Expander Graphs: Notes on Notes

Joshua Erde

Department of Mathematics,
TU Graz.

Note: *These are my own personal notes which take most of the survey article "Expander graphs and their applications" of Hoory, Linial and Wigderson verbatim and add in some extra detail in places. They are meant solely as an aid for lecturing, and are not intended for further distribution. I'm making them available in case they are useful for my students. Any errors contained within are entirely my own.*

Contents

1 Introduction

So, very roughly, expander graphs are graphs which are simultaneously very sparse (and so have few edges), but very well-connected. These graphs were first studied in the early 70s, but since then it has turned out that the property of being an expander is significant in many different contexts, mathematical, but also physical and computational. In computer science they come up very naturally in the design and analysis of communication networks, but also have surprising applications to many other topics, such as error-correcting codes, the theory of pseudorandomness and the analysis of Monte-Carlo algorithms. In mathematics, outside of graph theory, they have found use in the theory of Metric embeddings (embedding metric spaces into Euclidean spaces) which also turns out to have applications in computing, the convergence of Markov chains, Sieve theory in both an arithmetic (i.e number theoretic) and algebraic (i.e. group theory) setting and to the study of hyperbolic manifolds, to name just a small selection of topics. We will touch on a few of these applications in the course.

Beyond this, and partly because of it this, it seems that expansion is a fundamental concept that deserves to be investigated in it's own right. One reason for the ubiquity of, and the interest in expander graphs comes from the fact that the notion of expansion can be cast in multiple different forms: Combinatorially, expander graphs can be defined in terms of their high connectivity - to disconnect a larger part of the graph we have to delete many edges. Geometrically, they can be defined in terms of the isoperimetric constant - small sets have to have large boundaries. Probabilistically, they can be defined in terms of the the mixing time of random walks - if we take a random walk on the graph, how long do we have to walk until we're 'fully lost'. Finally, algebraically, expanders can be defined by thinking about the graph as an operator, determined by it's adjacency matrix, and the spectrum of this operator (the eigenvalues).

These four perspectives offer different tools to study expander graphs from, as well as suggesting interesting questions about the connections between the various definitions (i.e., qualitatively how does one expansion notion affect the others)

Let me start then by giving a rough overview of the structure of course.

1. Three motivating examples
    - We start by giving essentially an introductory lecture, or series of lectures, where we give three motivating examples (where expander graphs arise in surprising contexts). Namely for constructing error correcting codes, for deterministically improving random algorithms and for analysing the hardness of linear transformations (how efficiently can we compute linear transformations). The hope is that this gives us a little taste of what expander graphs are about, and how they are used, to motivate the rest of the course.
2. A spectral view of expansion
    - We then move on to defining expansion algebraically, by relating the expansion to the eigenvalues of a graph. We'll talk about various extremal questions about spectral expansion and also the 'Expander mixing lemma' which is an important tool relating the spectrum to the edge distribution and notions of pseudorandomness.
3. Random walks on expander graphs

- We then talk about the probabilistic definition of expansion, looking at random walks on expander graphs and showing that they rapidly mix, i.e approach the limiting distribution quickly. We can use this intuition to use random walks to sample vertices efficiently from the graph, and we'll give some applications of this idea in computer science

4. A geometric view of expansion

   - We then move on to the geometric definition of expansion, where we relate expansion to the geometric notion of isoperimetry. An important idea here is a discrete analogue of the Laplace operator, whose spectrum is related to that of the graph. Cheeger's inequality in Riemannian geometry relates the spectrum of the Laplace operator to the isoperimetric constant, and the discrete analogue here is fundamental to relating this view to the algebraic view.

5. Extremal problems on spectrum and expansion

   - Having introduced the various types of expansion, we focus on the extremal aspects - how large can the various notions of expansion be, both in themselves, and with respect to each other.

6. The Margulis construction, the Zig-zag product and lossless conductors.

   - We finally give some explicit constructions of expander graphs. The first is the Margulis construction, which is algebraic in nature, and to prove it's expansion we'll use some discrete fourier analysis. We then demonstrate a more flexible way of building expander graphs explicitly and inductively using a graph product, here the analysis will depend on Entropic methods, an idea coming from Shannon's seminal work on communication. Finally we will consider the problem of constructing optimal vertex expanders using the zig-zag product, which turns out to be related to randomness preserving and enhacing objects such as conductors and extractors.

7. Application to metric embeddings

   - To round off the course we'll briefly talk about the relationship of expanders with problems of metric embeddings, showing how the graph metric of expanders graphs are examples of metrics which are in some sense as far from Euclidean as possible, and how the question of embedding metric spaces with low distortion can be used to infrom algorithms for determining the expansion of a graph.

2 Three motivating problems

2.1 Hardness for linear transformations

A very natural question to consider in the context of computing is, given some library or set of basic operations and some goal function, how simple can an algorithm be that computes this goal only using these operations.

For example, we might be given some multinomials or linear transformations over a field, and we use to compute the output from a given input, but only using addition and multiplication.

One very simple model for this would be a *straight line program*, we have a set of variable names $X = \{x_i\}$ and some collection of basic operations $F = \{f_e\}$ and we have a finite sequence of instructions (or gates) of the form $I_i = `x_i = f_e(x_j, x_k)'$, with the restriction that the variable x_i appearing on the left hand side does not occur previously in the sequence.

It's natural to model this as a directed graph D (or at least a labelled directed graph) where we put an edge from x_j and x_k to x_i for each gate $x_i = f_e(x_j, x_k)$. The sources of D are the *input nodes* and the sinks the *output nodes*.

One particularly simple type of straight line program would be a *linear program* where the allowed operations are of the form $f_{a,b}(x_j, x_k) = ax_j + bx_k$ for some $a, b \in \mathbb{R}$.

The following problem is then very natural

Question 2.1. *Let A be an $n \times n$ matrix over a field $\mathbb{F}$ (nb. this turns out toe be interesting for finite fields as well). What is the least number of gates in a linear program, or equivalently what is the least number of edges in the associated graph, which computes the transformation $x \mapsto Ax$?*

Not only is this question obviously very intersting theoretically, it has genuine technological significance. Indeed if we let $a_{ij} = \omega^{ij}$ where ω is a primitive nth root of unity then this transformation is the *discrete Fourier transform* or DFT, which is fundamental to many modern technologies involving signal processing, spectral analysis, machine learning etc.

A direct computation of the DFT from the definitions would use $O(n^2)$ gates, however the so called *fast fourier transform* of FFT, designed by Cooley and Tukey, improves this to $O(n \log n)$, which is actually a huge improvement in practical terms.

Since you need at least $O(n)$ gates just to specify the input, the FFT is close to the theoretical limit in terms of complexity. Whilst this $\log n$ gap might seem small, it is quite significant, the existence of a *VFFT* using only $O(n)$ gates would have genuine technological consequences, whereas on the other hand establishing the necessity of $\Omega(n \log n)$, or even just $\omega(n)$ gates would be a huge theoretical breakthrough.

For every finite field $\mathbb{F}$ it is fairly easy to show that *most* $n \times n$ matrices A require $\Omega\left(\frac{n^2}{\log n}\right)$ gates, just based on a double counting argument - Count the number of possible linear programs with a given number of gates and count the number of matrices. However, as is often the case

with computational complexity, even though we know that computationally hard functions are abundant, it is very difficult to exhibit specific, explicit linear transformations which require more than $O(n)$ gates.

It was discovered by Valiant that for certain types of matrices A, the graph of any linear program which computes this transformation must have certain connectivity properties. A general property of A which guarantees this is the following: We say A is *super regular* if every square submatrix of A has full rank.

Roughly one can show that the outputs of such a transformation are so intricately dependent on the inputs, that it is impossible to separate them in the graph easily. This leads to the definition of a *super concentrator* a graph $G = (V, E)$ with two sets I and O of *inputs* and *outputs* such that for every k and every $S \subseteq I$ and $T \subseteq O$ with $|S| = |T| = k$ there exists a set of k vertex disjoint paths in G from S to T.

It is a simple exercise to show that the underlying graph of any linear program which computes the transformation of a super regular matrix is a superconcentrator. Valiant conjectured that any super concentrator must have $\omega(n)$ many edges. Note that the DFT is not super regular, although Valiant also showed it must satisfy a slightly weaker connectivity property (in terms of bilinear programs for example the convolution is super regular)

However, Valiant himself disproved this conjecture, using expander graphs to construct super concentrators with $O(n)$ edges, which can be used to give an explicit super regular matrix A that has a linear program of linear complexity. We'll go into more details of this at the end of this section.

This might seem like the end of the story, but it turned out that Valiant was thinking along the right lines, and in a closely related setting these superconcentrators turned out to be essential. Roughly if we consider programs with more than two input per gate allowed, but where we restrict the *depth* of the program (length of a path from input to output) then $\omega(n)$ lower bounds for the number of edges in the corresponding graphs can be proven. From this it can be deduced that for certain types of linear transformations there are superlinear bounds for the complexity of computing them using bounded depth linear circuits.

2.2 Error correcting codes

A fundamental problem in communication is noise. Suppose Alice has a message consisting of k bits which she would like to transmit to Bob over some noisy communication channel. The noise means that Bob might receive a different message to the one that Alice sends.

A very simple model of this noise (although not a particular good one) would to be to assume that the noise won't change more than a certain fraction p of the bits (more reasonably you might model it as a probability that a bit is corrupted, although for large k and fixed p it's very unlikely that significantly more or less bits are corrupted than expected).

Question 2.2. *Alice and Bob are communicating over a noisy channel where a fraction p of the bits sent might be altered. What is the smallest number of bits that Alice can send so that Bob can unambiguously recover the a k-bit message?*

This question was considered by Claude Shannon in his ground breaking paper "A mathematical theory of communication", which introduced many fundamental topics in computer science, in a surprisingly developed form, although the exact form of the question is due to Hamming (whose name you might recognise from the Hamming distance).

To solve this problem Shannon suggested creating a dictionary/code $\mathcal{C} \subseteq \{0,1\}^n$ of size $|\mathcal{C}| = 2^k$ and usng a bijective mapping (an *encoding*) $\phi : \{0,1\}^k \to \mathcal{C}$. To send a message $x \in \{0,1\}^k$, Alice transmits the n-bit encoded message $\phi(x) \in \mathcal{C}$.

Bob will then receive some string $y \in \{0,1\}^n$, which might differ from $\phi(x)$ in a fraction p of the bits. Bob's *decoding* function is simple: Look for the 'closest ' codeword $z \in \mathcal{C}$ to y, where we use the *Hamming metric*, the number of bits which differe, to compute this distance and deduce that the message was $\phi^{-1}(z)$.

When can this go wrong? Well, if two of the codewords are too close together, then Bob can mistake one message for another. More precisely, we need that the Hamming distance between all codewords is at least $2pn$.

So, we've reduced the problem of communicating over this noisy channel to a very concrete combinatorical problem - Given k and p we need to choose n so that we can find a set of 2^k points in $\{0,1\}^n$ which are all at least pn away from each other.

Given a dictionary $\mathcal{C}$ its *rate* is given by $R = \frac{\log |C|}{n}$ and its *distance* by

$$\delta = \frac{\min_{c_1 \neq c_2 \in \mathcal{C}} d_H(c_1, c_2)}{n}.$$

So, the rate measures it's efficient in terms of utilising the channel (not sending too many bits to transmit a short message) and the distance measures it's ability to overcome the noise of the channel.

Question 2.3. *Is it possible to design arbitrarily large dictionaries $\mathcal{C}$ whose rate and distance are at least R_0 and δ_0 for come absolute constants R_0 and δ_0?*

Moreover, can we make these codes explicit, and can we efficiently encode and decode them?

This problem and its relatives (optimising the code's parameters and the efficiency of the algorithms to encode and decode) in this and other models of noise form the basis of Coding theory. It took over 20 years until even the basic question above was resolved, but we are able to give a simple solution to this problem using expander graphs.

2.3 Deterministic error amplification

Randomised algorithms were first considered in the context of primality testing. Rabin, and Solovay and Strassen gave algorithms which when inputted with a k bit integer x and a string r of k random bits, and efficiently computes a function $f(x,r)$ which take the values 0 and 1. If $f = 1$ then x is a prime with probability at least $\frac{1}{2}$, and if $f = 0$ then x is composite.

This bound of $\frac{1}{2}$ might seem unsatisfactory, and we might want to improve it. One easy way to improve it would just be to repeat the algorithm many times, which many different strings

r. Repeating it t times would reduce the probability of a false positive to 2^{-t}. On the other hand the running time, and the number of random bits that we use will also increase by a factor of t. It is intriguing to consider whether we can *deterministically* reduce this error, without having to introduce any more random bits (whilst it might seem easy to produce random bits, actually producing genuinely random strings in the quantity required for modern computing is a non-trivial problem, and the *amount of randomness* needed by a random algorithm should be considered a resource in much that same way memory space and computation time is).

These primality testing algorithms belong to a class of algorithms known as **RP**, the randomised polynomial time algorithms. Let $\{0,1\}^*$ denote the set of all finite binary strings. Then we say a language $\mathcal{L} \subseteq \{0,1\}^*$ is in the class **RP** if there exists a randomised algorithm A with a polynomial (in $|x|$) running time such that if $x \in \mathcal{L}$, then $A(x,r) = 1$, whereas if $x \notin \mathcal{L}$ then the probability that $A(x,r) = 1$ is at most $\frac{1}{16}$ (where $\frac{1}{16}$ is chosen for notational convenience, clearly it doesn't matter what constant we choose here), where r is a uniformly chosen random string of k bits, where k is polynomial in $|x|$. In this case we say that $\mathcal{L}$ has a *(1-sided error) randomised polynomial time membership algorithm.*

Question 2.4. *Assume that $\mathcal{L}$ has a (1-sided error) randomised polynomial time membership algorithm. How many random bits are needed in order to reduce the probability of error to be $\leq \epsilon$?*

Perhaps surprisingly we will see that using expander graphs we can achieve any desired level of accuracy *without using* **any** *additional random bits*!

2.4 Magical graphs

In the previous section we presented three seemingly unrelated problems. We will now introduce a new object, which we call a *magical graph* for now, that will allow us to solve all these problems. These graphs exhibit an *expansion* property, a combinatorial isoperimetric inequality.

Let $G = (L, R, E)$ be a bipartite graph, where L and R are the left and right vertex sets. We say that G is an $(n, m; d)$*-magical graph* if $|L| = n$ and $|R| = m$, every vertex in L has exactly d neighbours in R (G is left regular), and

(1) $|\Gamma(S)| \geq \frac{5d}{8}|S|$ for all $S \subseteq L$ with $|S| \leq \frac{n}{10d}$;

(2) $|\Gamma(S) \geq |S|$ for every $S \subseteq L$ with $|S| \leq \frac{n}{2}$.

Note that, since the graph is d-left-regular, $|\Gamma(S)| \leq d|S|$, and so ((1)) says that most sets have almost the maximum amount of neighbours (up to a constant multiple) that they can have. We obviously can't ask for this to be true for sets of size much larger than $\frac{n}{d}$, and so ((2)) is just asking that these larger size sets at least don't have too bad 'expansion'.

It was shown by Pinsker that such graphs (or at least, graphs with similar 'expansion' properties) do in fact exist. The proof is a novel 'probabilistic' proof of existence.

If people haven't seen the probabilistic method before, the general idea is that in order to show that an object with certain properties, we choose an object from the set of all objects according

to some probability distribution, and bound from below the probability that the random object satisfies these properties. If this probability is non-zero then we can deduce that at least one object exists with the desired properties! In fact, as we will see, the proof shows that most graphs will be magical.

Theorem 2.5. *There exists a constant n_0 such that for every $d \geq 32$ and $n \geq n_0$, $m \geq \frac{3n}{4}$ there exists an $(n, m; d)$-magical graph.*

Proof. We take a set L of size n and a set R of size m and we choose a random bipartite (multi-)graph by choosing, for each vertex $v \in L$ a set of d neighbours, each chosen independently and uniformly at random from R. We make this choice independently for each vertex $v \in L$. Since the size of L and R are as desired, and the graph will be left d-regular, it remains to show that properties ((1)) and ((2)) hold.

Let us deal with ((2)) first, as it is slightly simpler. $S \subseteq L$ have cardinality $\frac{n}{10d} < s := |S| \leq \frac{n}{2}$ and $T \subseteq R$ with $t := |T| = s$. We are interested in the 'bad' events where $\Gamma(S) \subseteq T$, if such an event occurs, then our graph is not magical. In other words, if we let $X_{S,T}$ be the indicator random variable of the event that $\Gamma(S) \subseteq T$ then we are interested in the random variable $X := \sum_{S,T} X_{S,T}$, if $X > 0$ then the graph is not magical, and if $X = 0$ then the graph satisfies ((2)).

For any S and T it is easy to verify that $\mathbb{E}(X_{S,T}) = \mathbb{P}(\Gamma(S) \subseteq T) = \left(\frac{t}{m}\right)^{sd}$. Indeed, for each $v \in S$ the probability one of its neighbours lies in T is $\frac{t}{m}$ and so the probability that $N(v) \subseteq T$ is $\left(\frac{t}{m}\right)^d$, since these events are independent for different neighbours, and so the probability that $\Gamma(S) \subseteq T$ is $\left(\frac{t}{m}\right)^{sd}$, since these events are independent for different $v \in S$.

Hence, by the union bound, and using the standard estimate that $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$, we see that

$$\begin{aligned}
\mathbb{P}(X > 0) &\leq \sum_{S,T} \mathbb{P}(X_{S,T} = 1) = \sum_{S,T} \left(\frac{t}{m}\right)^{sd} \\
&\leq \sum_{s=\frac{n}{10d}}^{\frac{n}{2}} \binom{n}{s}\binom{m}{s}\left(\frac{s}{m}\right)^{sd} \\
&\leq \sum_{s=1}^{\frac{n}{10d}} \left(\frac{en}{s}\right)^s \left(\frac{em}{s}\right)^s \left(\frac{s}{m}\right)^{sd} \\
&= \sum_{s=1}^{\frac{n}{10d}} \left(\frac{e^2 n s^{d-2}}{m^{d-1}}\right)^s \\
&< \frac{1}{10},
\end{aligned}$$

where the last inequality can be shown by showing that in this range the sth term is at most 20^{-s}.

Property ((1)) can be shown via similar method. Let $S \subseteq L$ have cardinality $s := |S| \leq \frac{n}{10d}$, let $T \subseteq R$ be a set of size $t := |T| = \frac{5ds}{8}$ and let $Y_{S,T}$ be the indicator random variable of the event that $\Gamma(S) \subseteq T$. Then as before, if $Y = \sum_{S,T} Y_{S,T}$ we have that G satisfy the fourth property if and only if $Y = 0$. As in the previous case, for any fixed S and T, $\mathbb{P}(\Gamma(S) \subseteq T) = \left(\frac{t}{m}\right)^{sd}$ and so

we can calculate

$$\begin{aligned}\mathbb{P}(Y>0) &\leq \sum_{S,T}\mathbb{P}(Y_{S,T}=1) = \sum_{S,T}\left(\frac{t}{m}\right)^{sd}\\ &\leq \sum_{s=1}^{\frac{n}{10d}}\binom{n}{s}\binom{m}{\frac{5ds}{8}}\left(\frac{5ds}{8m}\right)^{sd}\\ &\leq \sum_{s=1}^{\frac{n}{10d}}\left(\frac{en}{s}\right)^{s}\left(\frac{8em}{5ds}\right)^{\frac{5ds}{8}}\left(\frac{5ds}{8m}\right)^{sd}\\ &< \frac{1}{10},\end{aligned}$$

where again the last inequality can be shown by (carefully) showing that in this range the sth term is at most 20^{-s}.

In particular

$$\mathbb{P}(G \text{ is magic}) = \mathbb{P}(X=0 \text{ and } Y=0) \geq 1-(\mathbb{P}(X>0)+\mathbb{P}(Y>0)) \geq \frac{9}{10}.$$

□

We note that a downside to the proof of Theorem 2.5 is that the proof of existence is non-constructive. To resolve the problems fully we would need explicit constructions of such graphs. The issue of finding such constructions is an important aspect of the field, and we will return to it later, but for now let us present how we can solve our problems using the existence of these magic graphs as a black box.

2.5 A super concentrator with $O(n)$ edges

We will see how we can use these magical graphs to construct superconcentrators. These graphs exhibit incredibly high connectivity, despite only having $O(n)$ edges. The search for super concentrators with n inputs and Kn edges with K as small as possible is still ongoing, and has motivated some important advances in the area. The current best bound is from Alon and Capalbo of $K = 44$.

A *matching* in a graph G is a set of pairwise vertex-disjoint edges. Given two subsets $X, Y \subseteq V(G)$ we say a matching M is *from X to Y* if every edge in M has one endpoint in X and the other in Y, and each $x \in X$ is contained in some edge in M.

We will use the following well-known result

Theorem 2.6 (Hall's theorem)**.** *Let $G = (X, Y, E)$ be a bipartite graph. Then G contains a matching from X to Y if and only if $|\Gamma(A)| \geq |A|$ for all $A \subseteq G$*

If G is an $(n, m; d)$-magical graph then $|\Gamma(S)| \geq |S|$ for every $S \subseteq L$ with $|S| \leq \frac{n}{2}$. Hence, in particular, if we look at the subgraph of G between S and R then, since every $A \subseteq S$ has size

at most $|S| \leq \frac{n}{2}$, we can conclude that $|\Gamma(A) \geq |A|$. Hence this graph satisfies Hall's condition, and so there is a matching from S to $\Gamma(S)$.

We will use this fact to recursively construct a super concentrator C' with n vertices on each side. For $n \leq n_0$ suitably small we can simply take $K_{n,n}$ which is clearly a superconcentrator and has at most $\frac{n_0}{2}n$ edges.

For $n \geq n_0$ we build our superconcentrator C' with n inputs and outputs using three building blocks:

(i) Two copies $G_1 = (L_1, R_1, E_1)$ and $G_2 = (L_2, R_2, E_2)$ of an $(n, \frac{3n}{4}, d)$-magical graph;

(ii) A superconcentrator C connecting the input set R_1 to the output set R_2. Note that, since these sets have size $\frac{3n}{4} < n$, such a superconcentrator exists by assumption;

(iii) A perfect matching M between L_1 and L_2.

We consider the graph H with $V(H) = L_1 \cup R_1 \cup L_2 \cup R_2$ whose edge set is $E(G_1) \cup E(G_2) \cup E(C) \cup M$. We claim that H is a super concentrator with input set L_1 and output set L_2, and also that it only has linearly many (in n) edges.

So, let $S \subseteq L_1$ be a set of input vertices and $T \subseteq L_2$ be a set of output vertices such that $|S| = |T| = k$. We wish to show that there are k vertex-disjoint paths between S and T in H.

If $k > \frac{n}{2}$, then many vertices in S must be matched to vertices of T in the matching M, explicitly at least $k - \frac{n}{2}$ vertices. Deleting the matched vertices from S and T leaves two sets of size at most $\frac{n}{2}$.

Hence, it will be sufficient to show that we can find such paths when $k \leq \frac{n}{2}$ without using the edges of M. In this case, since G_1 and G_2 are magical, there are matchings M_1 and M_2 in G_1 and G_2 from S to $\Gamma(S)$ and T to $\Gamma(T)$, respectively. Let S' and T' be the endpoints of these matchings in $\Gamma(S) \subseteq R_1$ and $\Gamma(T) \subseteq R_2$.

Since C is a super concentrator, and $|S'| = |S| = |T| = |T'|$ the set of input vertices $S' \subseteq R_1$ can be connected to the set of output vertices $T' \subseteq R_2$ via a family of k vertex-disjoint paths in C. Extending these paths by the matchings M_1 and M_2 we obtain a family of k vertex-disjoint paths from S to T in C'. It follows that C' is a super concentrator.

It remains to estimate the number of edges $e(n)$ used in this construction. We obtain the following recursion

$$e(n) \leq \begin{cases} 2nd + n + e\left(\frac{3n}{4}\right) & \text{for } n \geq n_0 \\ n^2 & \text{for } n \leq n_0 \end{cases}.$$

Solving this recursion gives $e(n) \leq Kn$ where K only depends on d and n_0, and so we have a super concentrator with $O(n)$ edges as desired.

