric, so suppose that $f > g$ on S. Then

$$\mu_f(S) - \mu_g(S) = \int_S f\,d\mu - \int_S g\,d\mu = \int_S (f-g)\,d\mu > 0,$$

so $\mu_f \neq \mu_g$. □

Now we show that $\mu_\eta = \mu_{\hat{f}\otimes\hat{g}}$. To verify this, we need only check a basis for the measure space, so we'll check product sets $S \times T$, where $S, T \subseteq [0,1]$. Using that $\eta_n \xrightarrow{w^*} \eta$ and that $\chi_S\chi_T \in L^1([0,1]^2)$, we have

$$\begin{aligned}
\mu_\eta(S\times T) &= \int\limits_{[0,1]^2} \chi_S(x)\chi_T(y)\eta(x,y)\,dx\,dy \\
&= \lim_{n\to\infty} \int\limits_{[0,1]^2} \chi_S(x)\chi_T(y)\eta_n(x,y)\,dx\,dy \\
&= \lim_{n\to\infty} \int\limits_{[0,1]} \chi_S(x) f_n(x)\,dx \int\limits_{[0,1]} \chi_T(y) g_n(y)\,dy \\
&= \mu_{\hat{f}}(S)\mu_{\hat{g}}(T) = \mu_{\hat{f}\otimes\hat{g}}(S\times T).
\end{aligned}$$

It follows that $\eta = \hat{f} \otimes \hat{g}$. Using weak* convergence of (η_n) in equation (A.2) shows that $H(\hat{f},\hat{g})$ is maximal.

To finish the proof, we show that we can choose a maximal f and g that are $\{0,1\}$-valued. Let f and g be functions so that $H(f,g)$ is maximal and define

$$f_1(x) = \begin{cases} 1 & \text{if } f(x) \geq \frac{1}{2} \\ 0 & \text{if } f(x) < \frac{1}{2}. \end{cases} \qquad \text{and} \qquad f_2(x) = \begin{cases} 2f(x) - 1 & \text{if } f(x) \geq \frac{1}{2} \\ 2f(x) & \text{if } f(x) < \frac{1}{2}. \end{cases}$$

Then $f = \frac{1}{2}f_1 + \frac{1}{2}f_2$. Using the triangle inequality and the maximality of $H(f,g)$, we have

$$H(f,g) \leq H\left(\frac{1}{2}f_1 + \frac{1}{2}f_2, g\right) \leq \frac{1}{2}H(f_1,g) + \frac{1}{2}H(f_2,g) \leq H(f,g).$$

Equality must hold in each step, so in particular $H(f_1,g) = H(f,g)$ is maximal and f_1 is $\{0,1\}$-valued. (This is a more general technique called convex optimization; the same steps as above show that the vertices of any convex combination will also be optimal.) We can now follow the same procedure by fixing f_1 and finding a $\{0,1\}$-valued g_1 so that $H(f_1,g_1)$ is maximal.

Every characteristic function χ_S is a measurable function from $[0,1]$ to $[0,1]$, so

$$\|W\|_\square = \sup_{S,T\subseteq[0,1]} \left| \int_{S\times T} W(x,y)\,dx\,dy \right| \leq \sup_{f,g\colon [0,1]\to[0,1]} \left| \int_{[0,1]^2} f(x)W(x,y)g(y)\,dx\,dy \right|.$$

But f_1 and g_1 are characteristic functions that achieve the supremum on the right, so the sets $S = f_1^{-1}(1)$ and $T = g_1^{-1}(1)$ maximize the supremum for $\|W\|_\square$. This finishes the proof. □

A.4 Subderivatives

Definition A.15. Let $\phi\colon [a,b]\to\mathbb{R}$ be a convex function. A *subderivative* of ϕ at $x_0\in[a,b]$ is a point $c\in[a,b]$ so that $\phi(x)\geq c(x-x_0)+\phi(x_0)$ for every $x\in[a,b]$.

We think of a subderivative as a line passing through $(x_0,\phi(x_0))$ that lies below the curve of ϕ in the plane. For example, $\phi(x)=|x|$ has no derivative at $x=0$, but it does have a subderivative. In fact, any real number in the interval $[-1,1]$ is a subderivative of $|x|$ at $x=0$.

Definition A.16. The *subdifferential* of a convex function at x_0 is the set of its subderivatives at x_0.

Proposition A.17. *Let $\phi\colon [a,b]\to\mathbb{R}$ be a convex function and $x_0\in[c,d]$. Both of the limits*

$$c = \lim_{x\to x_0^-} \frac{\phi(x)-\phi(x_0)}{x-x_0} \quad \textit{and} \quad d = \lim_{x\to x_0^+} \frac{\phi(x)-\phi(x_0)}{x-x_0}$$

exist, $c\leq d$, and the subdifferential of ϕ at x_0 is the interval $[c,d]$.

Proof sketch. Using convexity of ϕ, you can show that if $x<y$, then

$$\frac{\phi(x)-\phi(x_0)}{x-x_0} \leq \frac{\phi(y)-\phi(x_0)}{y-x_0}.$$

It follows that

$$\lim_{x\to x_0^-} \frac{\phi(x)-\phi(x_0)}{x-x_0} = \sup_{x<x_0} \frac{\phi(x)-\phi(x_0)}{x-x_0} \quad \text{and} \quad \lim_{x\to x_0^+} \frac{\phi(x)-\phi(x_0)}{x-x_0} = \inf_{x>x_0} \frac{\phi(x)-\phi(x_0)}{x-x_0},$$

so both limits exist. Once you have c and d at hand, the remaining two statements also follow from convexity. (Each piece can be proven by assuming the negation and obtaining a contradiction to convexity.) □

Corollary A.18. *If $\phi\colon [a,b]\to\mathbb{R}$ is a convex function and $x_0\in[a,b]$, then there is a real number m so that $\phi(x)\geq m(x-x_0)+\phi(x_0)$ for every $x\in[a,b]$.*

References

[1] C. Borgs, J. Chayes, L. Lovász, V. Sós, and K. Vesztergombi. Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing. *Advances in Math.*, 219:1801–1851, 2008.

[2] L. Lovász. *Large Networks and Graph Limits*, volume 60 of *Colloquium Publications*. American Mathematical Society, 2012.

[3] Y. Zhao. Mit 18.217 lecture notes: Graph theory and additive combinatorics, Fall 2019. http://yufeizhao.com/gtac/gtac.pdf.