A short word about computational aspects: If we have an algorithm to construct these magical graphs which takes time $t(n)$ then the above recursive construction will give an algorithm which constructs a superconcentrator with input/output size n in time $O(t(n))$.

We also note that this is just one of a host of applications of expanders to network construction problems. Another important one comes in the work of Ajtai, Komlos and Szemeredi, who build what is called the AKS sorting network, which sorts n items using a very small number $O(n \log n)$ of comparisons (and in particular each item is only involved in a small number $O(\log n)$ of comparisons).

2.6 Construction of good error correcting codes

We will need the following useful property of magical graphs.

Claim 2.7. Suppose G is a $(n, \frac{3n}{4}, d)$-magical graph and $S \subseteq L$ is such that $s = |S| \leq \frac{n}{10d}$. Then there exists some vertex $u \in R$ with a unique neighbour in S, i.e., such that $|N(u) \cap S| = 1$.

Proof. Consider the number of edges between S and $\Gamma(S)$, which we write as $e(S, \Gamma(S))$. On the one hand, by left regularity, clearly $e(S, \Gamma(S)) = ds$. However, on the other hand, since by (1) $|\Gamma(S)| \geq \frac{5d}{8}$, the average number of neighbours that a vertex in $|\Gamma(S)|$ has in S is at most $\frac{8}{5} < 2$. But, every vertex in $\Gamma(S)$ has at least one neighbour in S, and hence there must be at least one vertex in $\Gamma(S)$ with exactly one neighbour in S (in fact, there has to be many). □

We will use the magical graph G to construct a code $\mathcal{C} \subseteq \{0,1\}^n$ with rate at least $\frac{1}{4}$ and distance at least $\frac{1}{10d}$. To this end, let us think of the graph G in terms of an $n \times m$ matrix A with rows indexed by L and columns by R where $(A)_{ij} = 1$ if and only if the vertex i in L is adjacent to the vertex j in R. Our code is given by the right kernel of A over $\mathbb{F}_2$

That is, $\mathcal{C} = \{x \in \{0,1\}^n \colon Ax = 0\}$. Clearly $\mathcal{C}$ is a linear subspace of $\{0,1\}^n$ and, since $m \leq \frac{3n}{4}$, it's dimension is at least $\frac{n}{4}$. In particular, $|\mathcal{C}| \geq 2^{\frac{n}{4}}$. Hence the *rate* $R = \frac{\log |\mathcal{C}|}{n} \geq \frac{1}{4}$ as claimed.

To prove a lower bound on the distance of the code, we note that, by the linearity of $\mathcal{C}$, the distance will be equal to the smallest size of a non-zero codeword, in other words, the smallest support of a non-zero $x \in \mathcal{C}$. Let $x \neq 0$ have support $S = \{i \in L \colon x_i = 1\}$. If $|S| < \frac{n}{10d}$ then, as we saw, there must be some $j \in R$ such that $|N(j) \cap S| = 1$. But then $(Ax)_j = 1$, and so $x \notin \mathcal{C}$! It follows that for every $x \in \mathcal{C}$, $|x| \geq \frac{n}{10d}$ and hence the distance δ of the code is at least $\frac{1}{10d}$.

This is a special example of a so-called LDPC (Low density parity check) code. The idea for such a code was first suggested by Gallagher, in fact predating and motivating the work of Bassalygo, Pinsker and Margulis who first explicitly defined and cosntructed expander graphs. Whilst falling out of favour for a while, these codes are now believed to have simultaneously the best coding paramaters and algorithmic performance in various settings.

However we note that is is only in the last 20 years that the art of explicitly constructing expanders has reached the point where the implementation of this simple argument is feasible. We also note that this algorithm can also be shown to have very efficient (linear time) decoding, and we will revisit them later in the course.

As in the previous application, the time complexity of constructing this magical graph is dominating the time to construct the code here, driving home the point that as much as the

existence of these graphs and codes, it is efficient algorithms to generate them which are interesting. The next application calls for an even more concise and efficient description of these graphs.

2.7 Deterministic error amplification

Recall that we have a language $\mathcal{L} \subseteq \{0,1\}^*$ which is in RP, so that $\mathcal{L}$ has a randomised polynomial time membership algorithm with a 1-sided error, some function f which, given a string x and a random string r, calculates a function $f(x,r)$ such that $f(x,r) = 1$ whenever $x \in \mathcal{L}$ but $f(x,r) = 1$ when $x \notin \mathcal{L}$ with probability at most $\frac{1}{16}$.

To reduce the probability of error we will carry out *dependent sampling* of random strings using these magical graphs. This comes from an idea of Karp, Pippenger and Sipser. Our goal is to reduce the failure probability below some fixed threshold ϵ whilst using as few random bits as possible. Let us fix $x \notin \mathcal{L}$ and consider the set of 'bad' strings $B = \{r \in \{0,1\}^k \colon f(x,r) = 1\}$, those which fail on input together with x. We would like to make it very likely that one of the strings r that we consider is not in B, however the only information we have about $B \subseteq \{0,1\}^k$ is that it is not too big, $|B| \leq \frac{n}{16}$ where $n = 2^k$.

For a given integer d we will give an algorithm for the membership problem which evaluates f only d times and fails with probability $\epsilon \leq \frac{1}{10d}$. Note, if we just choose a random string r each time, then the failure probability is much smaller, $\left(\frac{1}{16}\right)^d$. The advantage here will be that we need to choose far fewer random strings.

Let G be an $(n,n;d)$-magical graph, where $n = 2^k$ as before, and identify L and R with $\{0,1\}^k$ arbitrarily. To decide whether a given string $x \in \mathcal{L}$ we first randomly sample a k-bit string $r \in L$ and let $r_1, \ldots, r_d \in R$ be its neighbours (note that whilst these are random, since G is fixed, they are determined by the choice of r). We evaluate $f(x,r_i)$ for each i, and the algorithm outputs 1 (so claims $x \in \mathcal{L}$)) if $f(x,r_i) = 1$ for all i, otherwise we say $x \notin \mathcal{L}$.

We want to bound then the probability of failure for an arbitrary input x. Clearly the only way this algorithm can be wrong is if $x \notin \mathcal{L}$ but $f(x,r_i) = 1$ for all i, i.e., $r_i \in B$ for all i. In other words, $N(r) \subseteq B$. Let $S \subseteq L$ be the set of strings r such that $N(r) \subseteq B$, so that our algorithm fails if and only if we chose r from S. It remains to estimate the probability that $r \in S$ (or, in other words, the size of S).

However, if $|S| > \frac{n}{10d}$ then let $S' \subseteq S$ be of size exactly $\frac{n}{10d}$. In this case by (1) we have that $|\Gamma(S')| \geq \frac{5d}{8}|S'| \geq \frac{n}{16}$ by our expansion property, however since $\Gamma(S') \subseteq \Gamma(S) \subseteq B$,

$$|B| \geq |\Gamma(S')| \geq \frac{n}{16}$$

or, in other words, the probability that $f(x,r) = 1$ is at least $\frac{1}{16}$, a contradiction! It follows that $|S| \leq \frac{n}{10d}$ and so, the moment of magic, the probability that $r \in S$, which is the probability our algorithm fails, is at most $\frac{1}{10d}$.

By choosing d appropriately large, we can reduce the error as small as we like. The key point here being that we only ever needed to choose the initial k-bit random string r!

Unlike in the previous examples the size of the magical graph n is exponential in the size of the problem considered (this parameter k). This means that to efficiently implement this algorithm, our encoding of the magical graph will have to be much more efficient than in the previous applications, we can't hope to know the entire graph! Specifically, to ensure that our computations happen in polynomial time, we will have to have an oracle which, given a k-bit string r can generate the d neighbours of r in polynomial time. We will later see that even this level of explicitness is achievable!

Also, we note that the reduction in error is rather inefficient in terms of the number of strings we evaluate. By using random strings each time we achieve an exponential decay in the failure probability. We will later see that a more refined dependent sampling procedure using expanders, using still many fewer random bits, can achieve such an exponential decay in terms of the number of evaluations.

Finally, we note that we were only dealing here with 1-sided errors. In practise, most stochastic algorithms can fail on both inputs inside and outside of $\mathcal{L}$, and the above strategy does not work as written. However, we will later see that a small modification achieves the same level of error reduction in this context as well.

3 Graph expansion and eigenvalues

3.1 Edge expansion and a combinatorial definition of expansion

Unless we say explicitly otherwise, all graphs we consider with be regular, with regularity d, and undirected. We will write $|G| = n$ for the number of vertices and $||G|| = e$ for the number of edges in a graph. Graphs are not necessarily simple, they may have multiple edges and loops. We write $E(S,T)$ for the set of edges between S and T and $e(S,T)$ for the number of edges with an endpoint in S and the other in T (formally, when S and T intersect we will count the edges with both endpoints in $S \cap T$ twice in this quantity).

An important concept is the the *edge boundary* of S, $\partial(S) = E(S, S^c)$, which gives rise to the *expansion ratio* of G, denote by $h(G)$, which is

$$h(G) = \min_{S:\, |S| \leq \frac{n}{2}} \frac{|\partial S|}{|S|}.$$

so, a graph has a large expansion ratio (or in other words is a good expander) if every "small" set of vertices has a large number of edges in it's boundary.

The are a few important ways we can extend this definition. The first is to consider other notions of boundary, such as the vertex boundary, giving rise to vertex expansion which we discuss later. We can also consider expansion ratios as a function of the sets considered, either changing in some continuous fashion, or as a 'cut-off' requiring expansion for only certain types of sets.

Definition. A sequence of d-regular graph $G_1, G_2, \ldots$ where d is fixed and $n_i = |G_i|$ is increasing with i is a *family of expander graphs* if there exists some constant $\alpha > 0$ such that $h(G_i) \geq \alpha$ for all i.

We will concerned with the construction of families of expander graphs, and in particular the *explicit* construction of such graphs. There are two particular notions of 'explicitness' that we will keep in mind in this regard. In the first we require that the n-vertex graph can be generated 'from scratch' in time polynomial in n. In the stronger version we instead require that the neighbourhood of any given vertex be computable is time that is polynomial in the description length of the vertex (which is usually polynomial in $\log n$).

These definitions might seem at first a little odd, but they arise very naturally when considering the algorithmic implications of expander graphs, where the performance of algorithms which use expanders will depend on efficiently obtaining the relevant information about the graph structure of the expanders that are used.

Definition. Let (G_i) be a family of expander graphs such that the n_i are not increasing too quickly (for example $n_{i+1} \leq n_i^2$).

(1) The family is called *mildly explicit* if there is an algorithm that generates the jth graph in the family G_j is time polynomial in j.

(2) The family is called *very explicit* if there is an algorithm that on input of an integer i, a vertex $v \in V(G_i)$ and $k \in [d]$ computes the kth neighbour of v in G_i, where this algorithm's

run-time should be polynomial in its input length (the number of bits needed to express the triple i, v, k).

To demonstrate these definitions let us give briefly a description of two families of expander graphs.

The first is a family of 8-regular graphs G_i, one for each integer i. The vertex set is the discrete torus $\mathbb{Z}_m \times \mathbb{Z}_m$ and the neighbours of a vertex (x, y) are given by their image under certain linear transformations (and their affine shifts). Explicitly we take the following matrices

$$T_1 = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \text{ and } T_2 = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}, \text{ and let } e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and } e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Then the neighbours of a vertex $v = (x, y)$ are given by $T_1 v, T_2 v, T_1 v + e_1$ and $T_2 v + e_2$, as well as the four neighbours of v obtained by the four inverse transformations, where all operations are taken mod i.

This family of graphs, due to Margulis, is the first explicitly constructed family of expander graphs. His original proof that they are expanders was based on representation theory (note the algebraic definition of the graph) and did not give an explicit bound on the expansion ratio h, although such a bound was later derived by Gabber and Galil using tools from harmonic analysis. We will discuss these graphs in more detail later, and give a proof of their expansion using discrete fourier analysis. Note that this family is very explicit - if you tell me the graph G_i and a vertex (x, y), I can tell you very easily the neighbours of (x, y) in G_i (in fact in constant time).

A second family is a family of 3-regular expanders G_i one for each prime number p. Here the vertex set is $\mathbb{Z}_p$ and a vertex x is connected to $x + 1, x - 1$ and x^{-1}, where again calculations are mod p.

Here the proof of expansion relies on some very deep number theoretical results, specifically the Selberg 3/16 theorem, and we won't discuss this is detail. However this family is only mildly explicit, since in order to determine the structure of G_i we first have to actually determine p_i the ith prime.

3.2 Graph spectrum and an algebraic definition of expansion

As you might have guessed from the structure of the examples given, there are deep links between graph expansion and algebraic structures. We will investigate these links further, giving an algebraic definition for expansion.

The *adjacency matrix* of an n-vertex graph G, which we denote by $A = A(G)$, is an $n \times n$ matrix whose rows and columns are indexed by $V(G)$ and whose uvth entry is the number of edges in G between u and v.

Since A, by definition, is real and symmetric, it has n real eigenvalues whch we denote by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, which we can associate with an orthonormal system of eigenvectors $v_1, v_2, \ldots, v_n$ where $Av_i = \lambda_i v_i$ for each i. We often refer to the eigenvalues of A as the *spectrum*

of the graph (here we can think about the graph as this matrix A, which is an operator acting on vectors in $\mathbb{R}^V$, or weightings on the vertices of V. As we will see, the action that this matrix has can tell us a lot about the structure of the graph G).

For example the following simple facts, which we'll prove on the exercise sheet

- $\lambda_1 = d$ and the corresponding eigenvector is $v_1 = \frac{1}{\sqrt{n}}\mathbf{1} = (\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}) := \boldsymbol{u}$;
- G is connected iff $\lambda_1 > \lambda_2$;
- G is bipartite iff $\lambda_1 = -\lambda_n$.

More importantly for us, the graphs second eigenvalue is closely related to its expansion parameter.

Theorem 3.1. *Let G be a d-regular graph with spectrum $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Then*

$$\frac{d-\lambda_2}{2} \leq h(G) \leq \sqrt{2d(d-\lambda_2)}$$

This can be thought of as a discrete version of a famous theorem of Cheeger in Riemannian geometry (we'll talk more about this connection later) and was proved by Dodziuk, and Alon and Milman. We'll give a proof of this theorem later in course, but for now let us just note that qualitatively it gives use a two-sided connection between the quantity $d - \lambda_2$, or perhaps more instructively $\lambda_1 - \lambda_2$, which we call the *spectral gap* and the quantity $h(G)$. A lower or upper bound on either will give us a bound on the other, and in particular $h(G)$ is bounded away from zero if and only if $d - \lambda_2$ is (thinking of this in the context of a family of expander graphs).

3.3 The expander mixing lemma

Stronger information can be derived from considering a related quantity, which we denote by $\lambda = \lambda(G) = \max\{|\lambda_2|, |\lambda_n|\}$, which can be seen to the be the largest absolute value of an eigenvalue outside of $|\lambda_1| = d$. If λ_2 is small then the spectral gap is large, but here we insist that *all* other eigenvalues are small in absolute value.

The following really useful lemma the relates this quantity λ to the distibution of edges in the graph G.

Lemma 3.2 (Expander mixing lemma)**.** *Let G be a d-regular graph with n vertices and let $\lambda = \lambda(G)$. Then for all $S, T \subseteq V$*

$$\left| e(S,T) - \frac{d|S||T|}{n} \right| \leq \lambda\sqrt{|S||T|}.$$

Before we prove this lemma, let us say a few words of interpretation. The left hand side measures the difference between $e(S,T)$ and the 'expected' number of edges between S and T in a random graph with the same density (each edge is present with probability $\frac{d}{n}$ and there are $|S||T|$ many potential edges between S and T. Hence, when λ is very small, the deviation (or

discrepancy) between these two quantities is small, and so the distribution of the edges in this graph is very uniform, or close to random.

In some sense, many properties of random graphs can be shown to hold deterministically in graphs where λ is small, and for this reason (and others) they are often known as *pseudorandom* graphs. There are other, equivalent, more combinatorial definitions of pseudorandomness that seem to imply that these graphs are a very natural object of study.

We also note that, whilst the bounds in Theorem 3.1 are quite far apart when the spectral gap is much smaller than d (and so the expansion ratio is not perhaps very well tied to the spectral gap), the relationship in the expander mixing lemma is much closer to being tight, demonstrating a much closer relationship between λ and the distribution of edges in G, as demonstrated by this very recent partial converse of the expander mixing lemma given by Bilu and Linial.

Lemma 3.3. *Let G be a d-regular graph and suppose that there is some positive ρ such that for all disjoint $S, T \subseteq V$*

$$\left| e(S,T) - \frac{d|S||T|}{n} \right| \leq \rho\sqrt{|S||T|}.$$

Then $\lambda \leq O\left(\rho\left(1+\log\left(\frac{d}{\rho}\right)\right)\right)$, and this bound is tight.

Let us give then a proof of the expander mixing lemma.

Proof of Lemma 3.2. Let $\mathbf{1}_S$ and $\mathbf{1}_T$ be the characteristic vectors of S and T in $\mathbb{R}^V$. That is, $(\mathbf{1}_S)_v = 1$ if $v \in S$ and 0 otherwise. Let $\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_n}$ be an orthonormal basis of eigenvectors for the adjacency matrix and let us express $\mathbf{1}_S = \sum_i \alpha_i \boldsymbol{v_i}$ and $\mathbf{1}_T = \sum_j \beta_j \boldsymbol{v_j}$.

We can count the number of edges from S to T in the following way

$$e(S,T) = (\mathbf{1}_S)^T A \mathbf{1}_T = (\sum_i \alpha_i \boldsymbol{v_i})^T A (\sum_j \beta_j \boldsymbol{v_j}).$$

However, since the $\boldsymbol{v_i}$ are orthonormal eigenvectors of A, it follows that

$$e(S,T) = \sum_i \lambda_i \alpha_i \beta_i.$$

Since $\alpha_1 = \langle \mathbf{1}_S, \boldsymbol{v_1} \rangle = \langle \mathbf{1}_S, \boldsymbol{u} \rangle = \frac{|S|}{\sqrt{n}}$, and similarly $\beta_1 = \frac{|T|}{\sqrt{n}}$, and $\lambda_1 = d$

$$e(S,T) = \frac{d|S||T|}{n} + \sum_{i \geq 2} \lambda_i \alpha_i \beta_i.$$

Hence,

$$\left| e(S,T) - \frac{d|S||T|}{n} \right| \leq \left| \sum_{i \geq 2} \lambda_i \alpha_i \beta_i \right| \leq \sum_{i \geq 2} |\lambda_i \alpha_i \beta_i| \leq \lambda \sum_{i \geq 2} |\alpha_i \beta_i|.$$

Finally then, by Cauchy-Schwartz, we can bound

$$\begin{aligned}\left|e(S,T) - \frac{d|S||T|}{n}\right| &\leq \lambda \sum_{i\geq 2} |\alpha_i \beta_i| \\ &\leq \lambda \sqrt{\sum_{i\geq 2} |\alpha_i|^2 \sum_{j\geq 2} |\beta_j|^2} \\ &= \lambda ||\alpha||_2 ||\beta||_2 \\ &= \lambda ||1_S||_2 ||1_T||_2 = \lambda \sqrt{|S||T|}.\end{aligned}$$

□

In what follows it will sometimes be convenient to consider the *normalised second eigenvalue* $\frac{\lambda(G)}{d}$. We will refer to d-regular graphs with n vertices as (n,d)-*graphs*, and if they further have a normalised second eigenvalue of at most α, i.e., if $\lambda(G) \leq \alpha d$, then we refer to them as (n,d,α)-*graphs*.

3.4 How large can the spectral gap be?

So, we will think of graphs with a large spectral graph, or equivalently a small λ_2, as having good expansion properties. It is natural to ask how large a spectral gap we can achieve in general. Clearly this questions depends on the relationship between d and n. We are mostly interested in the case where d is fixed, but n is growing. In this range the question is essentially answered by this theorem of Alon and Boppana.

Theorem 3.4. *Let d be fixed, then for every (n,d)-graph G*

$$\lambda_2(G) > 2\sqrt{d-1} - o_n(1).$$

To illustrate how things can differ when d can grow as a function of n, consider the complete graph on n vertices, which we denote by K_n, which has $d = n-1$. Clearly the adjacency matrix $A(K_n) = J - I$ where J is the all-ones matrix and I is the identity.

It is an exercise in elementary linear algebra to show that the spectrum of K_n is give by $\lambda_1 = n-1$ and $\lambda_i = -1$ for all $i \geq 2$, and hence $\lambda_2(K_n) = 1$ is much smaller than $\sqrt{d} \approx \sqrt{n}$.

We will discuss the bound more and give a proof later in the course, but for now let us give a simple argument for a slightly weaker statement.

Lemma 3.5. *For every (n,d)-graph G*

$$\lambda(G) \geq \sqrt{d}(1 - o_n(1)).$$

Proof. Let A be the adjacency matrix of G. It is relatively easy to show that the trace of A^k is the number of closed walks of length k in G. In particular, $\mathrm{tr}(A^2) \geq dn$. Indeed, for each vertex and each edge incident to that vertex we can form a closed walk of length 2 by moving back and forth along that edge.

On the other hand

$$\mathrm{tr}(A^2) = \sum_i \lambda_i^2 \leq d^2 + (n-1)\lambda^2.$$

Rearranging gives $\lambda^2 \geq d\frac{n-d}{n-1} = d(1 - o_n(1))$. □

3.5 Four perspectives on expansion

We've now developed enough of the background to give a bit of a broader view of the main questions and topics that we're going to cover in the course.

As we've seen expansion can be defined in a purely combinatorial manner, but is closely related to the spectral theory of graphs. As we shall see shortly, another equivalent probabilistic way to think about expansion will come from rapidly mixing random walks.

In each of these frameworks we'll consider mostly four types of questions

- **Extremal:** How large/small can the relevant expansion parameters be?
- **Typical:** How are these parameters distributed in a typical/random graph?
- **Explicit Construction:** Can one construct graphs for which these paramaters (nearly) obtain their optimum?
- **Algorithmic:** Given a graph, can you efficiently evaluate/estimate its expansion parameters?

It also will be natural to consider some **comparative** questions: What can we conclude about, say, combinatorial expansion paramaters from spectral information etc.

Let us start in the next section by introducing this probabilistic perspective to expansion.

4 Random walks on expanders graphs

If we take a random walk on a graph (choosing uniformly a random neighbour in each step) then our position at time t is some probability distribution on the vertices of the graph (depending on the starting vertex). It's not too hard to show that, independent of the starting vertex, for a long enough walk this distribution approaches a limiting distribution, which will in fact be uniform for an (n, d)-graph.

In other words, one way to approximate a random sample of a vertex in a graph is to take a long enough random walk. A key property of expander graphs is that the random walk converges very quickly to this limiting distributing which we call the stationary distribution, and so in order to choose a random sample, it's sufficient to take a random walk of quite a short length.

In many theoretical, and practical, problems one needs to draw samples from some distribution $\mathcal{F}$ on a finite (but very large) set V. One way to do so is to consider a graph with vertex set V so that the stationary distribution on G is $\mathcal{F}$. A clever choice of G can guarantee that it is feasible to efficiently simulate this random walk and furthermore that the distribution given by the walk *rapidly* converges to $\mathcal{L}$. This type of method, in some contexts known as the "Monte-Carlo" method, pops up in various fields such as statistical physical and combinatorial optimisation.

The main idea behind this chapter will be that the set of vertices visited by a length t random walk on an expander will 'look like' (in some sense) a set of t vertices sampled uniformly and independently from the graph. The computational signifance of this is that the number of random bits required to generate a length t walk on a d-regular graph is significantly smaller than the number of random bits required to sample t random vertices when $d \ll n$.

4.1 Rapid mixing of walks

A *walk* on a graph $G = (V, E)$ is a sequence of vertices $v_1, v_2, \ldots$, such that v_{i+1} is a neighbour of v_i for each i. If v_{i+1} is selected uniformly at random from among v_i's neighbours, independently for each i, this is called a *random walk* on G.

It is a well-known fact that for every finite connected non-bipartite graph G the distributions π_i converge to a limit, or *stationary* distribution, and it is easy to show that this distribution depends only on the degree of the vertex (and so in particular is uniform on regular graphs). In this subsection we are interested in *how quickly* this distribution converges to the limit. There are several sensible/interesting ways to measure the distance between distributions, and we will consider a few different norms and entropy measures. The main point is that in an expander the distance to the uniform measure shrinks substantially with *every* step of the random walk, and in fact this condition in some way characterises quantitatively graph expansion.

Let us begin by making a few useful definitions. A vector $\boldsymbol{p} \in \mathbb{R}^V$ is called a *probability distribution vector* if $p_i > 0$ for all i and $\sum_i p_i = 1$. The *uniform distribution* is given by $\boldsymbol{u} = \frac{1}{n}(1, 1, \ldots, 1)$.

Definition. A *random walk* on a finite graph $G = (V, E)$ is a discrete-time stochastic process

$(X_0, X_1, \ldots)$ taking values in V. The vertex X_0 is sampled from some initial distribution $\boldsymbol{\pi^1}$ on V, and X_{i+1} is chosen uniformly at random from the neighbours of X_i. We write $\boldsymbol{\pi^i}$ for the distribution of X_i.

If G is a d-regular graph with adjacency matrix A then its *normalised adjacency matrix* is given by $\hat{A} = \frac{1}{d}A$. Here are some simple comments on this random walk:

- The random walk on G is a *Markov chain* with state set V and transition matrix $\hat{A}$.
  - A Markov chain is a sequence $(X_1, X_2, \ldots,)$ of random variables such that for any n and $x_1, x_2 \ldots, x_n$

    $$\mathbb{P}(X_n = x_n | X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}) = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}),$$

    when both conditional probabilities are well defined, i.e., it is a memoryless stochastic process.
- $\hat{A}$ is real, symmetric and double stochastic (all row and column sums are equal to one).
- If $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_n$ are the eigenvalues of $\hat{A}$ then $\hat{\lambda_i} = \frac{\lambda_i}{d}$ for all i, and the eigenvectors are the same. In particular $\hat{\lambda}_1 = 1$ and if G is an (n, d, α)-graph then $\max\{|\hat{\lambda}_2|, |\hat{\lambda}_n|\} \leq \alpha$.
- If we sample a vertex x from some probability distribution $\boldsymbol{p}$ on V and then move to a random neighbour y of x, the resulting probability distribution on V is given by $\hat{A}\boldsymbol{p}$.
- The matrix $\hat{A}^t$ is the transition matrix of the Markov chain defined by random walks of length t. In other words, $(\hat{A}^t)_{ij}$ is the probability that a random walk starting at i is at j after t steps, and so $\boldsymbol{\pi^i} = \hat{A}^i\boldsymbol{\pi^1}$.
- The stationary distribution of the random walk on G is the uniform distribution, namely, $\boldsymbol{u}\hat{A} = \hat{A}\boldsymbol{u} = \boldsymbol{u}$.

4.1.1 Convergence in the ℓ_p norms

The inner product of two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is denote by

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_i x_i y_i.$$

Given $p \geq 1$ the ℓ_p norm is given by

$$||\boldsymbol{x}||_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}},$$

where it is a simple exercise, using for example Jensen's inequality, to check that this is a norm. These are discrete versions of the L_p norms from functional analysis. We are particularly interested in the following special cases

- $||\boldsymbol{x}||_1 = \sum_i |x_i|$;

- $||\boldsymbol{x}||_2 = \sqrt{\sum_i |x_i|^2}$;
- $||\boldsymbol{x}||_\infty := \max_i\{|x_i|\} = \lim_{p\to\infty} ||\boldsymbol{x}||_p$.

Our first useful observation is that if G is an (n, d, α)-graph and $\alpha < 1$, then regardless of the initial distribution $\boldsymbol{\pi^1}$, the random walk converges in ℓ_1 exponentially fast to its limiting distribution $\boldsymbol{u}$. This will in fact follows from a similar bound in ℓ_2, which follows from the fact that in ℓ_2 the distance to the uniform distribution will shrink by a factor of α at each step.

Lemma 4.1. *Let G be an (n, d, α)-graph with normalised adjacency matrix $\hat{A}$. Then for every probability distribution vector $\boldsymbol{p}$,*

$$||\hat{A}\boldsymbol{p} - \boldsymbol{u}||_2 \leq \alpha||\boldsymbol{p} - \boldsymbol{u}||_2 \leq \alpha.$$

Proof. We first note that the uniform distribution $\boldsymbol{u}$ is invariant under the action of $\hat{A}$. So, in particular, $\boldsymbol{u}$ is an eigenvalue of $\hat{A}$ with eigenvalue one, and all other eigenvalues of $\hat{A}$ have absolute value at most α. Hence it follows that if $\boldsymbol{x}$ is orthogonal to $\boldsymbol{u}$ then

$$||\hat{A}\boldsymbol{x}||_2 \leq \alpha||\boldsymbol{x}||_2.$$

However,

$$\langle \boldsymbol{u}, \boldsymbol{p} - \boldsymbol{u}\rangle = \langle \boldsymbol{u}, \boldsymbol{p}\rangle - \langle \boldsymbol{u}, \boldsymbol{u}\rangle = \frac{\sum_i p_i}{n} - \frac{\sum_i 1}{n^2} = \frac{1}{n} - \frac{1}{n} = 0,$$

and so $\boldsymbol{u}$ is orthogonal to $\boldsymbol{p} - \boldsymbol{u}$. It follows that

$$||\hat{A}\boldsymbol{p} - \boldsymbol{u}||_2 = ||\hat{A}(\boldsymbol{p} - \boldsymbol{u})||_2 \leq \alpha||\boldsymbol{p} - \boldsymbol{u}||_2 \leq \alpha,$$

where the last line follows from the fact that

$$||\boldsymbol{p} - \boldsymbol{u}||_2^2 = \langle \boldsymbol{p} - \boldsymbol{u}, \boldsymbol{p} - \boldsymbol{u}\rangle = \langle \boldsymbol{p} - \boldsymbol{u}, \boldsymbol{p}\rangle = ||\boldsymbol{p}||_2^2 - \langle \boldsymbol{p}, \boldsymbol{u}\rangle = ||\boldsymbol{p}||_2^2 - \frac{1}{n} \leq 1 - \frac{1}{n},$$

since $||\boldsymbol{x}||_2^2 = \sum_i x_i^2 \leq \sum_i x_i = 1$ for all probability distribution vectors. □

A simple consequence of repeated applications of Lemma 4.1 is the following:

Theorem 4.2. *Let G be an (n, d, α)-graph with normalised adjacency matrix $\hat{A}$. Then for any probability distribution vector $\boldsymbol{p}$ and any positive integer t*

$$||\hat{A}^t\boldsymbol{p} - \boldsymbol{u}||_2 \leq \alpha^t||\boldsymbol{p} - \boldsymbol{u}||_2 \leq \alpha^t.$$

Again as a simple consequence we obtain the following ℓ_1 bound.

Theorem 4.3. *Let G be an (n, d, α)-graph with normalised adjacency matrix $\hat{A}$. Then for any probability distribution vector $\boldsymbol{p}$ and any positive integer t*

$$||\hat{A}^t\boldsymbol{p} - \boldsymbol{u}||_1 \leq \sqrt{n}\alpha^t.$$

Proof. A simple application of the Cauchy-Schwarz inequality show that for any vector $\boldsymbol{x} \in \mathbb{R}^n$

$$||\boldsymbol{x}||_1^2 = \left(\sum_i |x_i|\right)^2 = \left(\sum_i 1 \cdot |x_i|\right)^2 \leq \left(\sum_i 1\right)\left(\sum_i |x_i|^{\frac{1}{2}}\right) = n||\boldsymbol{x}||_2^2.$$

Hence, by Theorem 4.2 we have that

$$||\hat{A}^t\boldsymbol{p} - \boldsymbol{u}||_1 \leq \sqrt{n}||\hat{A}^t\boldsymbol{p} - \boldsymbol{u}||_2 \leq \sqrt{n}\alpha^t.$$

□

In particular, once $t \gg \log n$, the bound $\sqrt{n}\alpha^t \ll 1$ and so the random walk will be very close to uniform.

Why are we particularly interested in the ℓ_1 norm? Well, it's reasonably common to measure the distance between two probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ on a set X in terms of their *total variation distance*

$$d_{TV}(\boldsymbol{p}, \boldsymbol{q}) = \max_{B \subseteq X} |\mathbb{P}_{\boldsymbol{p}}(B) - \mathbb{P}_{\boldsymbol{q}}(B)| = \frac{1}{2}||\boldsymbol{p} - \boldsymbol{q}||_1.$$

To put it another way, if the ℓ_1 distance between two distribution is sufficiently small, then the two distributions assign nearly equal probabilities to *every* event in the probability space.

4.1.2 Convergence in entropy

Another important perspective of a random walk is offered by the *entropy* of the associated probability distributions. the concept of entropy is a fundamental concept in theory of communication, capturing the amount of 'information', or alternatively 'uncertainty', that a distribution carries. When we take a random step in our random walk we 'inject' a little bit more randomness into our distribution π_i, indeed precisely the $\log d$ random bits that are necessary to specify which of the d neighbours of our current vertex we travel to.

One should expect that this injection increases the amount of randomness in the distribution. Note that, in some ways the uniform distribution is the 'most random' distribution, whereas as we approach a distribution which is supported on fewer and fewer points we are becoming closer and closer to deterministic. Hence, we should expect that each step increases the entropy of the distributions $\boldsymbol{\pi^i}$, and makes them closer to uniform. This turns out to be true in fact in any regular graph, and expanders are the graphs for which this increase is significant.

The entropy viewpoint will be important as a heuristic later in the course, in particular for the *zig-zag construction* and its use in constructing various types of expanding graphs and computational frameworks.

In the same way that different norms capture different aspects of the probability distributions, there are several variations on the theme of entropy. Given a probability distribution $\boldsymbol{p}$ on $[n]$ we define:

- **Shannon entropy:** $H(\boldsymbol{p}) = -\sum_i p_i \log(p_i)$.
- **Rényi 2-entropy:** $H_2(\boldsymbol{p}) = -2\log(||\boldsymbol{p}||_2)$.
- **Min entropy:** $H_\infty(\boldsymbol{p}) = -\log(||\boldsymbol{p}||_\infty)$.

As with the p-norms, there is in fact a family of entropies H_α for $\alpha > 1$ and the Shannon entropy is the limit as $\alpha \to 1$, whereas the min entropy is the limit and $\alpha \to \infty$. We note in particular that

Proposition 4.4. *For any probability distribution p on $[n]$*

$$H_\infty(\boldsymbol{p}) \leq H_2(\boldsymbol{p}) \leq 2H_\infty(\boldsymbol{p}).$$

Proof. This follows from the fact that

$$\max_i p_i = \max_i p_i \left(\sum_i p_i \right) \geq \sum_i p_i^2 \geq \max_i p_i^2.$$

□

We note also that $H_1(\boldsymbol{p}) \geq H_2(\boldsymbol{p}) \geq H_\infty(\boldsymbol{p})$, the first of which follows from Jensen' inequality. Hence the min entropy is the strongest measure of randomness, if H_∞ is large then so is H_2 and H_1, whereas there are distribution with a fixed min entropy for which the shannon entropy is arbitrarily large. This concept of min entropy will turn out to be useful later when we discuss objects known as randomness extractors.

These measures of randomness do share a some basic properties, which we collect below, writing $\tilde{H}$ for a 'generic' entropy measure

- $\tilde{H}(\boldsymbol{p}) \geq 0$ with equality iff the distribution is supported on a single element.
- $\tilde{H}(\boldsymbol{p}) \leq \log n$ with equality iff the distribution is uniform.
- For any doubly stochastic matrix X (non-negative entries with row and column sums equal to one), $\tilde{H}(X\boldsymbol{p}) \geq \tilde{H}(\boldsymbol{p})$, with equality iff $\boldsymbol{p}$ is uniform.

An immediate consequence of this final statement is that the entropy of the distributions π_i increase with every step of the random walk on a regular graph. Making this statement quantitative depends on the particular choice of entropy measure. Below we will do so for the H_2 measure, which will be simple due to it's relation to the ℓ_2 norm. However we note that since the H_2 and H_∞ measures are at a most a constant multiple away from each other, the same qualitative statement will hold for the H_∞ measure, which will be useful for us later.

For the Shannon entropy H_1 the precise relationship between the increase in H_1 in each step of a random walk and the spectral constant α is not known. However, there is a closely related constant, known as the *Log-Sobolev* constant, which in some way is related to this per step increase in energy, and there are known quantitative relations between it and α.

We will decompose our arbitrary distributions as $\boldsymbol{p} = \boldsymbol{u} + \boldsymbol{f}$ where $\boldsymbol{f} \perp \boldsymbol{u}$. We define a measure μ of how close $\boldsymbol{p}$ is to the uniform distribution via $\mu = \frac{||\boldsymbol{f}||}{||\boldsymbol{p}||} \leq 1$ (here and in what follows using the 2-norm), so that $\mu = 1$ iff $\boldsymbol{p}$ is uniform.

Then, noting that $\hat{A}\boldsymbol{f} \perp \boldsymbol{u}$ and $||\boldsymbol{u}||^2 = ||\boldsymbol{p}||^2 - ||\boldsymbol{f}||^2 = (1-\mu^2)||\boldsymbol{p}||^2$,

$$||\hat{A}\boldsymbol{p}||^2 = ||\hat{A}(\boldsymbol{u}+\boldsymbol{f})||^2 = ||\boldsymbol{u} + \hat{A}(\boldsymbol{f})||^2 = ||\boldsymbol{u}||^2 + ||\hat{A}\boldsymbol{f}||^2 \leq \left((1-\mu^2) + \alpha^2\mu^2\right)||\boldsymbol{p}||^2.$$

Hence,

$$\begin{aligned} H_2\left(\hat{A}\boldsymbol{p}\right) &= \log_2\left(||\hat{A}\boldsymbol{p}||^2\right) \\ &\geq \log_2\left(\left((1-\mu^2) + \alpha^2\mu^2\right)||\boldsymbol{p}||^2\right) \\ &= H_2\left(\boldsymbol{p}\right) - \log\left(1-\mu^2) + \alpha^2\mu^2\right) \\ &= H_2\left(\boldsymbol{p}\right) - \log\left(1-(1-\alpha^2)\mu^2\right). \end{aligned}$$

It follows that the 2-entropy of the sequence $\boldsymbol{\pi^1}, \boldsymbol{\pi^2}, \ldots$ is non-decreasing, and in fact is strictly increasing as long as the distributions $\boldsymbol{\pi^i}$ are not uniform. Furthermore, for better expanders (i.e., smaller α) the 2-entropy is growing faster.

4.2 Random walks resemble independent sampling

The general setup we will consider is an abstract sampling problem in which an unknown set B in a universe V of size n is 'bad' in some sense, and we want to sample the universe so as to avoid the bad set as much as possible. In particular, we're interested in doing so in an efficient way with regards to the number of random bits that we used.

In the introductory lecture, the set B represented the set of bad choices for a probabilistic algorithm, namely those for which the output disagreed with ground truth. Already there we saw the advantages of imposing, without any particular respect to the underlying universe, an expander graph structure on V. Here we will develop this idea further and show that, using such ideas, we can choose a sample from V based on a random walk in the graph. Perhaps surprisingly, the chance of hitting B with this (highly dependent) sample, will be very close to the probability of hitting B with a completely independent sample, where the quantitative truth of this statement will depend on the degree and expansion of the underlying expander graph.

Suppose then that we have an (n, d, α)-graph $G = V(, E)$ where the vertices in some subset $B \subseteq V$ are 'bad'. All we know about B is its cardinality $|B| = \beta n$. As in our previous example with 1-sided errors, we wish to sample at least one vertex outside of B. If we uniformly sample $t+1$ vertices $x_0, x_1, \ldots, x_t$ from V, then the probability that we 'fail', and fail to sample a vertex outside of B, is clearly β^{t+1}. Whilst this is tending to 0 exponentially fast, we have to use $(t+1)\log n$ random bits to make our sequence of samples. We will see that a similar level of accuracy can be obtained using substantially fewer bits. The basic idea will be to sample a starting vertex v uniformly, and take the vertex set of a random walk of length t from v as our sample. We will find that the chance of failure in this set-up will still be exponentially small in t.

In particular, to achieve an error rate of $\leq \epsilon$ we only need to use $\log n + O\left(\log \frac{1}{\epsilon}\right)$ many random bits, as opposed to $O\left(n \log \frac{1}{\epsilon}\right)$ many random bits. Note that the number of extra bits to achieve this improvement is now independent of $|V| = n$!

As a starting point let us interpret the expander mixing lemma as the case $t = 1$ of this approach. The conclusion of the lemma is that, for any subsets $S, T \subseteq V(G)$ of an (n, d, α)-graph,

$$\left| e(S,T) - \frac{d|S||T|}{n} \right| \leq \lambda d \sqrt{|S||T|} \leq \alpha d n.$$

It will be useful to rewrite this as

$$\left| \frac{e(S,T)}{dn} - \frac{|S||T|}{n^2} \right| \leq \alpha.$$

Let us imagine two different experiments. In the first we sample an ordered pair (i, j) of vertices uniformly in V^2 and call it a success if $i \in S$ and $j \in T$. Clearly the probability of success is $\frac{|S||T|}{n^2}$.

In the second experiment we randomly pick $i \in V$ and then choose j uniformly from the neighbours of i. This time the probability of success is

$$\begin{aligned} \sum_{s \in S, t \in T} \mathbb{P}(i = s, j = t) &= \sum_{s \in S} \mathbb{P}(i = s) \sum_{t \in T} \mathbb{P}(j = t | i = s) = \sum_{s \in S} \frac{1}{n} \sum_{t \in T} \frac{1}{d} \mathbb{1}_{t \sim s} \\ &= \frac{1}{dn} \sum_{s \in S, t \in T} \mathbb{1}_{t \sim s} = \frac{|E(S,T)|}{dn}. \end{aligned}$$

The expander mixing lemma then tells us that the success probabilities of these two, very different, methods of picking a pair of vertices are approximately equal, they differ by at most α.

Let us extend this intuition to longer random walks. Let us fix a (n, d, α)-graph G and a subset $B \subseteq V$ of cardinality $|B| = \beta n$. We consider the following stochastic process: Let $X_0 \in V$ be chosen uniformly at random and then perform a random walk $X_0, X_1, \ldots, X_t$ of length t on G. Denote by (B, t) the event that the random walk is confined to B, i.e., $X_i \in B$ for all i.

Theorem 4.5 (Ajtai-Komlós-Szemerédi and Alon-Feige-Widgerson-Zuckerman)**.** *Let $G = (V, E)$ be an (n, d, α)-graph and let $B \subseteq V$ be a subset of cardinality $|B| = \beta n$. Then for any $t \in \mathbb{N}$*

$$\mathbb{P}((B, t)) \leq (\beta + \alpha)^t.$$

Note that, if we picked the vertices $X_0, X_1, \ldots, X_t$ the probability would only be slightly smaller, β^t.

Let $P = P_B$ be the orthogonal projection on the subspace of coordinates belonging to B (working over $\mathbb{R}^V$). In other words, $P_{ij} = 1$ if $i = j \in B$ and 0 otherwise. We will use the following simple observation.

Lemma 4.6. *The probability of the event (B, t) be can expressed as*

$$\mathbb{P}((B, t)) = \left\| (P\hat{A})^t P \boldsymbol{u} \right\|_1.$$

Proof. The entry $(\hat{A}^t)_{xy}$ is the number of walks of length t from x to y divided by d^t, which can alternatively be thought of as the probability of reaching y from a random walk of length t started at x. If we consider instead $(P\hat{A})^t$ then the effect of the projection is that we only consider walks which are confined to B (after their first vertex). Hence if we apply $(P\hat{A})^t$ to $P\boldsymbol{u}$ then we get a vector supported on B, such that for each $b \in B$, $\left((P\hat{A})^t P\boldsymbol{u}\right)_v$ is $\frac{1}{n}$ times the probability that a random walk starting at v is confined to B, or in other words, the probability that $X_0 = v$ and $X_i \in B$ for all i.

In particular

$$\left\|(P\hat{A})^t P\boldsymbol{u}\right\|_1 = \sum_{v\in B}\left((P\hat{A})^t P\boldsymbol{u}\right)_v = \mathbb{P}((B,t)).$$

□

We also need the following lemma.

Lemma 4.7. *For any vector* $\boldsymbol{v}$

$$\left\|P\hat{A}P\boldsymbol{v}\right\|_2 \leq (\beta+\alpha)||\boldsymbol{v}||_2.$$

Proof. Note that, if we replace $\boldsymbol{v}$ with $P\boldsymbol{v}$ then, since P is idempotent, the left hand side is unchanged, whereas the right hand side becomes smaller since P is a contraction in ℓ_2. Hence we may assume that $\boldsymbol{v}$ is zero outside of B.

Similarly, we may assume that $\boldsymbol{v}$ is non-negative, and by the linearity of both sides we can assume that $||\boldsymbol{v}||_1 = 1$, and so

$$P\boldsymbol{v} = \boldsymbol{v} = \frac{\langle \boldsymbol{v}, \boldsymbol{u}\rangle}{\langle \boldsymbol{u}, \boldsymbol{u}\rangle}\boldsymbol{u} + \boldsymbol{z} = \boldsymbol{u} + \boldsymbol{z}$$

where $\boldsymbol{z} \perp \boldsymbol{u}$.

It follows that

$$P\hat{A}P\boldsymbol{v} = P\hat{A}\boldsymbol{u} + P\hat{A}\boldsymbol{z} = P\boldsymbol{u} + P\hat{A}\boldsymbol{z}.$$

and so

$$\left\|P\hat{A}P\boldsymbol{v}\right\|_2 \leq \|P\boldsymbol{u}\|_2 + \left\|P\hat{A}\boldsymbol{z}\right\|_2.$$

We claim that $\|P\boldsymbol{u}\|_2 \leq \beta||\boldsymbol{v}||_2$ and $\left\|P\hat{A}\boldsymbol{z}\right\|_2 \leq \alpha||\boldsymbol{v}||_2$, which together imply the claim.

For the first, since $||\boldsymbol{v}||_1 = 1$, and the support of $\boldsymbol{v}$ has size $\leq |B| \leq \beta n$, it follows from the Cauchy-Schwarz inequality that

$$1 = ||\boldsymbol{v}||_1 = \sum_{i\in B} 1 \cdot v_i \leq \sqrt{\sum_{i\in B} 1} \cdot \sqrt{\sum_{i\in B} v_i^2} = \sqrt{\beta n}||\boldsymbol{v}||_2.$$

However, since $\|P\boldsymbol{u}\|_2 = \sqrt{\frac{\beta}{n}}$ it follows that

$$\|P\boldsymbol{u}\|_2 \leq \beta||\boldsymbol{v}||_2.$$

For the second, we note that

$$\left\|P\hat{A}\boldsymbol{z}\right\|_2 \leq \left\|\hat{A}\boldsymbol{z}\right\|_2 \leq \alpha||\boldsymbol{z}||_2 \leq \alpha||\boldsymbol{x}||_2$$

using the fact that P is a contraction in ℓ_2, that $\boldsymbol{z}$ is orthgonal to $\boldsymbol{u}$ and so a combination of eignenvectors of $\hat{A}$ of eignenvalue at most α in modulus, and finally that $\boldsymbol{x}$ is the orthogonal sum of $\boldsymbol{u}$ and $\boldsymbol{z}$ and hence $||\boldsymbol{z}||_2 \leq ||\boldsymbol{z}||_2$. □

Using these two lemmas we can prove Theorem 4.5.

Proof of Theorem 4.5. By Lemma 4.6, Cauchy-Schwarz and Lemma 4.7 and we have that

$$\begin{aligned}\mathbb{P}((B,t)) &= \left\|(P\hat{A})^t P\boldsymbol{u}\right\|_1 \\ &\leq \sqrt{n}\left\|(P\hat{A})^t P\boldsymbol{u}\right\|_2 \\ &\leq \sqrt{n}(\beta+\alpha)^t \left\|\boldsymbol{u}\right\|_2 \\ &= (\beta+\alpha)^t.\end{aligned}$$

□

Theorem 4.5 shows that, for a good enough expander, the probability that a random walk of length t is contained in B is not much larger than the probability that a random sample of $t+1$ points is contained in B. There are two slight deficiencies with this argument, the first is that we are 'missing' a factor of β, and also we don't have a corresponding lower bound. We state without proof the following tighter version (it is perhaps best to think of β as being fixed, and α as being potentially much smaller).

Theorem 4.8. *If $\beta > 6\alpha$, then*

$$\beta(\beta+2\alpha)^t \geq \mathbb{P}((B,t)) \geq \beta(\beta-2\alpha)^t.$$

We note that it is also possible to derive "time dependent' versions of the upper bound in Theorem 4.5 (using simple adaptations of the proof), which we will need later in the course.

Theorem 4.9. *Let $G=(V,E)$ be an (n,d,α)-graph and let $B \subseteq V$ be a subset of cardinality $|B| = \beta n$. Then for any $t \in \mathbb{N}$ and any subset $K \subseteq \{0,1\ldots,t\}$*

$$\mathbb{P}(X_i \in B \text{ for all } i \in K) \leq (\beta+\alpha)^{|K|-1}.$$

Occasionally we will also have to deal with a situation where the excluded set of vertices depends on the time step, for which the following version is also useful.

Theorem 4.10. *Let $G=(V,E)$ be an (n,d,α)-graph, $t \in \mathbb{N}$ and let $B_0, B_1, \ldots, B_t \subseteq V$ be subsets of cardinality $|B_i| = \beta_i n$. Then*

$$\mathbb{P}(X_i \in B \text{ for all } i \in [t]) \leq \prod_{i=0}^{t-1}\left(\sqrt{\beta_i\beta_{i+1}}+\alpha\right).$$

This again follows from a slight adaptation of the previous arguments, using the fact that, writing P_i for the projection onto B_i

$$\mathbb{P}(X_i \in B \text{ for all } i \in [t]) = \left\| \prod_{i=1}^{t} \left(P_i \hat{A} \right) P_0 \boldsymbol{u} \right\|_1 \quad \text{and} \quad \left\| P_{i+1} \hat{A} P_i \boldsymbol{v} \right\|_2 \leq \left(\sqrt{\beta_i \beta_{i+1}} + \alpha \right) ||\boldsymbol{v}||_2.$$

As before, this simple approach seems to give away a factor of $\sqrt{\beta_0 \beta_t}$, which is important for certain applications.

4.3 Applications

4.3.1 Efficient error reduction in probabilistic algorithms

We return to the topic raised in the Introduction of reducing the error in probabilistic algorithms 'efficiently' (with respect to the number of extra random bits used).

Let A be a probabilistic algorithm for a language $\mathcal{L} \subseteq \{0,1\}^*$. Let us first deal with the simple case of one-sided errors, where A can only make errors on inputs $x \notin \mathcal{L}$, in which case $\mathcal{L}$ is in the class $\boldsymbol{RP}$. We will then deal with the case where A can make errors on inputs both inside and outside of $\mathcal{L}$, in which case $\mathcal{L}$ is in the corresponding class $\boldsymbol{BPP}$.

Recall that our 'benchmark' is given by simply repeating the algorithm t many times on the same input x with independent random strings $r_1, \ldots, r_t \in \{0,1\}^k$, where the output is given by the conjunction of the answers in the one-sided error case. In the two-sided error case it makes more sense to take the majority opinion as the output.

In both cases the error probability is decreasing exponentially in t, whereas the number of extra bits is increasing linearly in kt. We will show how to achieve an exponential reduction in the error probability in both cases using many fewer random bits.

One-sided error : Let A then be a 1-sided error randomised polynomial time membership algorithm for the language $\mathcal{L} \in \boldsymbol{RP}$. Given $x \in \{0,1\}^m$ the algorithm samples a string $r \in \{0,1\}^k$ uniformly at random, where k is polynomial in $|x|$, and computes a boolean function $A(x,r)$ such that $A(x,r) = 1$ if $x \in \mathcal{L}$ and $A(x,r) = 1$ with probability at most β if $x \notin \mathcal{L}$.

Let $G = (V, E)$ be an (n, d, α)-graph with $V = \{0,1\}^k$, with α sufficiently small compared to β. Note that the choice for α will put a lower bound on d (we will see later that we can take $d = O(\alpha^{-2})$).

Given $x \notin \mathcal{L}$ let $B(x) \subseteq \{0,1\}^k$ be the of strings r such that $A(x,r) = 1$, where $|B(x)| \leq \beta 2^k$. Let A' be the following membership algorithm:

(1) Choose $v_0 \in V$ uniformly at random;

(2) Take a random walk $(v_0, v_1, \ldots, v_t)$ in G of length t starting at v_0;

(3) Return $\bigwedge_{i=1}^{t} A(x, v_i)$.

Note that for this algorithm A' to be efficient, it has to be possible to efficiently compute the random walk, hence the importance of G being able to find explicit G.

By Theorem 4.5 the probability that A' will fail, which is the probability that $v_i \in B$ for each i, is at most $(\beta+\alpha)^t$. Compared to the naive algorithm from the Introduction, we achieve an exponential reduction in error probability, whilst only using $m + t\log d = m + O(t)$ random bits.

Two-sided error : We will show that the same trick can amplify the success probability of a randomised membership algorithm that makes errors on all types of inputs. More formally, a language $\mathcal{L}$ belongs to the class $\boldsymbol{BPP}$ if there is a polynomial time randomised algorithm A to decide whether a given input x belongs to $\mathcal{L}$ which is incorrect with probability at most β. As before we can reduce our error probability by running A multiple times on independently sampled strings and taking a *majority vote*. That is, we sample t strings $r_1, r_2, \ldots, r_t$ uniformly and independently and we look at the numbers $A(x,r_1), A(x,r_2), \ldots, A(x,r_t)$. If more than half of them are one then we say that $x \in \mathcal{L}$ and if more than half of them are zero we say that $x \notin \mathcal{L}$.

How accurate will this algorithm be? Well, if $x \in \mathcal{L}$ then the number of ones is distributed as $\mathrm{Bin}(t, 1-\beta)$, and similarly if $x \notin \mathcal{L}$ the number of zeroes is distributed as $\mathrm{Bin}(t,1-\beta)$, and so this is really a question about the *tails* of the binomial distribution $\mathbb{P}(\mathrm{Bin}(t,1-\beta) > \Delta)$.

A standard bound is the Chernoff bound, which tells us that, under the right parameterisation, these tails decrease exponentially.

Lemma 4.11. *Let $X \sim Bin(n,p)$, let $0 \le \delta \le 1$ and let $\mu := \mathbb{E}(X)$ then*

$$\mathbb{P}(|X-\mu| \ge \delta\mu) \le \exp\left(-\frac{\delta^2\mu}{3}\right).$$

It follows that the probability that A' fails is shrinking exponentially in t. More precisely

$$\begin{aligned}\mathbb{P}(A' \text{ fails}) &= \mathbb{P}\left(\mathrm{bin}(t,1-\beta) \le \frac{t}{2}\right) \le \mathbb{P}\left(|\mathrm{Bin}(t,1-\beta) - (1-\beta)t| \ge \left(\frac{1}{2}-\beta\right)t\right) \\ &\le \exp\left(-\frac{\left(\frac{1}{2}-\beta\right)^2(1-\beta)t}{3}\right).\end{aligned}$$

To save on randomness we can again use random walks on an expander graph.

As before, let us suppose we have an input $x \in \{0,1\}^m$ for which A uses k-bit random strings $r \in \{0,1\}^k$ and let G be an (n,d,α)-graph with vertex set $V = \{0,1\}^k$. Again, let $B(x)$ be the set of strings for which $A(x,r)$ gives the wrong output. Our modified algorithm A' then works as follows:

(1) Choose $v_0 \in V$ uniformly at random;

(2) Take a random walk $(v_0, v_1, \ldots, v_t)$ in G of length t starting at v_0;

(3) Return *majority*$\{A(x,v_i)\colon 0 \le i \le t\}$.

The algorithm will then fail only if a majority of the v_is belong to $B(x)$. Fix a set of indices $K \subseteq \{0, 1, \ldots, t\}$ of cardinality $|K| \geq \frac{t+1}{2}$. Then by Theorem 4.9 the probability that $v_i \in B$ for all $i \in K$ is at most $(\beta + \alpha)^{|K|-1} \leq (\beta + \alpha)^{\frac{t-1}{2}}$.

Assuming that $\beta + \alpha$ is small enough (less than $\frac{1}{8}$ say) we be apply the union bound over all possible choices of K to conclude that

$$\mathbb{P}(A' \text{ fails}) \leq 2^t(\beta + \alpha)^{\frac{t-1}{2}} \leq 2^t 2^{\frac{-3(t-1)}{2}} \leq 2\sqrt{2}2^{-\frac{t}{2}}.$$

So, again we achieve an exponential reduction in the error probability using only $m + O(t)$ random bits.

In general, this idea that a random walk on an expander graph gives a good approximation of a random set can be used in many algorithmic applications, for example it gives a quick and easy way to approximate the average of any function over a space.

4.3.2 Hardness of approximating maximum clique size

We now turn to a different application of random walks on expanders to computational complexity, showing how they are used in enhancing hardness of approximation factors of the clique problem. Let's begin with a little background on hardness of approximation.

A *clique* in a graph G is a subset $S \subseteq V(G)$ in which all pairs of vertices are adjacent. The *clique number*, denoted $\omega(G)$, is defined as the largest size of a clique in G. It is an important computational problem to be able to compute, or estimate this parameter of a graph. Specifically, given a graph G and an integer k we are asked to determine whether $\omega(G) \geq k$.

There is no known subexponential (on all inputs G and k) algorithm for this problem, and it seems unlikely that one will exist. Indeed, determining the clique number of a graph is known to be ***NP**-hard*, if there is a polynomial time algorithm to compute the clique number of a graph, then there is a polynomial time algorithm to solve every problem in $\boldsymbol{NP}$, i.e., $\boldsymbol{P} = \boldsymbol{NP}$ (which, as far we know, seems unlikely to be the case). Here, roughly, $\boldsymbol{P}$ is the class of 'yes-no' problems which can be solved in polynomial time, whereas $\boldsymbol{NP}$ is the class of 'yes-no' problems whose solutions can be 'verified' in polynomial time. There are many natural and important problems which are known to be $\boldsymbol{NP}$-hard.

When it is hard to find an optimum solution to a problem (i.e., the size of the *largest* clique), a natural relaxation is to seek an approximate solution. A famous result in the 90s, the so called $\boldsymbol{PCP}$ theorem, which in some way says that solutions to problems in $\boldsymbol{NP}$ can be checked very efficiently by a randomised algorithm, implies that for many difficult optimisation problems, it is still difficult to approximate their solutions (unless $\boldsymbol{P} = \boldsymbol{NP}$). In particular, it is hard to approximate $\omega(G)$ to even a constant factor.

Theorem 4.12 (Feige-Goldwasser-Lovász-Safra-Szegedy)**.** *There are constants $0 < \delta_2 < \delta_1 < 1$ such that it is $\boldsymbol{NP}$-hard to decide for a given n vertex graph whether $\omega(G) \geq \delta_1 n$ or $\omega(G) \leq \delta_2 n$.*

(that is, the problem of determining for any graph for which one of the two hold, which holds, is $\boldsymbol{NP}$-hard). In fact, the simplest form of the $\boldsymbol{PCP}$ theorem is almost equivalent to this statement.

The purpose of this section will be to show that even obtaining a *very rough* approximation to $\omega(G)$ is still $\boldsymbol{NP}$-hard. More precisely, we will show that it is $\boldsymbol{NP}$-hard to approximate $\omega(G)$ to within a (multiplicative) factor of n^ϵ for some fixed $\epsilon > 0$.

Theorem 4.13. *There exists a constant $\epsilon > 0$ such that, if there exists a polynomial time algorithm A whose output on every n-vertex graph G satisfies*

$$n^{-\epsilon}\omega(G) \leq A(G) \leq n^\epsilon \omega(G),$$

then $\boldsymbol{P} = \boldsymbol{NP}$.

The idea of the proof will be that if such an algorithm A existed, we could use it to create an efficient algorithm B to approximate the clique number to within a constant factor. Such a conversion is called a *reduction*. If this reduction is deterministic, then the existence of algorithm B would imply by Theorem 4.12 that the $\boldsymbol{NP}$-hard problem of approximating the clique number to within a constant factor had a polynomial time solution, and hence $\boldsymbol{P} = \boldsymbol{NP}$.

We will first present a *probabilistic reduction* between the two problem, so that the algorithm B is a randomised algorithm with 2-sided errors. Note that this would still has interesting complexity consequences, namely that the existence of A would imply that $\boldsymbol{NP} \subseteq \boldsymbol{BPP}$. Moreover, it is a relatively standard fact (using the self-reducibility of $\boldsymbol{NP}$-hard problems) that the inclusion $\boldsymbol{NP} \subseteq \boldsymbol{BPP}$ implies that $\boldsymbol{NP} \subseteq \boldsymbol{RP}$. After describing this reduction we will show how to eliminate the randomness from B, for which we will use walks on expander graphs.

We note that, in fact, a much stronger hardness result is known about approximating the clique number, which requires much more advanced techniques. Indeed, Hastad showed that efficiently approximating $\omega(G)$ to within a factor of $n^{1-\delta}$ for any $\delta > 0$ would imply that $\boldsymbol{RP} = \boldsymbol{NP}$, and quite recently this was derandomised by Zuckerman to show that such an approximation is even $\boldsymbol{NP}$-hard. Note that, a factor n approximation is trivial, so this result is essentially tight!

The probabilistic reduction:

We wish to convert an algorithm A for approximating $\omega(G)$ to within a factor of n^ϵ to a probabilistic algorithm B to distinguish between the two cases of Theorem 4.12. In order to apply B to a given n-vertex graph $G = (V, E)$, consider a graph H with vertex set V^t where $t = \log n$. Two vertices $(v_1, v_2, \ldots, v_t)$ and $(u_1, u_2, \ldots, u_t)$ will be adjacent in H if the subgraph of G induced on $\{v_1, v_2, \ldots, v_t, u_1, u_2, \ldots, u_t\}$ is a clique. It will turn out that the property of $\omega(G)$ being below $\delta_2 n$ or above $\delta_1 n$ will be amplified significantly in H, and this amplification is so strong that a random subset of $m = \text{poly}(n)$ vertices in H tends to behave very differently with respect to the clique number of the induced graph.

Algorithm B then works as follows:

(1) Choose m random vertices from V^t and compute the subgraph H' of H induced on these vertices;

(2) Apply algorithm A to H';

(3) We return 1 if $A(H') > \frac{1}{2}\delta_1^t m^{1-\epsilon}$, and otherwise 0;

where 1 indicates that $\omega(G) \geq \delta_1 n$ and 0 indicates that $\omega(G) \leq \delta_2 n$.

In order to analyse this algorithm we will need the following simple combinatorial observation.

Claim 4.14. Every clique in H is contained in a clique of the form S^t where S is an inclusion maximal clique in G. In particular, $\omega(H) = \omega(G)^t$.

Proof. Clearly if S is a clique in G then S^t is a clique in H, and hence $\omega(H) \geq \omega(G)^t$. Conversely, let S' be a clique in H, and let $S \subseteq V(G)$ be the set of the vertices of G which appear as an entry in some t-tuple in S'. Then clearly S must form a clique in G, and $|S'| \leq |S|^t$, and so also $\omega(H) \leq \omega(G)^t$. □

So, we need to show two things, for an appropriate choice of $m = \text{poly}(n)$ and $\epsilon = \Theta(1)$.

(a) If $\omega(G) \geq \delta_1 n$, then with a large probability $A(H') \geq \frac{1}{2}\delta_1^t m^{1-\epsilon}$;

(b) If $\omega(G) \leq \delta_2 n$, then with a large probability $A(H') \leq 2\delta_2^t m^{1+\epsilon} \leq \delta_1^t m^{1-\epsilon}$;

For the first claim, we assume that there is a clique $K \subseteq G$ of size $\delta_1 n$, and hence by Claim 4.14 there is a clique $K' \subseteq H$ of size $(\delta_1 n)^t$. The number of vertices of K' contained in H' is binomially distributed with mean

$$\frac{m|V(K')|}{|V(H)|} \geq \delta_1^t m.$$

However then, by the Chernoff bounds, the probability that H' contains less than $\frac{1}{2}\delta_1^t m$ is at most

$$\exp\left(-\frac{\delta_1^t m}{12}\right) = o_n(1).$$

Hence, with high probability $\omega(H') \geq \frac{1}{2}\delta_1^t m$ and so $A(H') \geq \frac{1}{2}\delta_1^t m^{1-\epsilon}$

For the second claim we need to show that it is very unlikely that the m vertices we sample from H include a large clique. For the analysis of this it suffices to consider the *inclusion-maximal* cliques in H, of which, by Claim 4.14, there are at most 2^n. For any inclusion maximal clique L in H, it has size at most $(\delta_2 n)^t$ by assumption and so again the number of vertices of K contained in H' is stochastically dominated by a binomial random variable with mean

$$\frac{m|V(K)|}{|V(H)|} \geq \delta_2^t m.$$

Then, again by the Chernoff bound, for any fixed K the probability that H' contains at least $2\delta_2^t m$ many vertices of K is at most

$$\exp\left(-\frac{\delta_1^t m}{3}\right),$$

and so by a union bound the probability that H' contains at least $2\delta_2^t m$ of any maximal clique (which is the probability that H' contains a clique of size at least $2\delta_2^t m$) is at most

$$2^n \exp\left(-\frac{\delta_1^t m}{3}\right) = o_n(1)$$

as long as $m\delta_2^t = mn^{\log \delta_2} \gg n$. In this case, with high probability $\omega(H') \geq 2\delta_2^t m$ and so $A(H') < 2\delta_2^t m^{1+\epsilon}$.

It remains to choose $\epsilon = \Theta(1)$ and $m = \text{poly}(t)$ such that $m\delta_2^t = mn^{\log \delta_2} \gg n$ and $2\delta_2^t m^{1+\epsilon} \leq \frac{1}{2}\delta_1^t m^{1-\epsilon}$. The first just requires that m is a large enough power of n, for example $m = n^{-\log \delta_2 + 2}$. For the second, after rearranging and taking logs, we see that it is sufficient that

$$2\epsilon \leq \frac{t}{\log m} \log \frac{\delta_1}{4\delta_2} = \frac{1}{2 - \log \delta_2} \log \frac{\delta_1}{4\delta_2},$$

which is just a function of δ_1 and δ_2.

The deterministic reduction: Let us show then how we can use expander graphs to *derandomise* the previous reduction. Again we are assuming the existence of a polynomial time algorithm A, but now we will derive the existence of a deterministic alogorithm B' to distinguish between the two cases of Theorem 4.12.

The only difference between B' and the B described in the preceding section if that B' will use *derandomised sampling* to construct the graph H'. To do so, let us choose some (n, d, α)-graph G' on the same vertex set as G. In order to select the vertices in H', we no longer take a random sample of t-tuples from $V(G)^t$. Instead we consider the set of all t-tuples which represent a length $(t-1)$ walk in the graph G'. This graph has $m = nd^{t-1}$ vertices and, since d is fixed and we again take $t = \Theta(\log n)$, m is polynomial in n. The edge set of H' is formed as before, by taking edges between t-tuples whose union is a clique in G.

We have already established that random walks on G' should 'approximate' in some sense random subsets of $V(G)$. We need to give an analogue of this principle in the current setting as well.

Claim 4.15. If $\omega(G) \leq \delta_2 n$, then $\omega(H') \leq (\delta_2 + 2\alpha)^t m$.

Proof. As before, a clique in H' corresponds to a collection of $(t-1)$-walks in G' which are confined to some fixed clique in G. If we perform a uniform random walk of length $t-1$, choosing a starting vertex uniformly at random, we are equally likely to achieve any vertex in H' and hence, the size of the largest clique in H' is equal to m times the maximum over all cliques S of G of the probability that such a walk is confined to S. Since by assumption $|S| \leq \delta_2 n$, it follows from Theorem 4.8 that

$$\frac{\omega(H')}{m} = \max_{S \text{ a clique}} \mathbb{P}(\text{A random walk contained in } S) \leq (\delta_2 + 2\alpha)^t.$$

□

The complementary statement that we need is then:

Claim 4.16. If $\omega(G) \geq \delta_1 n$, then $\omega(H') \geq (\delta_1 - 2\alpha)^t m$.

Proof. Let S be a clique in G of cardinality $|S| \geq \delta_1 n$. Again by Theorem 4.8 a random walk of length $t-1$ in G is confined to S with probability at least $(\delta_1 - 2\alpha)^t$. The conclusion then follows as before □

The rest of the proof follows in a similar manner as before, and we leave the details to the exericise sheet.

5 A geometric view of expander graphs

An attractive feature of expander graph is the multiple different viewpoints we can consider them from. In the previous sections we saw how we can consider these objects combinatorially, algebraically and probabilistically. In this section we add to this a geometric perspective.

5.1 Some background on isoperimetry

We will relate the expansion of a graph G to notions of isoperimetry. The basic idea of isoperimetry is the idea of relating the boundary of a set to it's volume. In any context where both volume and boundary have a definition, the isoperimetric problem asks over all sets of the same (iso) volume, which has the smallest boundary (perimeter)?

We will have all have seen the prototypical example of a such a question, which was considered as early as ancient Greece.

Question 5.1. *Of all simple closed curves in the plane enclosing a fixed area, which has the shortest length?*

The answer to this problem was known to the Greeks, that a circle is the unique minimiser, but it took a very long time for this to be proved! The first rigorous proof was claimed by Steiner, using an idea known as Steiner symmetrisation. Whilst a flaw was pointed out by Weierstrass, several correct proofs were found soon after, some of which used Steiner's ideas. Analogous ideas have turned out to be very useful in the study of isoperimetric inequalities on graphs. For this reason, and for histroical signigicance, we briefly describe the approach.

We first note that there is no loss of generality in restricting our attention to convex domains, since the convex hull of a closed planar domain has larger area, but a smaller circumference. Given a convex domain K we can *symmetrise* it as follows. We describe a symmetrisation around the x-axis, but this can be perfomed with respect to any line through the origin.

K has some projection $[x_1, x_2]$ onto the x-axis, and for each $x \in [x_1, x_2]$ we consider the set of points K_x of the form $(x, y) \colon y \in \mathbb{P}$, which again must be some interval $[y_1, y_2]$. We translate this set so that is is symmetric about the x-axis, i.e., $K'_x = [\frac{y_1 - y_2}{2}, \frac{y_2 - y_1}{2}]$. The symmetrised set is that $K = \bigcup_{x \in [x_1, x_2]} K'_x$. It can be shown that this transformation preserves the area of K, and does not increase the circumference. In particular, the circumference of an optimal K (if one should exists, which was the mistake pointed out by Weierstrass) must be invariant under these operations. In fact, a little more work shows that K itself must be invariant under these operations, and from there is can be shown that K is a disc.

Discrete versions of these symmetrisation arguments have been useful in various discrete geometry problems, in particular the idea of *compressions*, which have been very useful in studying certain types of hypergraphs, can maybe be though of as a generalisation of these ideas.

5.2 Graph isoperimetric problems

A simple, but quite important, observation is that a graph has some very natural notions of boundary and volume. A natural measure of the volume of a set of vertices S is its cardinality $|S|$. There are perhaps three main notions of boundary we might want to consider, the *edge boundary* of a set S, $\partial_E(S) = \{e\colon e \cap S \neq \emptyset \text{ and } e \cap S^c \neq \emptyset\}$ and the *inner* and *outer vertex boundary* of a set

$$\partial_V^i(S) = \{v \in S\colon N(v) \cap S^c \neq \emptyset\} \qquad \text{and} \qquad \partial_V^o(S) = \{v \in S^c\colon N(v) \cap S \neq \emptyset\}.$$

For some reason, the inner boundary is rarely considered, and we will only ever refer to $\partial_V(S) = \partial_V^o(S)$. For each of these notions of boundary we can define the *isoperimetric parameter*

$$\Phi_E(G,k) = \min\{\partial_E(S)\colon S \subseteq V, |S| = k\} \text{ and } \Phi_V(G,k) = \min\{\partial_V(S)\colon S \subseteq V, |S| = k\}.$$

It is desirable to compeltely understand these parameters for natural families of graphs, or to computationally determine or estimate these parameters for given G and k. In the broadest generality these computational problems are very difficult (co-$\boldsymbol{NP}$-hard), and much research has been dedicated to finding efficient algorithms to approximate these quantities.

Let us illustrate these notions with an important family of graphs for which the isoperimetric problems have been completely solved (using this notion of compressions), the hypercubes. The d-dimensional cubes Q^d is a graph on vertex set $V(Q^d) = \{0,1\}^d$ where two vertices are adjacent if they differ in precisely one coordinate, i.e., if they have hamming distance one. The cube arises very naturally in a whole host of combinatorial contexts. One particular reason for this is its natural interpretation as the Hasse diagram of the subset relation on a d element set.

The isoperimetric problem has been solved in a very explicit manner in these graphs, not only are the values of the functions $\Phi_E(Q^d,k)$ and $\Phi_V(Q^d,k)$ known for each k, the minimising sets of vertices are well understood.

- For the edge isoperimetric problem the minimisers can be seen to be subcubes. We have the inequality $\Phi_E(Q^d,k) \geq k(d - \log k)$, and equality can be seen to hold when S is the set of vertices of a subcube of Q^d.
- For the vertex isoperimetric problem the minimisers can be seen to be *Hamming balls*, balls of a fixed radius around the origin. Here we see that if $k = \sum_{i=1}^{j} \binom{d}{i}$ then $\Phi_E(Q^d,k) = \binom{d}{j+1}$.

5.3 The discrete Laplacian

In classic vector analysis the Laplace operator $\Delta(f) = \operatorname{div}(\operatorname{grad}(f))$ is a differential operator which arises naturally in the differential equations describing many physical phenomena. Informally, it measures how well the value $f(p)$ is approximated by the average of f over small spheres centered at p.

We will define a discrete analogue which will turn out to be useful for several reasons. In order to do so, it will make sense to first introduce discrete analogues of the gradient and divergence

operates in graphs. The correct definition of the Laplacian will then be apparent. Suppose we have an undirected graph $G = (V, E)$ and we fix an arbitrary orientation of the edges (note: this choice of orientation will not affect any of the definitions).

Let K be the $V \times E$ incidence matrix of G where the entry

$$K_{u,e} = \begin{cases} +1 & \text{if the edge } e \text{ exits the vertex } u \\ -1 & \text{if the edge } e \text{ enters the vertex } u \\ 0 & \text{otherwise.} \end{cases}$$

We can then define the *gradient* operator. Given any function $f : V \to \mathbb{R}$, which we view as a row vector in $\mathbb{R}^V$, the gradient operator maps $\boldsymbol{f}$ to $\boldsymbol{f}K$, a vector indexed by E. The gradient measures the change of $\boldsymbol{f}$ along the edges of the graph, so that if $e = uv$ then $(\boldsymbol{f}K)_e = f_u - f_v$.

The *divergence* operator takes a function $g : E \to \mathbb{R}$ on the edge set of G, now viewed as a column vector in $\mathbb{R}^E$, and maps it to $K\boldsymbol{g}$, a vector indexed by V. If we think of $\boldsymbol{g}$ as describing a flow, then its divergence at a vertex is the net outbound flow, namely,

$$(K\boldsymbol{g})_v = \sum_{e \text{ exits } v} g_e - \sum_{e \text{ enters } v} g_e.$$

We can then define the *Laplacian*, which maintaining the analogy should map a function $f : V \to \mathbb{R}$ to $\boldsymbol{f}KK^T\boldsymbol{f}^T$. The matrix $L = L_G = KK^T$ is accordingly called the (discrete) Laplacian of G. This matrix is also sometimes known as the Tutte matrix of G in the graph theory literature and appears in various contexts there, such as the matrix-tree Theorem adna randomised matching algorithms.

A simple calculation shows that L is a symmetric matrix with rows and columns indexed by V where

$$L_{u,v} = \begin{cases} -1 & \text{if } (u, v) \in E \\ \deg(u) & \text{if } u = v. \end{cases}$$

Indeed, $L_{u,v} = \sum_e K_{u,e}K_{v,e}$ and so if $u \neq v$ then the summand is non-zero iff $e = (u, v)$, where it takes the value -1. Conversely, if $u = v$ then the summand is one for each e which is incident to $u = v$.

From this it is easy to deduce the following equality, which will be useful later.

$$\begin{aligned} \boldsymbol{f}L\boldsymbol{f}^T &= \sum_u (\boldsymbol{f}L)_u f_u \\ &= \sum_u \deg(\mathrm{u}) f_u^2 + \sum_u \sum_{(u,v)\in E} -f_v f_u \\ &= \sum_{(u,v)\in E} f_u^2 + f_v^2 - 2f_u f_v \\ &= \sum_{(u,v)\in E} (f_u - f_v)^2. \end{aligned} \tag{5.1}$$

In particular we have the following consequence.

Proposition 5.2. *For every graph G the matrix L_g is positive semi-definite. Its smallest eigenvalue is 0, and the corresponding eigenfunction is $\boldsymbol{u}$.*

In fact, we can say much more about the spectrum of L. Though it is possible to develop some of this theory for irregular graphs, the regular case is simpler to state an analyse.

Lemma 5.3. *The Laplacian L_G of a d-regular graph G satisfies:*

- $L = L_G = dI - A(G)$;
- *The spectrum of L is in $[0, 2d]$.*
- *The smallest eigenvalue of L is zero.*
- *The spectral gap of G is equal to the smallest positive eigenvalue of L.*

5.4 The Cheeger constant and inequality

Many of the things we do here for graphs have been studied previously in the framework of Riemannian manifolds. It won't be important for us, but these are spaces that look locally like $\mathbb{R}^n$ and carry a differentiable structure with a smoothly varying notion of inner product among tangent vectors. This allows us to carry out many familiar operations from calculus, compute volumes and distances. The following notion is then an idea of expansion that arose in the study of these spaces.

Definition. The *Cheeger constant* of a compact n-dimensional Riemannian manifold M is

$$h(M) = \inf_A \frac{\mu_{n-1}(\partial A)}{\min\{\mu_n(A), \mu_n(M \setminus A)\}},$$

where this infimum is taken over open subsets of M, ∂A is the boundary of A and μ_i is the ith dimensional measure.

The analogy with edge expansion should hopefully be obvious: We partition M into two parts A and its complement $M \setminus A$, and consider the ratio between two quantites

- The ($(n-1)$-dimensional) measure of the boundary of A; and
- The minimum of the (n-dimensional) measures of A and its complement.

(of course, as in the definition of $h(G)$, we could alternatively take the minimum over open sets of 'volume' at most $\frac{1}{2}$ of the whole space).

As discussed above, it is possible to develop much of the theory of differential calculus over Riemannian manifolds. In particular, associated with any Riemannian manifold M there is a linear differential operator known as the Laplacian, which is defined on real functions $f\colon M \to \mathbb{R}$ in the familiar way $\Delta(f) = \mathrm{div}(\mathrm{grad}(f))$. If $\Delta f = \lambda f$ we say that f is an eigenfunction of the Laplacian with eigenvalue λ. It can be shown that also in this setting the eigenvalues of λ are all non-negative, and that its smallest eigenvalue is zero, corresponding to the constant eigenfunction. A fundamental theme in this area is then the connection between expansion (in the form of the Cheeger constant h) and the spectrum of this operator.

Theorem 5.4 (Cheeger)**.** *Let M be a compact Riemannian manifold, and let λ be the smallest positive eigenvalue of its Laplacian. Then $\lambda \geq \frac{h^2}{4}$.*

5.5 Expansion and the spectral gap

Note the (not coincidental) similarities between the previous inequality and the upper bound of Theorem 3.1. Recall that the expansion ratio of a graph G is defined as

$$h(G) = \min_{S\colon |S| \leq \frac{n}{2}} \frac{|\partial S|}{|S|}.$$

Theorem 5.5 (Dodziuk/ Alon and Milman / Alon)**.** *Let G be a d-regular graph with spectrum $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Then*

$$\frac{d-\lambda_2}{2} \leq h(G) \leq \sqrt{2d(d-\lambda_2)}$$

Before we, finally, give a proof of this theorem, let us say a few words about it's optimality, and related concepts.

Firstly we note that both upper and lower bounds are tight.

1. For the lower bound recall our discussion of the isoperimetric inequality on the hypercube Q^d. Taking S to be the $(d-1)$-dimensional subcube we see that $h(G) \leq 1$ and conversely,

$$h(G) = \min_{S\colon |S| \leq \frac{n}{2}} \frac{|\partial S|}{|S|} \geq \min_{S\colon |S| \leq \frac{n}{2}} \frac{|S|(d - \log|S|)}{|S|} \geq \min_{S\colon |S| \leq \frac{n}{2}} (d - \log|S|) = 1.$$

   On on the other hand it is reasonably easy to show that Q^d has eigenvalues $d, d-2, d-4, \ldots, -d$ (with varying multiplicites), and so the spectral gap is $d - \lambda_2 = 2$.

2. For the upper bound we consider an n-vertex cycle C_n. The edge expansion ratio can be seen to be $\Theta\left(\frac{1}{n}\right)$, attained with S being half the cycle. On the other hand the adjacency matrix of C_n is a circulant matrix, whose eigenvalues can be easily shown to be $\omega_j + \omega_j^{n-1} = 2\cos\left(\frac{2\pi j}{n}\right)$, where $\omega_j = e^{\frac{2\pi i j}{n}}$ for $j = 0, \ldots, n-1$. In particular,

$$\lambda_2 = 2\cos\left(\frac{2\pi}{n}\right) = 2 - \Theta\left(\frac{1}{n^2}\right)$$

   and so $2 - \lambda_2 = \Theta\left(\frac{1}{n^2}\right)$.

We also note one particularly important generalisation of this theorem, to general *reversible Markov chains*. For such chains we can defined a weighted analogue of edge-expansion, called *conductance*, and give similar bounds on this in terms of the spectral gap of the adjacency matrix of the chain. This was done by Jerrum and Sinclair, and had a big impact on the analysis of the convergence of Monto-Carlo algorithms.

Proof of Theorem 5.5. The easy direction is the lower bound, whose proof is similar to the proof of the Expander mixing lemma, where we now leverage the fact that the two sets we consider are complementary.

Recall that the first eigenvector of a regular graph is $\boldsymbol{u}$, and so to bound the second eigenvalue from below, it suffices to exhibit a vector $\boldsymbol{f} \perp \boldsymbol{u}$ which 'scales' by a large factor. More precisely we will find such a vector $\boldsymbol{f}$ whose *Rayleigh quotient* is large

$$\frac{\boldsymbol{f}A\boldsymbol{f}^T}{||\boldsymbol{f}||_2^2} \geq d - 2h(G),$$

from which it follows that $\lambda_2 \geq d - 2h(G)$.

The vector $\boldsymbol{f}$ we will consider comes from an optimal set S for edge expansion, i.e., $h(G) = \frac{e(S,S^c)}{|S|}$ and $|S| \leq \frac{n}{2}$. We take then $f = |S^c|\mathbf{1}_S - |S|\mathbf{1}_{S^c}$, where as before $\mathbf{1}_X \in \mathbb{R}^V$ is the indicator function of a subset $X \subseteq V$. Note that, indeed, $\boldsymbol{f} \perp \boldsymbol{u}$.

It is simple to express the Rayleigh quotient in terms of graph parametes:

$$||\boldsymbol{f}||_2^2 = |S^c|^2|S| + |S|^2|S^c| = |S||S^c|(|S| + |S^c|) = n|S||S^c|,$$

$$\begin{aligned}\boldsymbol{f}A\boldsymbol{f}^T &= \sum_u (\boldsymbol{f}A)_u f_u = \sum_u f_u \sum_v f_v A_{u,v} = \sum_{(u,v)\in E} 2f_u f_v \\ &= 2\left(e(S)|S^c|^2 + e(S^c)|S|^2 - |S||S^c|e(S,S^c)\right).\end{aligned}$$

Now, since G is d-regular, we can substitute $2e(S) = d|S| - e(S,S^c)$ and $2e(S^c) = d|S^c| - e(S,S^c)$, to see that

$$\begin{aligned}\boldsymbol{f}A\boldsymbol{f}^T &= \left(d|S| - e(S,S^c)\right)|S^c|^2 + \left(d|S^c| - e(S,S^c)\right)|S|^2 - 2|S||S^c|e(S,S^c) \\ &= d|S||S^c|^2 + d|S^c||S|^2 - e(S,S^c)\left(|S^c|^2 + |S|^2 - 2|S||S^c|\right) \\ &= nd|S||S^c| - n^2 e(S,S^c).\end{aligned}$$

from which it follows that

$$\lambda_2 \geq \frac{\boldsymbol{f}A\boldsymbol{f}^T}{||\boldsymbol{f}||_2^2} = \frac{nd|S||S^c| - n^2 e(S,S^c)}{n|S||S^c|} = d - \frac{ne(S,S^c)}{|S||S^c|} \geq d - 2h(G),$$

using that $h(G) = \frac{e(S,S^c)}{|S|}$ and $|S^c| \geq \frac{n}{2}$ by assumption.

So, let us move on to the difficult direction, the upper bound, that high expansion implies a large spectral gap. The key idea here is to consider the eigenvector $\boldsymbol{v} = \boldsymbol{v_2}$ associated with λ_2. It turns out that the weighting $\boldsymbol{v} \in \mathbb{R}^V$ will be close in some way to an edge cut with few crossing edges. Indeed, if $\boldsymbol{v} \in \{-1,1\}^V$ then it would be relatively clear which cut to consider, however in general we can split the vertices according to some threshold value, and a natural one to consider would be 0.

So, let us consider $\boldsymbol{w} = \boldsymbol{v}^+$, the vector defined by $w_x = \max\{v_x, 0\}$ and $V^+ = \mathrm{supp}(\boldsymbol{w}) = \{x \colon w_x > 0\}$. Without loss of generality we may assume that $|V^+| \geq \frac{n}{2}$, since otherwise we can consider $-\boldsymbol{v}$, which is also an eigenvector with eigenvalue λ_2. We now aim to prove two bounds on the Rayleigh quotient of $\boldsymbol{w}$, but a trick here will be to consider, rather than the adjacency matrix A, the Laplacian L:

(i) $\frac{\boldsymbol{w}L\boldsymbol{w}^T}{||\boldsymbol{w}||_2^2} \leq d - \lambda_2$;

(ii) $\frac{h(G)^2}{2d} \leq \frac{\boldsymbol{w}L\boldsymbol{w}^T}{||\boldsymbol{w}||_2^2}$.

from which it clearly follows that $h \leq \sqrt{2d(d-\lambda_2)}$ as claimed.

We start by proving (i). We first note that, since $\boldsymbol{v}$ is an eigenvector of L with eigenvalue $d-\lambda_2$, for any $x \in V^+$ we can write

$$(\boldsymbol{w}L)_x = \big(\boldsymbol{w}(dI-A)\big)_x = dw_x - \sum_{y\in V} A_{xy}w_y = dv_x - \sum_{y\in V^+} A_{xy}v_y$$
$$\leq dv_x - \sum_{y\in V} A_{xy}v_y = (L\boldsymbol{v})_x = (d-\lambda_2)v_x.$$

Then, since $w_x = 0$ for all $w \notin V^+$ we see that

$$\boldsymbol{w}L\boldsymbol{w}^T = \sum_{x\in V}(\boldsymbol{w}L)_x w_x = \sum_{x\in V^+}(\boldsymbol{w}L)_x w_x \leq \sum_{x\in V^+}(d-\lambda_2)v_x^2 = (d-\lambda_2)\sum_{x\in V} w_x^2 = (d-\lambda_2)||\boldsymbol{w}||_2^2.$$

So, it remains to prove (ii). To this end, let us define an orientation on $E(G)$ given by orienting each edge xy as (x,y) if $w_x \geq w_y$ (breaking ties arbitrarily). We will use this orientation, as in the previous section, to define the $V \times E$ incidence matrix K of G.

Let us consider the following quantity

$$B_{\boldsymbol{w}} := \sum_{(x,y)\in E} |w_x^2 - w_y^2| = \sum_{(x,y)\in E} w_x^2 - w_y^2,$$

where the sum is taken over the oriented edges (x,y) of E. We will show that

$$h(G)||\boldsymbol{w}||_2^2 \leq B_{\boldsymbol{w}} \leq \sqrt{2d}||\boldsymbol{w}K||_2||\boldsymbol{w}||_2, \tag{5.2}$$

from which (ii) follows since $||\boldsymbol{w}K||_2^2 = \langle \boldsymbol{w}K, \boldsymbol{w}K\rangle = \boldsymbol{w}KK^T\boldsymbol{w}^T = \boldsymbol{w}L\boldsymbol{w}^T$. Note that the first inequality somehow relates $\boldsymbol{w}$ to the optimal cut in G.

Let us first prove the upper bound in (5.2). Using the Cauchy-Schwartz inequality we see that

$$B_{\boldsymbol{w}} = \sum_{(x,y)\in E} w_x^2 - w_y^2 = \sum_{(x,y)\in E}(w_x+w_y)(w_x-w_y) \leq \sqrt{\sum_{(x,y)\in E}(w_x+w_y)^2}\sqrt{\sum_{(x,y)\in E}(w_x-w_y)^2}.$$

However the second factor can be evaluated as

$$\sqrt{\sum_{(x,y)\in E}(w_x-w_y)^2} = \sqrt{\sum_{e\in E}(\boldsymbol{w}K)_e^2} = ||\boldsymbol{w}K||_2$$

and the first factor can be bounded as

$$\sqrt{\sum_{(x,y)\in E}(w_x+w_y)^2} \leq \sqrt{\sum_{(x,y)\in E} 2w_x^2 + 2w_y^2} = \sqrt{2d\sum_{x\in V} w_x^2} = \sqrt{2d}||\boldsymbol{w}||_2$$

where we used that $2x^2+2y^2-(x+y)^2 = x^2-2xy+y^2 = (x-y)^2 \geq 0$ for all x,y. It follows that $B_{\boldsymbol{w}} \leq \sqrt{2d}||\boldsymbol{w}K||_2||\boldsymbol{w}||_2$.

Let us then prove the lower bound in (5.2). For ease of notation we relabel $V(G) = [n]$ so that $w_1 \geq w_2 \geq \ldots \geq w_n$, in particular so that each edge ij with $i < j$ is oriented as (i, j). Our plan will be to rewrite $B_{\boldsymbol{w}}$ in terms of the coordinates of $\boldsymbol{w}$ and the sizes of the cuts $E([i], [i]^c)$ for $i \in V^+$. Bounding the sizes of these cuts by the expansion ratio (which is possible since $|V^+| \leq \frac{n}{2}$)) will lead to the claimed bound.

$$\begin{aligned} B_{\boldsymbol{w}} &= \sum_{(x,y)\in E} w_x^2 - w_y^2 = \sum_{(x,y)\in E, x<y} \sum_{i=x}^{y-1} w_i^2 - w_{i+1}^2 \\ &= \sum_{i=1}^{n} (w_i^2 - w_{i+1}^2) e([i], [i]^c) = \sum_{i\in V^+} (w_i^2 - w_{i+1}^2) e([i], [i]^c) \\ &\geq \sum_{i\in V^+} (w_i^2 - w_{i+1}^2) hi = h \sum_{i\in V^+} w_i^2 = h||\boldsymbol{w}||_2^2. \end{aligned}$$

The second last equality above is obtained by collapsing a telescoping sum and noting that $w_{i+1} = 0$ if $i = |V^+|$. □

5.6 Typical vertex-expansion

In this section we'll investigate the vertex expansion of small linear sets in a 'typical' (n, d, α)-graph. It will be convenient to consider the following, normalised, isoperimetric parameter

$$\Psi(G, k) = \min_{j\leq k} \frac{\Phi_V(G, j)}{j} = \min_{S\subseteq V, |S|\leq k} \frac{|\partial_V(S)|}{|S|}.$$

As a little toy example/warm up let us consider the same problem in the bipartite setting, and the slightly simpler parameter

$$\Psi^L(G, k) = \min_{S\subseteq L, |S|\leq k} \frac{|\partial_V(S)|}{|S|},$$

determining expansion of sets on the 'left side' of a bipartite graph. Note that, in the introduction we showed that existence of graphs where $\Psi^L(G, \frac{n}{10d}) \geq \frac{5d}{8}$. An obvious limit to the expansion of these small sets would be d, although in fact we can say a little more. Indeed, if $S \subseteq L$ is $L \cap K$ for some connected set of vertices in G then it is easy to see that $|\partial_V(S)| \leq sd + (s-1)(d-1)$, leading to the $\Psi^L(G, s) \leq d - 1 + \frac{1}{s}$.

Theorem 5.6. *For every $\delta > 0$ there exists an $\epsilon > 0$ such that almost every d-regular bipartite graph G with n vertices in each partition class*

$$\Psi^L(G, \epsilon n) \geq d - 1 - \delta.$$

We note that a similar lower bound also then holds for the corresponding edge-isoperimetric parameter.

We also note that it is a non-trivial question how we sample uniformly such a graph. In the bipartite case it is significantly easier, and this will motivate our techniques for the non-bipartite case. Later in the course we will also discuss a different way to sample (n, d)-graphs.

We will do so using the so-called *configuration model*. In the bipartite case, we start with two vertex sets L and R of size n and for each $x \in L$ and $y \in R$ we take a collection B_x and B_y consisting of d *half-edges*. We choose a matching M, uniformly at random from all possible matchings, of the points in $\bigcup_{x\in L} B_x$ and $\bigcup_{y\in R} B_y$, and we build a graph $G = G(M)$ on L and R where the number of edges from x to y is the number of half-edges in B_x matched to a half-edge in B_y in the matching M.

However, in some way we've just reduced the problem of sampling such graphs to the problem of sampling uniformly a matching M from a complete bipartite graph $K_{dn,dn}$. But, it turns out that this is much easier to do so and, whilst, unlike in the simpler binomial random graph model, the existence of edges are not independent from each other, it's relatively easy to get a handle on the probability that this matching contains certain substructures using simple counting techniques (as we'll see later).

One useful observation, which we will apply later in the more general configuration model as well, is that we can generate the matching 'sequentially': given any ordering $e_1, e_2, \dots$ of the half-edges in $\bigcup_{x\in X} B_x$ we can first choose the neighbour f_1 of e_1 uniformly at random from $\bigcup_{x\in Y} B_y$, then choose the neighbour f_2 of e_2 uniformly at random from $\bigcup_{x\in Y} B_y \setminus \{f_1\}$, and so on. Not only does this process also give a uniform distribution on the matching obtained, but if we stop at any point in the process, the remaining edges of the matching are uniformly distributed over all possible matchings of the remaining half-edges. We call this, or many similar observations like this, the *principle of deferred decisions*.

However, there is a slight technical difficulty with this model, even though M is uniformly distributed it isn't true that $G(M)$ is uniformly distributed (the probability of obtaining a graph depends on the number of paralell edges). These issues are normally easy to deal with in practise: for example for fixed d it can be shown that with positive probability $G(M)$ is simple, and conditioned on $G(M)$ being simple it is uniformly distributed over all d-regular bipartite graphs, and hence any event which holds whp in the configuration model also holds whp for a uniformly chosen simple d-regular bipartite graph. We will mostly ignore such issues in our presentation.

Proof of Theorem 5.6. We generate a random d-regular bipartite graph G according to the bipartite configuration model described above. Let us write $\eta = d - 1 - \delta$ for the expansion ratio which we wish to prove. For sets $S \subseteq L$ and $T \subseteq R$, let us write $X_{S,T}$ for the indicator random variable of the event that $\Gamma(S) \subseteq T$. Hence, we aim to show that whp $X_{S,T} = 0$ for all $s = |S| \leq \epsilon n$ and $|T| = \eta s$ (ignoring floor signs here for notational convenience). We note first that it suffices to prove this statment for sets $S \subseteq L$, since it will also hold for sets $S \subseteq R$ by symmetry and then if we have an arbitrary set $S \subset V(G)$ then $S_1 = S \cap L \subseteq L$ And $S_2 = S \cap R \subseteq R$ have disjoint neighbourhoods. Note furter that is suffices to prove it for $|T| = \eta s$ since if $\Gamma(S) \subseteq T'$ with $|T'| > \eta s$ we can just choose some arbitrary $T \supseteq T'$ of size $|T| = \eta s$, for which then $X_{S,T} = 1$.

However, for fixed S and T, using our sequential method of exposing the edges in G, we can calculate the probability that $\Gamma(S) \subseteq T$ by exposing first the half-edges emanating from vertices in S. Indeed, there are ds many such edges, and they all must be matched to a set of td half-edges, and hence

$$\mathbb{P}(X_{S,T} = 1) = \frac{td}{nd}\frac{td-1}{nd-1}\cdots\frac{td-sd}{nd-sd} = \frac{(td)_{sd}(nd-sd)!}{(nd)!} \leq \left(\frac{td}{nd}\right)^{sd}$$

where $(n)_k = n(n-1)\dots(n-k+1)$ is the falling factorial. It follows that

$$\mathbb{P}(\Psi(G,\epsilon n) < \eta) = 2\mathbb{P}\left(\sum_{S,T} X_{S,T} > 0\right) \leq 2\sum_{s=1}^{\epsilon n} \binom{n}{s}\binom{n}{t}\frac{(td)_{sd}(nd-sd)!}{(nd)!}.$$

Then, as usual, we're left to approximate this sum. Using our trusty inequality $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$

$$\begin{aligned}\mathbb{P}(\Psi(G,\epsilon n) < \eta) \leq \leq 2\sum_{s=1}^{\epsilon n}\left(\frac{en}{s}\right)^s\left(\frac{en}{\eta s}\right)^{\eta s}\left(\frac{\eta s}{n}\right)^{sd} \leq 2\sum_{s=1}^{\epsilon n}\left(\frac{en}{s}\left(\frac{en}{\eta s}\right)^{\eta}\left(\frac{\eta s}{n}\right)^{d}\right)^s \\ = 2\sum_{s=1}^{\epsilon n}\left(c(\delta,d)\left(\frac{s}{n}\right)^{\delta}\right)^s,\end{aligned}$$

where $c(\delta, d)$ is some constant which only depends on δ and d. Given a $\delta > 0$, it is easy to pick an ϵ such that

$$\left(\frac{s}{n}\right)^{\delta} \leq \epsilon^{\delta} < \frac{1}{10}$$

and so the tail of this series is shrinking like a geometric series (and hence is $o(1)$. However, for small s (say $s \leq \log n$) the sum

$$\sum_{s=1}^{\log n}\left(c(\delta,d)\left(\frac{s}{n}\right)^{\delta}\right)^s \leq \log n\frac{c(\delta,d)\log n}{n} = o(1).$$

□

With a bit more work we can prove a similar bound for arbitrary (n,d)-graphs. Now, as you might expect, our configuration model consists of choosing a set B_x of d half-edges for each vertex $x \in V$, choosing uniformly at random a matching M on the set $\bigcup_{x\in V} B_x$ and taking the graph $G = G(M)$ where the number of edges from x to y is the number of half-edges in B_x matched to half-edges in B_y (and the number of loops at x is the number of matching edges fully contained in B_x).

Note that as before, we can also expose this matching sequentially according to some ordering $e_1, e_2, \dots$ of the half-edges in $\bigcup_{x\in V} B_x$. However, now we need to make the obvious concession that if at each stage we only choose a partner for a half-edge e_j if it has not already been paired to an earlier half-edge e_i in the process. Again, it is easy to see that the matching generated in this way is uniformly distributed, and that if we stop at any point in the process the remaining edges of the matching are uniformly distributed over all matchings of the remaining half-edges.

Again, there are some echnical difficulties arising from the fact that the distrbituion of $G(M)$ is not uniform (the probability of obtaining a graph depends on the number of loops and parallel edges), but they can be dealt with in a similar fashion.

Theorem 5.7. *For every $\delta > 0$ there exists an $\epsilon > 0$ such that almost every (n,d)-graph G*

$$\Psi(G,\epsilon n) \geq d - 2 - \delta.$$

Similarly we can see that this bound is optimal by considering any connected set with s vertices, which shows that $\Psi(G,s) \leq d - 2 + \frac{4}{s}$.

Proof of Theorem 5.7. The first, and easier part of the proof will be to demonstrate that whp G has good edge expansion. Namely, using the expression $\partial_E(S) = d|S| - 2e(S)$, which holds for any d-regular graph, we see that it is sufficient to bound

$$e(S) \leq (1 + \frac{\delta}{2})|S| \text{ for all } |S| \leq \epsilon n, \tag{5.3}$$

from which it will follow that $\partial_E(S) \geq (d - 2 - \delta)|S|$ for all such S. We will then show that

$$|\partial_V(S)| + e(S) \geq (d - 1 - \frac{\delta}{2})|S| \text{ for all } |S| \leq \epsilon n, \tag{5.4}$$

from which it will follow that $\Psi(G, \epsilon n) = \min_{S \subseteq V, |S| \leq \epsilon n} \frac{|\partial_V(S)|}{|S|} \geq (d - 2 - \delta)$.

So, let us begin by proving (5.3). As before, let us define random variables $Y_{S,K}$ for each set $|S| \leq \epsilon n$ and K is a set of half-edges emanating from vertices in S of size k. Then $Y_{S,K} = 1$ if and only if all the matching M induces a matching on K (i.e, the half-edges in K are all matched amongst themselves). Note that, if $|E(S)| \geq k$, then $Y_{S,K} = 1$ for some set K of size k.

However, we can reasonably easily evaluate $\mathbb{P}(Y_{S,K})$. Indeed, the first step is to note that the number of perfect matchings on a set of size ℓ can be expressed as $\ell!! := (\ell - 1)(\ell - 3) \dots 1$. It follows that

$$\mathbb{P}(Y_{S,K}) = \frac{M(k)M(dn - k)}{M(dn)} = \frac{k!!(dn - k)!!}{dn!!} = \frac{k - 1}{nd - 1}\frac{k - 3}{nd - 3} \cdots \frac{1}{nd - k - 1} \leq \left(\frac{k}{nd}\right)^{\frac{k}{2}},$$

and so the probability that there is some subset S which doesn't satisfy (5.3) is at most

$$\begin{aligned}
\mathbb{P}\left(\sum_{S,K} Y_{S,K} > 0\right) &\leq \sum_{s=1}^{\epsilon n} \sum_{k=2(1+\frac{\delta}{2})s}^{sd} \binom{n}{s}\binom{ds}{k}\frac{k!!(dn - k)!!}{dn!!} \\
&\leq \sum_{s=1}^{\epsilon n} \sum_{k=2(1+\frac{\delta}{2})s}^{sd} \left(\frac{en}{s}\right)^s \left(\frac{eds}{k}\right)^k \left(\frac{k}{nd}\right)^{\frac{k}{2}} \\
&\leq \sum_{s=1}^{\epsilon n} \sum_{k=2(1+\frac{\delta}{2})s}^{sd} \left(\frac{en}{s}\right)^s \left(\frac{ed}{2(1 + \frac{\delta}{2})}\right)^{sd} \left(\frac{s}{n}\right)^{\frac{k}{2}} \\
&\leq \sum_{s=1}^{\epsilon n} \sum_{k=2(1+\frac{\delta}{2})s}^{sd} c(\delta, d)^s \left(\frac{s}{n}\right)^{\frac{k}{2} - s} \\
&\leq \sum_{s=1}^{\epsilon n} sd \left(c(\delta, d)\left(\frac{s}{n}\right)^{\frac{\delta}{2}}\right)^s \\
&= o(1),
\end{aligned}$$

as long as ϵ is sufficiently small, say so that $c(\delta, d)\epsilon^{\frac{\delta}{2}} < \frac{1}{2}$. Again we have to deal with the terms for small s separately, but it can easily be seen that the sum over say $s \leq \log n$ is $o(1)$. Hence (5.3) holds whp as claimed.

The proof of (5.4) follows along similar lines, but involves a more careful calculation. Let us define random variables $Z_{S,R,K}$ for every pair of disjoint vertex sets S and R of size s and r, respectively and subsets K of cardinality k of the ds half-edges emanating from vertices of

S. The $Z_{S,R,K}$ is the indicator function of the event that M induces a matching on K, and the other $ds-k$ half-edges emanating from S are matched with half-edges emanating from R. Hence we are interested in whether $Z_{S,R,K}=1$ for some triple (S,R,K) with $0<s\leq \epsilon n$ and $r+\frac{k}{2}=\left(d-1-\frac{\delta}{2}\right)s=\eta s$.

As before, given (S,R,K), we can calculate this probability explicitly

$$\begin{aligned}\mathbb{P}(Z_{S,R,K}=1)&=\frac{k!!(rd)_{sd-k}(nd-2sd+k)!!}{(nd)!!}\\&=\frac{k!!rd(rd-1)(rd-2)\dots(rd-sd+k+1)}{(nd-1)(nd-3)\dots(nd-2sd+k+1)}\\&\leq\left(\frac{ds}{n-2s}\right)^{sd-\frac{k}{2}},\end{aligned}$$

where we used that each of the $ds-\frac{k}{2}$ factors in the numerator are at most $\max\{k,rd\}\leq d^2s$ and each of the $ds-\frac{k}{2}$ factors in the denominator are at least $nd-2sd$. Hence we can bound

$$\begin{aligned}\mathbb{P}\left(\sum Z_{S,R,K}\right)&\leq\sum_{s=1}^{\epsilon n}\sum_{r+\frac{k}{2}=\eta s}\binom{n}{s}\binom{n-s}{r}\binom{ds}{k}\frac{k!!(rd)_{sd-k}(nd-2sd+k)!!}{(nd)!!}\\&\leq\sum_{s=1}^{\epsilon n}\sum_{r+\frac{k}{2}=\eta s}\left(\frac{en}{s}\right)^s\left(\frac{en}{r}\right)^r\left(\frac{eds}{k}\right)^k\left(\frac{ds}{n-2s}\right)^{sd-\frac{k}{2}}\\&\leq\sum_{s=1}^{\epsilon n}\sum_{r+\frac{k}{2}=\eta s}\left(\frac{en}{s}\right)^s\left(\frac{en}{s}\right)^r\left(\frac{s}{r}\right)^r\left(\frac{s}{k}\right)^k(ed)^k\left(\frac{ds}{n-2s}\right)^{sd-\frac{k}{2}}.\end{aligned}$$

Now, since $\left(\frac{s}{r}\right)^{\frac{r}{s}}$ and $\left(\frac{s}{k}\right)^{\frac{k}{s}}$ are both bounded above by a constant, and $n-2s\geq\frac{n}{2}$ we can simplify this

$$\begin{aligned}\mathbb{P}\left(\sum Z_{S,R,K}\right)&\leq\sum_{s=1}^{\epsilon n}\sum_{r+\frac{k}{2}=\eta s}c(d,\delta)^s\left(\frac{s}{n}\right)^{sd-s-r-\frac{k}{2}}\\&\leq\sum_{s=1}^{\epsilon n}sd\left(c(d,\delta)\left(\frac{s}{n}\right)^{\frac{\delta}{2}}\right)^s\\&=o(1),\end{aligned}$$

again for $\epsilon>0$ sufficiently small in terms of d and δ. □

6 Extremal problems on spectrum and expansion

In this section we will consider various *extremal problems* to do with the expansion and spectrum of graphs. In particular we are interested in which (n,d)-graphs maximise the edge- and vertex-expansion ratios (and more generally for these ratios restricted to sets of a fixed size), which (n,d)-graphs maximise the spectral gap, and how large these parameters are on these maximisers.

A useful example to keep in mind here will be the infinite d-regular tree T_d. It will turn out that this is in some sense the 'best' example for both these problems, and not only will this give us a good indication of what sort of bounds to prove in the general case, it will also help inform our proofs.

6.1 The d-regular tree

6.1.1 The expansion of T_d

Consider the edge-expansion function $\Phi_E(T_d,k)$. Any minimising set S is clearly connected, and hence a subtree, and so $e(S)=|S|-1$ and hence

$$\Phi_E(T_d,k)=kd-(2k-1)=k(d-2)+2$$

and hence the expansion ratio satisfies

$$h(T_d)=\inf_{|S| \text{ finite}} \frac{e(S,S^c)}{|S|}=d-2.$$

Note that the above argument also implies that $\Phi_E(G,k)\leq k(d-2)+2$ for every (n,d)-graph and every k and so the relative expansion of a set of size k cannot exceed $d-2+\frac{2}{k}$. However, for finite G the expansion ratio must in fact be significantly smaller than this. To see this, consider a random subset $S\subseteq V(G)$ of size $\frac{n}{2}$. The expected size of $\partial_E(S)$ is $\frac{(1+o(1))d|S|}{2}$, since each edge of G belongs to $\partial_E(S)$ with probability $\frac{1}{2}+o(1)$. Hence there exists some set S of size $\frac{n}{2}$ with $\frac{\partial_E(S)}{|S|}\leq \frac{d}{2}+o(1)$, and so $h(G)\leq \frac{d}{2}+o(1)$. In fact, a more refined analysis will show that $h(G)\leq \frac{d}{2}-c\sqrt{d}$ for some absolute constant c for any (n,d)-graph G with $d\geq 3$ and n sufficiently large. As we will see later, there are (n,d,α)-graphs with $\alpha=O\left(\frac{1}{\sqrt{d}}\right)$, and so this result is tight up to the value of c (which follows from the Cheeger bound).

6.1.2 The spectrum of T_d

More interesting is the spectrum of T_d. Here we will have to treat the (infinite) adjacency matrix $A(T_d)$ of T_d as a linear operator on $\ell_2(V(T_d))$, the set of square summable real functions on $V(T_d)$, and look at the spectrum of this operator.

For such operators it is normal to define the spectrum as $\text{spec}(A)=\{\lambda\colon (A-\lambda I) \text{ is not invertible}\}$. This means that λ can be in spectrum for two reason - either $(A-\lambda I)$ has a non-trivial kernel,

or it is not onto. For finite matrices these two conditions are equivalent, and we can determine whether λ is in the spectrum by looking for an eigenvector with eigenvalue λ. In contrast $(A - \lambda I)$ has no eigenvectors, and it's entire spectrum comes from the second reason. In this context there is also an analogue of the distribution and multiplicity of the eigenvalues, coming from the *spectral measure*, which we will mention briefly later.

Theorem 6.1 (Carter)**.** *The spectrum of $A := A(T_d)$ is given by*

$$spec(A) = [-2\sqrt{d-1}, 2\sqrt{d-1}].$$

Sketch of proof. We follow the approach of Friedman. We fix some arbitrary vertex v as the root of the tree. Then it turns out that

$$\lambda \in \text{spec}(A) \iff \boldsymbol{\delta}(v) \notin \text{Range}(A - \lambda I),$$

where $\boldsymbol{\delta}(v)$ is the characteristic function of $\{v\}$. The necessity of this condition is clear. The sufficiency is not hard to show, but we will not prove it. So, we wish to work out when there exists some function $\boldsymbol{f} \in \ell_2(V)$ satisfying

$$\boldsymbol{\delta}(v) = (A - \lambda I)\boldsymbol{f}. \tag{6.1}$$

We say that a function $\boldsymbol{f}$ on V is *spherical around* $v \in V$ if f_u depends only on the distance (in the graph metric) from u to v. The *spherical symmetrisation* of $\boldsymbol{g} \in \ell_2(V)$ around v is a spherical function $\boldsymbol{f}$ such that

$$\sum_{d(u,v)=r} f_u = \sum_{d(u,v)=r} g_u$$

for every $r \geq 0$. It is easy to see that if $\boldsymbol{g} \in \ell_2(V)$ is a solution to (6.1), then $\boldsymbol{f}$ is in $\ell_2(V)$ as well and also satisfies (6.1).

Hence we may assume wlog that $\boldsymbol{f}$ is spherical, and hence determined by some sequence $x_0, x_1, \ldots$ such that $f(u) = x_i$ whenever $d(u, v) = i$. This reduces (6.1) to the following recurrence relation

$$\lambda x_0 = dx_1 + 1 \text{ and } \lambda x_i = x_{i-1} + (d-1)x_{i+1} \text{ for } i \geq 1. \tag{6.2}$$

The general solution to such a recurrence is given by $x_i = \alpha\rho_1^i + \beta\rho_2^i$ where

$$\rho_{1,2} = \frac{\lambda \mp \sqrt{\lambda^2 - 4(d-1)}}{2(d-1)}$$

are the roots of the quadratic equation $\lambda\rho = 1 + (d-1)\rho^2$.

Then, if the discriminant is negative, i.e., if $|\lambda| \leq 2\sqrt{d-1}$, then the roots are complex with absolute value

$$\sqrt{\frac{\lambda^2 + 4(d-1) - \lambda^2}{4(d-1)^2}} = \frac{1}{\sqrt{d-1}}.$$

However, in this case $\boldsymbol{f} \notin \ell_2$. Indeed, $|x_i| = \Theta\left((d-1)^{-\frac{i}{2}}\right)$ and then, since there are $\Theta\left((d-1)^i\right)$ vertices at distance i from v, it follows that $||\boldsymbol{f}||_2 = \infty$. Hence, no such function $\boldsymbol{f} \in \ell_2$ exists, and so $\lambda \in \text{spec}(A)$. A similar observation holds for $\lambda = 2\sqrt{d-1}$.

Conversely, we claim that when $|\lambda| > 2\sqrt{d-1}$ then (6.2) has a solution in ℓ_2, implying that λ is not in the spectrum of A. To see this, we observe that

$$2(d-1)\frac{d}{d\lambda}\rho_1 = 1 - \frac{\lambda}{\sqrt{\lambda^2 - 4(d-1)}} < 0$$

for $|\lambda| > 2\sqrt{d-1}$, and so

$$\rho_1 \geq \frac{2\sqrt{d-1} - \sqrt{4(d-1) - 4(d-1)}}{2(d-1)} = \frac{1}{\sqrt{d-1}}.$$

So if we take $x_i = \alpha\rho_1^i$ for some any α then (6.2) holds for all $i \geq 1$, and $\boldsymbol{f} \in \ell_2$. It remains to choose α such that the first condition also holds, which says

$$\lambda\alpha = d\alpha\rho_1 + 1,$$

which has a solution iff $\lambda \neq d\rho_1$. However, $|\rho_1| < \frac{|\lambda|}{2(d-1)} \leq \frac{|\lambda|}{d}$ for all $d \geq 2$.

□

6.2 The Alon-Boppana lower bound

In this section we return to the question of how small λ_2 can be in a (large) d-regular graph. In Lemma 3.5 we gave a weak bound of $(1 - o(1))\sqrt{d}$, and now we will prove the stronger Alon-Boppana lower bound.

Theorem 6.2 (Alon-Boppana, Nilli, Friedman)**.** *There exists a constant c such that every (n,d)-graph G with diameter Δ satisfies*

$$\lambda_2(G) \geq 2\sqrt{d-1}\left(1 - \frac{c}{\Delta^2}\right).$$

Note that, for fixed d and n large, the diameter of G must be growing as a function of n. More explicitly, since the diameter of an (n,d)-graph is $\Omega(\log_{d-1} n)$ we obtain the following corollary.

Corollary 6.3. *For every (n,d)-graph G*

$$\lambda_2(G) \geq 2\sqrt{d-1}\left(1 - O\left(\frac{1}{\log^2 n}\right)\right).$$

We will present two different proofs of this theorem. The first is again via the probabilistic method, and in particular the *first moment method*, but will lead to a slightly weaker conclusion: only bounding $\lambda(G)$ rather than λ_2 and with a slightly weaker error term.

Proof I : Counting walks in T_d. Let A be the adjacency matrix of G. Clearly $\lambda(A)^{2k} = \lambda\left(A^{2k}\right)$ for every integer k, where $\lambda(A) = \max_{i\geq 2}|\lambda_i|$. We will give a lower bound on $\lambda\left(A^{2k}\right)$ by estimating the Rayleigh quotient of a judiciously chosen vector $\boldsymbol{f} = \boldsymbol{\delta}(s) - \boldsymbol{\delta}(t)$, where s and t are two vertices at distance Δ in G. That is, $f_s = -f_t = 1$ and $f_u = 0$ for all other u. Note that $\boldsymbol{f} \perp \boldsymbol{u}$, and hence

$$\lambda^{2k} \geq \frac{\boldsymbol{f}A^{2k}\boldsymbol{f}^T}{||\boldsymbol{f}||_2^2} = \frac{\left(A^{2k}\right)_{ss} + \left(A^{2k}\right)_{tt} - 2\left(A^{2k}\right)_{st}}{2}.$$

However, if we choose $k = \frac{\Delta-1}{2}$, so that there are no paths of length $2k$ from s to t, then the negative term in the numerator vanishes. Now, the positive terms in the numerator are counting the number of closed walks of length $2k$ which start and end at s or t, respectively.

However, we note that these must be at least the number of closed walks of length $2k$ which start and end at the root v of T_d. Indeed, it is easy to display a natural injection from walks in T_d rooted at v to walks in G rooted at a fixed vertex x. It follows that

$$\lambda \geq (t_{2k})^{\frac{1}{2k}},$$

where t_{2k} counts the number of closed walks of length $2k$ starting and ending at the root in T_d.

However the numbers t_{2k} have been studied in great detail. In fact, very good estimates, as well as a recursive definition and their generating functions are known. However all we will need is a very rough estimate (we will give a slightly more precise one later in the course). We can associate with each walk in T_d a *sign pattern*, a sequence in $\{+1, -1\}^{2k}$ where a step away from the root gives a $+1$ and a step towards the root gives a -1.

Clearly the sign pattern associated to the walk must satisfy the following two properties:

- It sums up to zero;
- Each partial sum is non-negative.

We call such a sign pattern *admissable*.

It is well known that the number of admissable sign pattterns of length $2k$ is given by the kth *Catalan number*

$$C_k = \frac{\binom{2k}{k}}{k+1}.$$

A simple way to see this is to associate every sign pattern with a*montonic path* in the grid, each $+1$ corresponding to a step to the right and each -1 to a vertical step. The path corresponding to an admissable sign pattern must end at (k, k) and never go strictly above the main diagonal $\{(x, x) \colon x \leq k\}$.

Now, there are $\binom{2k}{k}$ possible sign patterns of length k, and for every inadmissable sign pattern γ we can define a *reflected* sign pattern of length k γ' by flipping the pattern after the first point it goes strictly above the main diagonal. Clearly the walk corresponding to γ' ends at $(k-1, k+1)$. Furthermore, each walk starting at $(0, 0)$ and ending at $(k-1, k+1)$ must go strictly above the main diagonal, and this reflection process is reversible. Hence the number of inadmissable paths is equal to the number of monotonic paths ending at $(k-1, k+1)$, which is equal to $\binom{2k}{k-1}$. It follows that the number of admissable paths in equal to

$$\binom{2k}{k} - \binom{2k}{k-1} = \left(1 - \frac{k}{k+1}\right)\binom{2k}{k} = \frac{1}{k+1}\binom{2k}{k}.$$

Now, given an admissable sign pattern we claim there are at least $(d-1)^k$ different walks with this sign pattern, since for each of the k occurences of $+1$ we can choose the next step 'away'

from the root. It follows, using the consequence of Stirling's approximation that $\binom{2k}{k} \approx \frac{2^{2k}}{\sqrt{\pi k}}$, that

$$t_{2k} \geq C_k(d-1)^k = \Theta\left(\left(2\sqrt{d-1}\right)^{2k} k^{-\frac{3}{2}}\right)$$

and so, recalling that $k = \frac{\Delta-1}{2}$

$$\lambda \geq 2\sqrt{d-1}\left(\frac{\Delta-1}{2}\right)^{-\frac{6}{4(\Delta-1)}} = 2\sqrt{d-1}\left(1 - O\left(\frac{\log\Delta}{\Delta}\right)\right),$$

using the fact that

$$x^{-\frac{1}{y}} = e^{-\frac{\log x}{y}} = 1 - O\left(\frac{\log x}{y}\right),$$

as long as $x \ll e^y$. □

We can however improve this proof by making a more judicious choice for the function $\boldsymbol{f}$, which will come from a truncated eigenfunction of T_d.

Proof II: Using spherical functions. We follow here an argument of Friedman. Again our hope is to bound λ_2 in terms of the Rayleigh quotient, observing that

$$\lambda_2 = \min_{\boldsymbol{f}\perp\boldsymbol{u}} \frac{\boldsymbol{f}A\boldsymbol{f}^T}{||\boldsymbol{f}||_2^2},$$

using a clever choice of a test function $\boldsymbol{f}$. The idea is to try to use something close to an eigenfunction for a truncation of T_d.

Given two vertices s and t at distance Δ, we will construct a function $\boldsymbol{f}$ which is positive on vertices at distance $\leq k = \frac{\Delta}{2} - 1$ from s, negative on vertices at distance $\leq k$ from t and zero elsewhere, where the values which $\boldsymbol{f}$ takes are derived from the eigenfunction $\boldsymbol{g}$ with maximal eigenvalue for the d-regular tree of height k.

Let $S_i = \{v\colon d(s,v) = i\}$ and $T_i = \{v\colon d(t,v) = i\}$, noting that these sets are all disjoint, and let $Q = V(G) \setminus \bigcup_{0\leq i\leq k}(S_i \cup T_i)$ be the rest of the vertices of G. Note further that there are no edges between any S_i and T_j. Let us write $T_{d,k}$ for the d-regular tree of height k and let $A(T_{d,k})$ be its adjacency matrix. We use the following claim

Claim 6.4. Let μ be the largest eigenvalue of $A(T_{d,k})$. Then there is a unique eigenvector $\boldsymbol{g}$ with eigenvalue μ and $\boldsymbol{g}$ is non-negative and spherically symmetric.

This can be proved directly, or by appealing to the Perron-Frobenius theorem, and non-negativity and spherical symmetry can be verified as in Theorem 6.1. Let g_i be the value that $\boldsymbol{g}$ takes on the vertices on the ith level of $T_{d,k}$. It is clear that the g_i satisfy the following recursive relation and boundary conditions

$$\begin{aligned} \mu g_0 &= d g_1; \\ \mu g_i &= g_{i-1} + (d-1)g_{i+1} \text{ for } i = 1,\ldots,k; \\ g_{k+1} &= 0. \end{aligned} \tag{6.3}$$

We will need to know a little about the vector $\boldsymbol{g}$ and the value of μ. We can in fact determine the explicit solutions to the recursion (6.3). Let

$$g_i = (d-1)^{-\frac{i}{2}} \sin((k+1-i)\theta).$$

We claim that, for an appropriate choice of θ, $\boldsymbol{g}$ satisfies (6.3) with $\mu = 2\sqrt{d-1}\cos\theta$.

Indeed, for $0 < i \leq k$ we have

$$\begin{aligned} g_{i-1} + (d-1)g_{i+1} &= (d-1)^{-\frac{i-1}{2}} \big(\sin((k+2-i)\theta) + \sin((k-i)\theta)\big) \\ &= \sqrt{d-1}(d-1)^{-\frac{i}{2}} 2\sin((k+1-i)\theta)\cos\theta = \mu g_i, \end{aligned}$$

using that

$$\sin\alpha + \sin\beta = 2\sin\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2}.$$

The condition for $i = 0$ is then $\mu g_0 = d g_1$, that is,

$$2\sqrt{d-1}\cos\theta\sin((k+1)\theta) = \frac{d}{\sqrt{d-1}}\sin(k\theta).$$

or equivalently

$$h(\theta) = 2(d-1)\cos\theta\sin((k+1)\theta) - d\sin(k\theta) = 0. \tag{6.4}$$

Since $g_{k+1} = 0$, $\boldsymbol{g}$ gives a solution for any root θ of h. The largest possible value of μ will then come from the smallest positive root θ_0 of h, which satisfies $0 < \theta_0 < \frac{\pi}{k+1}$, since for very small values of θ h is positive, and at $\frac{\pi}{k+1}$ it is negative. Furthermore, it can be checked that g_i is non-negative and nonincreasing if $0 < \theta < \frac{\pi}{k+1}$. So, if we set $\theta = \theta_0$ then (6.3) is satisfied and $\boldsymbol{g} \geq 0$ and is decreasing, and so by Claim 6.4 is indeed the unique eigenvector of $A(T_{d,k})$ of largest eigenvalue.

Finally, to bound μ, we note that $\theta_0 < \frac{\pi}{k+1} \approx \frac{2\pi}{\Delta}$ and hence, using the Taylor expansion of cos, we see that

$$\cos\theta_0 > 1 - \frac{\theta_0^2}{2} \geq 1 - \frac{2\pi^2}{\Delta^2},$$

and so $\mu \geq 2\sqrt{d-1}\left(1 - \frac{2\pi^2}{\Delta^2}\right)$.

So, let us $\boldsymbol{f} \in \mathbb{R}^V$ as follows:

$$f_v = \begin{cases} c_1 g_i & v \in S_i, \\ -c_2 g_i & v \in T_i, \\ 0 & \text{else.} \end{cases}$$

where we choose c_1 and c_2 so that $\boldsymbol{f} \perp \boldsymbol{u}$. We first claim that

$$(A\boldsymbol{f})_v \geq \mu f_v \text{ for } v \in \bigcup S_i \qquad \text{and} \qquad (A\boldsymbol{f})_v \leq \mu f_v \text{ for } v \in \bigcup T_i.$$

Indeed, let $v \in S_i$ for some $i > 0$. Then of v's d neighbours some $p \geq 1$ belong to S_{i-1}, q belong to S_i and $d-p-q$ below to S_{i+1} and hence

$$(A\boldsymbol{f})_v = pc_1 g_{i-1} + qc_1 g_i + (d-p-q)c_1 g_{i+1} \geq c_1(g_{i-1} + (d-1)g_{i+1}),$$

since $\boldsymbol{g}$ is non-negative and nonincreasing. However, since $\boldsymbol{g}$ is an eigenvector of $A(T_{d,k})$ with eigenvalue μ,

$$(A\boldsymbol{f})_v \geq c_1(g_{i-1} + (d-1)g_{i+1}) = c_1(A(T_{d,k})\boldsymbol{g})_i = c_1\mu g_i = \mu f_v.$$

A similar argument works in the case that $v = s$ or that $v \in \bigcup T_i$.

Hence we can argue as follows

$$\begin{aligned}\boldsymbol{f}A\boldsymbol{f}^T = \sum_v f_v(A\boldsymbol{f})_v &= \sum_{v\in\bigcup S_i} f_v(A\boldsymbol{f})_v + \sum_{v\in\bigcup T_i} f_v(A\boldsymbol{f})_v + \sum_{v\in Q} f_v(A\boldsymbol{f})_v \\ &\geq \sum_{v\in\bigcup S_i} f_v\mu f_v + \sum_{v\in\bigcup T_i} f_v\mu f_v = \mu||\boldsymbol{f}||_2^2.\end{aligned}$$

However, since by our choice of c_1 and c_2 we have that $\boldsymbol{f} \perp \boldsymbol{u}$, it follows that $\lambda_1 \geq \mu$ as claimed. □

6.2.1 Extensions of the Alon-Boppana theorem

A quanititative variantiion of Theorem 6.2 states that in fact a constant fraction of the n eigenvalues of any (n,d)-graph must be almost as large as $2\sqrt{d-1}$.

Theorem 6.5 (Serre)**.** *For every $d \in \mathbb{N}$ and $\epsilon > 0$ there exists $c = c(\epsilon, d)$ such that every (n,d)-graph has at least cn eigenvalues greater than $2\sqrt{d-1} - \epsilon$.*

There are several proofs of the above theorem, and the best known bound for the constant c in the above theorem is $c \approx (d-1)^{-\pi\sqrt{\frac{2}{\epsilon}}}$. We give a short and elegant proof of Ciobă.

Proof. Let $A = A(G)$, and consider the matrix $(A + dI)^k$, where we will choose k later. Let n_ϵ be the number of eignevalues of A larger than $2\sqrt{d-1} - \epsilon$. On the one hand we can bound

$$\text{trace}(A + dI)^k = \sum_{i=1}^{n}(\lambda_i + d)^k \leq (2d)^k n_\epsilon + (d + 2\sqrt{d-1} - \epsilon)^k n, \tag{6.5}$$

and on the other hand.

$$\text{trace}(A + dI)^k = \sum_{j=0}^{k}\binom{k}{j}\text{trace}(A^j)d^{k-j} \geq \sum_{\ell=0}^{\frac{k}{2}}\binom{k}{2\ell}nt_{2\ell}d^{k-2\ell},$$

where $t_{2\ell}$ are the tree-numbers from the first proof of Theorem 6.2, and we've eliminated the (positive) terms for odd j. Using the estimates we previously gave on these numbers we can conclude that

$$\begin{aligned}\text{trace}(A + dI)^k &\geq \left(\frac{c'}{k^{\frac{3}{2}}}\right) n\sum_{\ell=0}^{\frac{k}{2}}\binom{k}{2\ell}(2\sqrt{d-1})^{2\ell}d^{k-2\ell} \\ &= \left(\frac{c'}{k^{\frac{3}{2}}}\right) n\big((d + 2\sqrt{d-1})^k + (d - 2\sqrt{d-1})^k\big) \\ &\geq \left(\frac{c'}{k^{\frac{3}{2}}}\right) n(d + 2\sqrt{d-1})^k,\end{aligned}$$

for some constant $c' > 0$. Together with (6.5) we conclude that

$$\frac{n_\epsilon}{n} \geq \frac{\left(\frac{c'}{k^{\frac{3}{2}}}\right)(d+2\sqrt{d-1})^k - (d+2\sqrt{d-1}-\epsilon)^k}{(2d)^k}.$$

Some analysis shows that this expression is positive for $k = \Omega\left(\frac{d}{\epsilon}\log\left(\frac{d}{\epsilon}\right)\right)$ (for a large enough leading constant). □

Determining the optimal parameter $c(\epsilon, d)$ for which Theorem 6.5 holds in general is an open problem.

Much less is known about the spectrum of irregular graphs. Whilst it is still true that the largest eigenvalue satisfies $d \leq \lambda_1 \leq \Delta$ where d is the average degree and Δ is the maximum degree, it is not necessarily true that the second eigenvalue is large as a function of d. Indeed, consider the *lollipop graph* L_n, which is formed by identifying some vertex of a clique K_n with a path of length n. It can be checked that the average degree of L_n is still $\theta(n)$, but its generalised second eigenvalue is small, $\lambda(L_n) \leq 2$. However, if the average degree of the graph is robust under 'local' changes, it can be shown that a similar bound as in the Alon-Boppana theorem holds even in irregular graphs.

Theorem 6.6 (Hoory)**.** *Let $d, r \geq 2$ be integers. Suppose that G is a graph whose average degree is at least d whenever a ball of radius r is deleted from G. Then*

$$\lambda(G) \geq 2\sqrt{d-1}\left(1 - \frac{c\log r}{r}\right),$$

where c is some absolute constant.

6.3 Ramanujan graphs

In light of the Alon-Boppana bound, it is interesting to know which graphs achieve the largest possible spectral graph. We say an (n, d)-graph G is *Ramanujan* if $\lambda(G) \leq 2\sqrt{d-1}$.

It is not even clear immediately that such graphs exists, why indeed should Theorem 6.2 be tight? However, it was a major result of Lubotzky, Philips, and Sarnak (who coined the term Ramanujan graphs) and independently of Margulis that arbitrarily large d-regular Ramanujan graphs exist when $d-1$ is a prime, and this was extended to prime powers by Morgenstern.

Theorem 6.7 (Lubotzky-Philips-Sarnak, Margulis, Morgenstern)**.** *For every prime p and every $k \in \mathbb{N}$ there exist infinitely many d-regular Ramanujan graphs with $d = p^k + 1$.*

We won't give a proof of this theorem, but will briefly describe the construction of Lubotzky, Philips, and Sarnak. Let p, q be distinct primes which are both congruent to 1 mod 4. We will define a $(\theta(q^3), p+1)$-graph $X^{p,q}$.

This graph will be built as a *Cayley graph*. Given a group G and a subset $S \subset G$ which is closed under inversion (i.e., writing the group operation as muliplication, $S = S^{-1}$) the Cayley graph $G(G, S)$ is a graph with vertex set S and edge set

$$E(C(G,S)) = \{(x, xs) \colon x \in G, s \in S\}.$$

For our purpose we take $G = PGL(2, q)$, which is the quotient of the group of 2×2 non-singular matrices over $\mathbb{F}_q$ by the scalar matrices. We fix some integer i with $i^2 = -1 \bmod q$ and define

$$S = \left\{ \begin{pmatrix} a_0 + ia_1 & a_2 + ia_3 \\ -a_2 + ia_3 & a_0 - ia_i \end{pmatrix} : a_0^1 + a_1^2 + a_2^2 + a_3^2 = p, \text{ with } a_0 > 0 \text{ odd and } a_1, a_2, a_3 \text{ even} \right\}$$

It follows from a theorem of Jacobi that there are exactly $p+1$ solutions to $a_0^1 + a_1^2 + a_2^2 + a_3^2 = p$, and so $|S| = p + 1$, and it can be verified that S is closed under inversion as needed.

The graph $X^{p,q}$ is then the connected component of e in $C(G, S)$ (it can be shown that $C(G, S)$ is either connected, or has two identical components depending on the quadratic residue $\left(\frac{p}{q}\right)$. In both cases it can be shown that the second largest eigenvalue is then bounded as in Theorem 6.7.

A very natural question to ask is whether this spectral gap can be obtained in graphs of arbitrary size.

Question 6.8. *Do there exist arbitrary large d-regular Ramanujan graphs for all $d \geq 3$?*

7 The spectrum of random graphs

In this section we return to the question of determing *typical* properties of (n, d)-graphs, in particular the typical spectrum of such a graph. A powerful way to approach such questions is by using the *probabilistic method* – if we have a method to sample an (n, d)-graph uniformly at random then we can consider the graph paramaters of this random graph as random variables, whoch we can investigate using probabilistic tools. Any statement which holds with high probability for a random (n, d)-graph must be true for the vast majority of (n, d)-graphs.

7.1 The bulk of the spectrum

Whilst we are most interested in the extreme eigenvalues λ_2 and λ_n, it will be much easier to analyse the ' bulk' of the spectrum - most of the eigenvalues which lie 'in the middle', and so we will do this first.

The adjacency matrix of a random (n, d)-graph is a random symmetric matrix, but the entries are heavily dependent on each other. If we loosen this requirement, and consider random symmetric matrices where the entries (in the upper triangle) are independent from each other, then there is a famous theorem of Wigner, *Wigner's semicircle law*, which describes the distribution of the eigenvalues.

Theorem 7.1 (Wigner)**.** *Let A_n be an $n \times n$ symmetricwhere matrix, where the off-diagonal entries are distributed independently as F, and the diagonal entries independently as G. Furthermore, assume that $var(F) = var(G) = \sigma^2$ and that F and G have finite moments. Define the* empirical eigenvalue distribution *as*

$$W_n(x) = \frac{1}{n}|\{i\colon \lambda_i(A_n) \leq x\}|,$$

where $\lambda_1(A_n) \geq \ldots \geq \lambda_n(A_n)$ are the eigenvalues of A_n. Then for every x,

$$W(x) = \lim_{n\to\infty} W_n(2x\sigma\sqrt{n}) = \frac{2}{\pi}\int_{-1}^{x} \sqrt{1-z^2}dz.$$

Roughly, Theorem 7.1 says that, under some weak conditions on the distributions of the entries, the eigenvalues of a large random symmetric matrix are distributed 'close to a semicircle'.

If we randomly generate and plot the eigenvalues of a random (n, d)-graph then we find that, for small d, as n gets large, the distribution no longer tends to a semicircle, but rather more of a saddle shape, with two peaks towards the extremes of its support. However, as both d and n get very large, the distribution will again tend to a semicircle. In fact, this behaviour is not limited to just random graphs, it will hold whenver the graph has very few short cycles. Specifically, if we let $C_k(G)$ be the number of cycles of length k in G then the following theorem holds for whenever $C_k(G) = o(|G|)$ for every fixed $k \geq 3$, a property that can be seen to hold whp for random d-regular graphs.

Theorem 7.2 (McKay)**.** *Let G_n be an infinite sequence of d-regular graphs such that $C_k(G) = o(|G|)$ for all $k \geq 3$. Define the empircal eigenvalue distribution as*

$$F(G_n, x) = \frac{1}{|G_n|}|\{i\colon \lambda_i(G_n) \leq x\}|.$$

Then for every x,

$$F(x) = \lim_{n\to\infty} F(G_n, x) = \int_{-2\sqrt{d-1}}^{x} \frac{d\sqrt{4(d-1)-z^2}}{2\pi(d^2-z^2)}dz.$$

Note that the limit distribution $F(x)$ is supported on $[-2\sqrt{d-1}, 2\sqrt{d-1}]$, which is the spectrum of T_d, see Theorem 6.1. The main idea behind the proof is that, since there are only very few short cycles, the neighbourhood of most vertices is almost a tree. Hence, for most vertices v, the number of closed walks of length k rooted at v is roughly equal to t_k, the analogous quantity for T_d. To make this argument we require a good estimate for t_k, as opposed to the lower bound we used in Theorem 6.2.

Lemma 7.3. *For every $s \in \mathbb{N}$*

$$t_{2s-1} = 0 \qquad \text{and} \qquad t_{2s} = \sum_{j=1}^{s} \binom{2s-j}{s} \frac{j}{2s-j} d^j (d-1)^{s-j}.$$

Proof. The first claim is obvious since T_d is bipartite, and hence contains no odd length closed walks. for the second, consider any closed walk W of length $2s$ rooted at v. We can associate to W a sequence $0 = \delta_0, \delta_1, \ldots, \delta_{2s} = 0$, where δ_i is the distance from v at time i. Clearly, the δ_i are all non-negative integers and $|\delta_i - \delta_{i-1}| = 1$ for all i.

It can be seen that the number of such sequences in which exactly j of the terms δ_i are equal to 0 is given by

$$\binom{2s-j}{s} \frac{j}{2s-j}.$$

This is a simple generalization of the Catalan numbers.

Given such a sequence, how many walks W correspond to this sequence? Well, each walk takes s steps away from v and s steps towards v; namely, there are s indices with $\delta_i - \delta_{i-1} = 1$ or -1, respectively. In a step towards v, the next vertex is uniquely determined, whereas on steps aways from d we have d choices when $\delta_i = 0$ (and the walk is currently at v), and $d-1$ choices when $\delta_i \neq 0$. Since these happen j and $s-j$ times, respectively, the conclusion follows. □

Now, if $C_k(G_n) = o(|G_n|)$ for every fixed k, then for every k almost every vertex has a cycle-free k-neighbourhood. Therefore there are $(1+o(1))|G_n|t_r$ closed paths of length r in G_n. However, on the other hand, the number of closed walks of length r in G_n is equal to the trace of $A(G_n)^r$, which is equal to the sum of the rth powers of the eigenvalues of X_i. Hence

$$t_r = (1+o(1))\frac{1}{|G_n|}\sum_{i=1}^{n} \lambda_i(G_n)^r$$

Hence, in the limit we see that $F(x)$ satisfies

$$\int x^r dF(x) = t_r$$

for all r. It remains then to recover $F(x)$ from its moments, which can be done for example by expanding F in the basis of Chebyshev polynomials.

7.2 The extreme eigenvalues

So, the proportion of eigenvalues of a typical (n,d)-graph which lie outside of the range $[-2\sqrt{d-1}, 2\sqrt{d-1}]$ is tending to 0 as n gets very large, but we this doesn't tell us much about the typical spectral gap, since there may be a vanishingly small proportion of the eigenvalues outside of this range (and indeed, $\lambda_1 = d$ always does).

However, the following theorem of Friedman actually tells us that the typical extremal eigenvalues do not lie too far from the bulk of the eigenvalues.

Theorem 7.4 (Friedman)**.** *Let G be a random (n,d)-graph, then for every $\epsilon > 0$ whp $\lambda(G) \leq 2\sqrt{d-1} + \epsilon$.*

Now that this almost gives an answer to Question 6.8, as it implies that nearly all (n,d)-graphs are very close to being Ramanujan. In fact, it is conjectured that a positive proportion of (n,d)-graphs should be Ramanujan, which is supported by computational evidence for small d.

The proof, which is too long and complex to include in detail, is based on the *trace method*, which we used earlier in our proof of Theorem 6.2: One estimates, by combinatorial means the trace of $A(G)^{2k}$ for some large k and substracts the 'leading' term $\lambda_1^{2k} = d^{2k}$ from it. For large enough k what remains will be dominated by the term λ^{2k}, and so we can estimate $|\lambda|$. On the one hand, the larger we choose k, the more dominant the contribution of λ^{2k} is and so the better our approximation comes, but on the other hand as k grows the enumeration of the number of closed walks become less precise, and so one needs to choose k carefully.

Instead, we will give a proof of a weaker bound using similar ideas.

Theorem 7.5 (Broder-Shamir)**.** *Let G be a random $(n,2d)$-graph. Then $\lambda(G) = O_p\left(d^{\frac{3}{4}}\right)$.*

Rather than working again with the *configuration model* we will instead work with a slightly simpler, and more convenient model, which generates regular graphs of even degrees called the *permutation model*.

We choose a random $2d$-regular graph on n vertices by choosing d Hamilton cycles in K_n independently and uniformly at random, and taking their union. It is easy enough to generate Hamlton cycles at random, we simply pick a random permutation of the vertices, and so this model is easy to generate. So, more explicitly, we choose d permutations $\pi_1, \ldots, \pi_d$ in the symmetric group S_n independently and uniformly at random and we form our graph $G(n,2d)$ on $[n]$ by taking

$$E(G) = \{(v, \pi_i(v)) \colon i \in [d], v \in [n]\}.$$

However, there is a problem with this model, it very clearly is not uniformly distributed on the set of all $(n,2d)$-graphs! Indeed, $G(n,2d)$ always contains a Hamilton cycle, by construction, but there exist $(n,2d)$-graphs which are no Hamiltonian. However, it can be shown that for $d \geq 2$ the distribution of $G(n,2d)$ is *contiguous* to the uniform distribution: Namely a family of events occurs whp in the one distribution iff it occurs whp in the other distribution. In other words, the distributions agree on asymptotically almost sure events.

Proof of Theorem 7.5. Let $G = G(n, 2d)$ be generated according to the permutation model with permutations $\pi_1, \dots, \pi_d$, and let $\hat{A}$ be the transition matrix of the random walk on G (i.e., $\frac{1}{2d}A(G)$). Let us write $1 = \mu_1 \geq \mu_2 \geq \dots \geq \mu_n$ for the eigenvalues of $\hat{A}$ and $\rho = \max\{|\mu_2|, |\mu_n|\}$. Since the eigenvalues of $\hat{A}^k$ are μ_i^k, we have that

$$\rho^{2k} \leq \operatorname{tr}\left(\hat{A}^{2k}\right) - 1$$

for any positive k. In particular, by Jensen's inequality we see that

$$\mathbb{E}(\rho) \leq \left(\mathbb{E}\left(\rho^{2k}\right)\right)^{\frac{1}{2k}} \leq \left(\mathbb{E}\left(\operatorname{tr}\left(\hat{A}^{2k}\right)\right) - 1\right)^{\frac{1}{2k}}.$$

We will apply this with a particular choice of k, that we will fix later, that will be large, but still much smaller than n.

Observe that the walks in G starting at the vertex 1 are in one to one correspondence to words over the alphabet $\Sigma = \{\pi_1, \pi_1^{-1}, \dots, \pi_d, \pi_d^{-1}\}$. Indeed, we can think of the directed edge $(v, \pi_i(v))$ as being labelled by π_i and the directed edge $(\pi_i(v), v)$ as being labelled with π_i^{-1}, and interpret a word as a sequence of edge labels to follow.

In this way a word $\omega \in \Sigma^{2k}$ can be thought of as a function $\omega : V \to V$ which maps a vertex $i \in V$ to the vertex j obtained by following the walk ω starting at i. In fact, as a composition of permutations this is also a permutation. Then, since $\left(\hat{A}^{2k}\right)_{i,i}$ can be interpreted as the probability that a random walk of length $2k$ starting at i also ends at i, if we pick a word $\omega \in \Sigma^{2k}$ uniformly at random we see that

$$\mathbb{E}\left(\operatorname{tr}\left(\hat{A}^{2k}\right)\right) = \mathbb{E}(|\{i \colon \omega(i) = i)\}|) = n\mathbb{P}(\omega(1) = 1),$$

where the last equality follows by symmetry.

We will analyse this probability in two parts, by considering the two sources of randomness, the random word ω over the alphabet Σ and the random permutations π_i, separately. We first consider the structure of the word ω as an element of the free group in d generators. We show that a random word (or at least, a random word once we have *reduced* consecutive pairs of the form $\pi_i\pi_i^{-1}$) is very unlikely to exhibit some nontrivial periodicity properties, and so almost all reduced words are *good* (for the right definition of good). We then study, for a fixed reduced *good* word, the probability that $\omega'(1) = 1$, where this probability is taken over our choices for $\pi_1, \dots, \pi_d \in S_n$.

So, given a word $\omega \in \Sigma^{2k}$ let $\omega' = \operatorname{red}(\omega)$ be the word obtained by repeatedly cancelling factors of the form $\pi_i\pi_i^{-1}$ from the word until non remain. Clearly $\mathbb{P}(\omega(1) = 1) = \mathbb{P}(\omega'(1) = 1)$, and so it will be sufficient to consider these reduced words. We say that a reduced word ω' is *bad* if it has the form $\omega' = \omega_a\omega_b^r\omega_a^{-1}$ for some words ω_a, ω_b and some $j \geq 2$. For example, the word $\pi_1\pi_3^{-1}\pi_2\pi_4\pi_2\pi_4\pi_3\pi_1^{-1} = (\pi_1\pi_3^{-1})(\pi_2\pi_4)^2(\pi_1\pi_3^{-1})^{-1}$ is bad. Note in particular that the empty word is bad.

It will turn out that it is feasible to calculate the probability, over choices of π_i, that $\omega'(1) = 1$ for a good reduced word, and also the probability, over choices of ω, that $\omega' = \operatorname{red}(\omega)$ is good. The first claim bounds this latter probability.

Claim 7.6. Let $\omega \in \Sigma^{2k}$ be chosen uniformly at random and let $\omega' = \operatorname{red}(\omega)$. Then

$$\mathbb{P}(\omega' \text{ is bad}) \leq k^2\left(\frac{2}{d}\right)^k.$$

Proof of claim. The basic idea of the proof is that if ω has a bad reduction, it can be determined from a very small subset of its letters: namely half of the pairs $\pi_i \pi_i^{-1}$ which we reduce and then the words ω_a and ω_b. In particular, this is less than half of the letters appearing in ω. However, implementing this idea is tricky. One important observation is that in order to generate ω uniformly at random we can choose each letter independently and uniformly at random from Σ.

First, let us oberve that all words ω of length $2k$ which reduce to a word ω' of length 2ℓ can be generated as follows: We start with a string which consists of $k-\ell$ left brackets and $k-\ell$ right brackets, where each initial segment contains at least as many left brackets as right brackets. Note that the brackets in such a sequence then can be paired up in a unique 'permissable' way, so that if we repeatedly delete adjacent pairs we end up with an empty string.

The *level* of an initial segment of this string is the difference between the number of left and right brackets that it contains. Note that the level is always non-negative. We now place a total of 2ℓ $*$s into our sequence, where we are only allowed to place (an abitrary number of) $*$s at a point where the level is 0 to obtain a string in $\{(,), *\}^{2k}$. However, we note that the string is in fact uniquely determined by the positions of the $k-\ell$ left brackets. Indeed, given an initial segment if the next entry is not a left bracket then either the level is 0, in which case the next entry cannot be a right bracket and so must be a $*$, or the level is not 0, in which case the next entry cannot be a $*$ and must be a right bracket. It follows that the total number of such strings we can build in this manner is $\binom{2k}{k-\ell}$. Let us write $\mathcal{S}(k, \ell)$ for the set of all such strings.

We say that a word $\omega \in \Sigma^{2k}$ *matches* such a string $S \in \mathcal{S}(k, \ell)$ if two conditions hold:

1. We can assign a letter in Σ to each left bracket and $*$, and then assign to each right bracket the inverse of the letter assigned to its paired left bracket, in such a way that the word we obtain is ω;
2. The sequence $\omega(*)$ assigned to the stars is equal to $\mathrm{red}(\omega) = \omega'$.

Given a fixed string $S \in \mathcal{S}(k, \ell)$, we want to consider the probability that ω matches S and $\mathrm{red}(\omega) = \omega'$ is bad. This is a little tricky to bound, since these two events are not necessarily independent, and also part of the probability that ω matches a string, specifically the part about the sequence assigned to the $*$s being reduced, is hard to work with. So, instead let us say that ω *weakly matches* S if ω just satisfies condition 1 above. Then it is clear that

$$\mathbb{P}(\omega \text{ matches } S \text{ and } \omega' \text{ is bad}) \leq \mathbb{P}(\omega \text{ weakly matches } S \text{ and } \omega(*) \text{ is bad})$$

since the latter condition is strictly less restrictive. However, now the two events are genuinely independent, since they depend on the value of letters at disjoints sets of positions in the word ω, and so

$$\mathbb{P}(\omega \text{ weakly matches } S \text{ and } \omega(*) \text{ is bad}) \leq \mathbb{P}(\omega \text{ weakly matches } S)\mathbb{P}(\omega(*) \text{ is bad})$$

Bounding the first quantity is easy - since we choose the letters of ω independently and uniformly at random, the probability that ω weakly matches S is at most the probability that the letter assigned to the left brackets match the letters assigned to the right brackets, which is at most $(2d)^{-k+\ell}$.

In order to bound the probability of the second event, let us split into cases according to the length of the words ω_a and ω_b in the decomposition $\omega(*) = \omega_a\omega_b^r\omega_a^{-1}$. Note that, since $\omega(*) \in \Sigma^{2\ell}$, the lengths of ω_a and ω_b determine j, which by assumption ≥ 2.

For a fixed r and s, the probability that there exist words ω_a and ω_b such that $|\omega_a| = r$ and $\omega_b = s$ and $\omega(*) = \omega_a\omega_b^r\omega_a^{-1}$ can be very naively bounded above is by $(2d)^{-\ell}$, since at the very least the second half of the word $\omega(*)$ is then determined by the first half. It follows that

$$\mathbb{P}(\omega(*) \text{ is bad}) \leq \sum_r \sum_s (2d)^{-\ell} \leq k^2(2d)^{-\ell}$$

and so

$$\mathbb{P}(\omega \text{ matches } S \text{ and } \omega' \text{ is bad}) \leq (2d)^{-k+\ell}k^2(2d)^{-\ell} = k^2(2d)^{-k}$$

Putting this all together we see that

$$\begin{aligned}
\mathbb{P}(\omega' \text{ is bad}) &= \sum_{\ell \leq k} \mathbb{P}(\omega' \text{ is bad and } |\omega'| = 2\ell) \\
&\leq \sum_{\ell \leq k} \sum_{S \in \mathcal{S}(k,\ell)} \mathbb{P}(\omega \text{ matches } S \text{ and } \omega' \text{ is bad}) \\
&\leq \sum_{\ell \leq k} \binom{2k}{k-\ell} k^2(2d)^{-k} \\
&= k^2(2d)^{-k} \sum_{\ell \leq k} \binom{2k}{k-\ell} \\
&= k^2(2d)^{-k}2^{2k-1} \\
&\leq k^2 \left(\frac{2}{d}\right)^k.
\end{aligned}$$

□

So, let us fix a good reduced word ω' of length $s \leq 2k$. We are interested in the probability, over our random choices of π_i, that $\omega'(1) = 1$. Recall that ω' is a word of length s over the alphabet Σ, and once we choose the permutations π_i it will correspond to a closed walk on the vertices $[n]$. At a high level then, the idea of the proof is to think of exposing this path, via our choice of π_i, 'as we go'. That is, we will only expose the value of $\pi_i(j)$ when it becomes relevant in this path.

More precisely, we note that we can expose sequentially the image or preimage of elements $i \in [n]$ under π_i uniformly at random from all 'possible' choices (i.e., values that are not excluded by the previous values we exposed) and the resulting permutation will be uniformly distributed on S_n.

So, our word ω' corresponds to a walk $v_0, v_1, \ldots, v_s$ and we 'uncover' the vertices one by one. For each i we have that $v_i = \sigma_i(v_{i-1})$, where σ_i is the ith letter in ω' (note that $\sigma_i = \pi_j^{\pm 1}$ for some j). We call step i in the walk *free* if the value of $\sigma_i(v_{i-1})$ has not yet been 'revealed' in a previous step, and is so undetermined. If this is the case, and t values of σ_i have previously been revealed (i.e., $\sigma_j = \sigma_i^{\pm 1}$ for t many free steps $j \leq i$), then we select v_i uniformly at random from among the $n - t$ elements not assigned to the range of σ_i. Otherwise the step is *forced* and we set v_i to be the, previously revealed, value of $\sigma_i(v_{i-1})$.

We note that,by our previous observations, the walk $v_0, v_1, \ldots, v_s$ generated in this way has the same distribution as if we chose the π_i independently and uniformly at random from S_n. We call a step i a *coincidence* if it is free and moreover the (randomly selected) vertex v_i coincides with a previous step on the path, i.e. $v_i \in \{v_0, \ldots, v_{i-1}\}$. If we let C_i be the event that step i is a coincidence then it is clear that

$$\mathbb{P}(C_i | v_0 = u_0, \ldots, v_{i-1} = u_{i-1}) \leq \frac{s}{n-s}$$

since, if step i is free, then we have at least $n - s$ choices for v_i (since at most s value of any π_j have been revealed), and for the step to be a coincidence this choice must coincide with one of the $i \leq s$ vertices already visited in the walk. Hence the probability of a coincidence at step i is at most $\frac{s}{n-s} \leq \frac{2k}{n-2k}$, independent of the preceding history of the process.

We first claim that if the event $\omega'(1) = 1$ holds, then at least one coincidence must occur. Indeed, since $v_s = 1$ there is some $k \in [n]$ such that there exists $i < j \in [s]$ with $v_i = v_j = k$. Let us choose such a k where j is minimal (and so v_i and v_j are the first two times that k appears in the walk). Then, step i must have been free, since k had not previously appeared in the walk. Furthermore, unless $\sigma_j \in \{\sigma_i, \sigma_{i+1}^{-1}\}$ then the jth step was also a coincidence, since we can't previously have assigned the value k for σ_j. In the first case, if $\sigma_j = \sigma_i$ then we must have that $v_{i-1} = \sigma_i^{-1}(k) = v_{j-1} = \sigma_j^{-1}(k)$, contradicting our choice of k, i and j. Similarly in the second case if $\sigma_j = \sigma_{i+1}^{-1}$ then we see that, since ω' is reduced, $j-1 \neq i+1$, and so $v_{i+1} = v_{j-1} = \sigma_{i+1}(k)$ contradicts our choice of k, i and j.

So, we can bound the probability that $\omega'(1) = 1$ by the sum of the following two events

(a) At least two coincidence occurred along the path;

(b) Exactly one coincidence occurred and $v_s = 1$.

The first probability is easily bounded. Conditioned on the positions i and j of the first two coincidences, the probability that the two coincidence occur is at most

$$\mathbb{P}(C_i)\mathbb{P}(C_j | C_i) \leq \left(\frac{2k}{n-2k}\right)^2.$$

However, there are at most s^2 choices for i and j and so the total probability that (a) occurs is at most $O\left(\frac{s^2k^2}{(n-2k)^2}\right) = O\left(\frac{k^4}{(n-2k)^2}\right)$.

Let us then bound the probability that $v_s = 1$ and there is exactly one coincidence is at most $\frac{1}{n-s+1} \leq \frac{1}{n-2k}$.

A *realisation* of ω' is any walk on $[n]$ which corresponds to ω' under a choice of the permutations π_i. Let us consider some realisation which satisfies the conditions in the lemma, i.e., $v_s = 1$ and there is exactly one coincidence.

Since ω' is reduced, any initial segment of such a realisation preceding the single coincidence is a simple path. Indeed, the first cycle we close must be because of a coincidence, since the last vertex visited v_i has only been determined as the image of v_{i-1} under σ_i, and the next step cannot be σ_i^{-1} since ω' is reduced. Hence, the step at which the coincidence takes place turns

this path into a 'lollipop', a cycle (possibly a loop) with a (possibly empty) tail. Since no more coincidences take place, and ω' is reduced, it is easy to check by the same argument as before, that no additional edges are visted by the walk after this point.

In particular, since there are no more coincidences and ω' is reduced, in order to eventually reach 1 the walk must from this point revolve around the cycle j times for some $r-1 \geq 0$ and then follow the tail to 0. However, since ω' is good, it follows that $r-1=0$!

It follows that $\omega' = \omega_a\omega_b\omega_a^{-1}$, where ω_a corresponds to the walk along the tail, and so may be empty, and the word ω_b corresponds to the walk around the cycle. Moreover, ω_b must be *cyclically reduced* (i.e., it is reduced and the first and last letters are non inverses of each other). Indeed, if v is the vertex where the coincidence occurs and the walk leaves v on an edge labelled π_i to a vertex w, then if the edge preceding the first coincidence were to be labelled π_i^{-1} then it cannot be a free choice, since we have already revealed that $\pi_i(v) = w$. However, then this condition determines the decomposition $\omega' = \omega_a\omega_b\omega_a^{-1}$.

Hence, if we fix an ω' which can result in such a realisation, there is a fixed decomposition $\omega' = \omega_a\omega_b\omega_a^{-1}$ and so, if we let $r = |\omega_a| + |\omega_b|$, the probability that the realisation satisfies the conditions in the lemma is at most the probability that the $(r-1)$th step is a free move to a *specific* previously visited vertex $v_{|\omega_a|}$. However the probability that this happens is then at most $\frac{1}{n-r} \leq \frac{1}{n-s+1}$ and the claim follows.

Putting the bounds of the two claims together we see that

$$\begin{aligned}\mathbb{P}(\omega(1)=1) = \mathbb{P}(\omega'(1)=1) &\leq \mathbb{P}(\omega' \text{ is bad}) + \mathbb{P}(\omega' \text{ is good and } \omega'(1)=1)\\ &\leq k^2\left(\frac{2}{d}\right)^k + O\left(\frac{k^4}{(n-2k)^2}\right) + \frac{1}{n-2k}.\end{aligned}$$

If we take $k = (2-\epsilon)\log_{d/2} n$ for an appropriately small ϵ then we see that

$$\mathbb{P}(\omega(1)=1) = \frac{k^2}{n^{2-\epsilon}} + \frac{1}{n} + O\left(\frac{k}{n^2}\right) + O\left(\frac{k^4}{n^2}\right)$$

and so

$$\begin{aligned}\mathbb{E}(\rho) \leq \left(\mathbb{E}\left(\operatorname{tr}\left(\hat{A}^{2k}\right)\right) - 1\right)^{\frac{1}{2k}} &= \left(n\mathbb{P}(\omega(1)=1) - 1\right)^{\frac{1}{2k}}\\ &= \left(\frac{k^2}{n^{1-\epsilon}} + O\left(\frac{k^4}{n}\right)\right)^{\frac{1}{(4-2\epsilon)\log_{d/2} n}}\\ &= (1+o(1))\left(\frac{2}{d}\right)^{\frac{1-\epsilon}{4-2\epsilon}}\\ &= (1+o(1))\left(\frac{2}{d}\right)^{\frac{1}{4}}\end{aligned}$$

for $\epsilon(d)$ sufficiently small. In particular, by Markov's inequality, with a probability tending to 0 with C we have that $\rho \leq C\left(\frac{2}{d}\right)^{\frac{1}{4}}$. In particular,

$$\lambda(G) = d\rho = O_p\left(d^{\frac{3}{4}}\right).$$

□

8 The Margulis construction

Finally, we come to the problem of actually constructing expander graphs, or, as turns out to be the harder part, proving their expansion properties.

In this section we will describe the first explicit construction of a family of expander graphs, which is particularly elegant.

They can be viewed as a discrete analogue of the following continuous object. Consider an infinite graph whose vertex set is the unit torus, viewed as $I \times I$ where I is the half open unit interval $[0, 1)$. The edges are defined by two linear transformations

$$T(x, y) = (x + y, y) \mod 1 \qquad \text{and} \qquad S(x, y) = (x, x + y) \mod 1,$$

where the neighbours of a point (x, y) are given by $S(x, y), T(x, y), S^{-1}(x, y)$ and $T^{-1}(x, y)$. Gabber and Galil showed the following theorem on the expansion of this infinite graph.

Theorem 8.1. *There exists an explicit $\epsilon > 0$ such that for any measurable $A \subseteq I \times I$ with Lebesgue measure $\mu(A) \leq \frac{1}{2}$*

$$\mu(A \cup \Gamma(A)) \geq (1 + \epsilon)\mu(A),$$

where $\Gamma(A) = S(A) \cup T(A) \cup S^{-1}(A) \cup T^{-1}(A)$ is the neighbourhood of A.

There are natural conjectures for what the optimally expanding sets should be. Namely if we take $A = \{(x, y) \colon |x| + |y| > t\}$ (where $|x| = \min\{x, 1 - x\}$) for some small t, then $A \cup \Gamma(A) = \{(x, y) \colon |x|, |y| < t\}$, so that

$$\mu(A \cup \Gamma(A)) = 4t^2 = 2\mu(A).$$

Conjecture 8.2 (Linial)**.** *For every $A \subseteq I \times I$ of Lebesgue measure $\mu(A) \leq \frac{1}{2}$,*

$$\mu(A \cup \Gamma(A)) \geq 2\mu(A).$$

There is some evidence for this conjecture, as it was recently shown, via a short elementary proof using symmetrisation, that this bound does hold if we take the similar graph defined on $\mathbb{R}^2$ instead.

In order to define our expander family, we will define discrete analogues of the graph defined above. This construction is very simple, and easy to generate, but proving its expansion properties turns out to be difficult!

We work again over a (discrete) torus and take the same linear transformations to define our edge set, however we will also take some affine shifts of these transformations (in some way as a discrete substitute for the continuity of the previous example). In fact, these extra edges are not necessary for the expansion of the graphs, but it will make the analysis easier.

Let $G_n = (V, E)$ be an 8-regular graph with vertex set $V = \mathbb{Z}_n \times \mathbb{Z}_n$ the discrete torus. We let

$$T_1 = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, T_2 = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}, \boldsymbol{e_1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \boldsymbol{e_2} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

then the neighbours of a vertex $\boldsymbol{v}$ are given by $T_v\boldsymbol{v}, T_2\boldsymbol{v}, T_1\boldsymbol{v}+\boldsymbol{e_1}, T_2\boldsymbol{v}+\boldsymbol{e_2}$, and the other four neighbours are given by the inverse transformations, where all calculations are performed mod n.

These are not quite, but are closely related to, the graphs that Margulis considered. He showed that this is a family of expander graphs, but his proof was existential and did not give a lower bound on the spectral gap. Later on, using harmonic analysis, Gabber and Galil gave the following explicit lower bound on the gap (in fact, an upper bound on the generalised second eigenvalue)

Theorem 8.3 (Gabber-Galil)**.** *For each $n \in \mathbb{N}$ the graph G_n satisfies $\lambda(G_n) \leq 5\sqrt{2} < 8$.*

We will prove a slightly weaker bound on λ (which is still strictly smaller than 8), following a simplification of a proof of Jimbo and Marouka due to Boppana. It is then a simple consequence of Cheeger's inequality that these graphs are a family of α-expanders for some fixed $\alpha > 0$. As always we are interested in the Rayleigh quotient of a vector $\boldsymbol{f}$ such that $\boldsymbol{f} \perp \boldsymbol{u}$. In this section it will be convenient to consider our vectors instead as functions $f\colon \mathbb{Z}_n^2 \to \mathbb{R}$. In this language we are interested in a function f such that $\sum_{\boldsymbol{x}} f(\boldsymbol{x}) = 0$ and

$$2\sum_{(\boldsymbol{x},\boldsymbol{y})\in E} f(\boldsymbol{x})f(\boldsymbol{y}) = fAf^T \leq 5\sqrt{2}||f||_2^2 = 5\sqrt{2}\sum_x f(x)^2,$$

Hence we can rewrite Theorem 8.3 in the following form

Theorem 8.4. *For any $f\colon \mathbb{Z}_n^2 \to \mathbb{R}$ such that $\sum_{\boldsymbol{x}} f(\boldsymbol{x}) = 0$, the following inequality holds:*

$$\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} f(\boldsymbol{x})\cdot\big(f(T_1\boldsymbol{x}) + f(T_1\boldsymbol{x}+\boldsymbol{e_1}) + f(T_2\boldsymbol{x}) + f(T_2\boldsymbol{x}+\boldsymbol{e_2})\big) \leq \frac{5}{\sqrt{2}}\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} f(\boldsymbol{x})^2. \tag{8.1}$$

In order to prove this, we will need to introduce some tools from fourier analysis.

8.1 A (very) brief introduction to discrete fourier anaylysis

We should stress that the following is very much a 'whistle-stop' introduction to the idea of discrete fourier analysis, which has (relatively) recently been used with much success in various areas of discrete mathematics. We will state, without proof, some of the basic facts and tools from the area, but we note that none of the results we mention are difficult to prove (and we will prove some of them on the example sheet).

A *character* of a group H is a homomorphism $\chi : H \to \mathbb{C}^*$ (the multiplicative group of $\mathbb{C}$), i.e, a function such that $\chi(gh) = \chi(g)\cdot\chi(h)$ for all $g, h \in H$. Let $\hat{H}$ denote the set of characters of H, which we forms an abelian group under pointwise multiplication. Note, in particular, that for any finite group the range of χ must be contained in the unit circle. Furthermore, when H is abelian (as will be the case in this section) we will write the group operation as $+$.

For example every group has the *trivial character* χ_0 which maps all elements to 1. It can be shown that the cyclic group $\mathbb{Z}_n$ has n characters given by the functions

$$\chi_k(h) = e^{\frac{2\pi i k h}{n}} \text{ for } 0 \leq k \leq n-1.$$

More generally, for the group $\mathbb{Z}_n^2$ we have n^2 characters, one for each $\boldsymbol{b} = (b_1, b_2) \in \mathbb{Z}_n^2$, where

$$\chi_{\boldsymbol{b}}(a_1, a_2) = \omega^{a_1 b_1 + a_2 b_2} = \omega^{\langle \boldsymbol{a}, \boldsymbol{b} \rangle},$$

for some primitive nth root of unity ω.

We write $\mathcal{F}(H)$ for the set of all complex functions on H, which is a linear space with an inner product given by

$$\langle f, g \rangle = \sum_{x \in H} f(x)\overline{g(x)}.$$

Broadly the idea of fourier analysis on a group H is to expand functions in $\mathcal{F}$ as linear combinations of characters.

Proposition 8.5. *Every finite abelian group H has $|H|$ distinct characters, which can be naturally indexed as $\{\chi_x \colon x \in H\}$, which form an orthogonal basis of $\mathcal{F}$. In particular every $f : H \to \mathbb{C}$ can be uniquely expressed as a sum $f = \frac{1}{|H|} \sum_{x \in H} \hat{f}(x)\chi_x$, where $\hat{f} \colon H \to \mathbb{C}$ is the* discrete fourier transform *of f, given by*

$$\hat{f}(x) = \langle f, \chi_x \rangle = \sum_{y \in H} f(y)\overline{\chi_x(y)} = \sum_{y \in H} f(y)\chi_x(-y).$$

Proof. We first note that for any character H and $b \in H$

$$\chi(b) \sum_{a \in H} \chi(a) = \sum_{a \in H} \chi(a + b) = \sum_{c \in H} \chi(c).$$

In particular, if χ is non-trivial then there exists b such that $\chi(b) \neq 1$ and it follows that

$$\sum_{a \in H} \chi(a) = (\chi(b) - 1) \sum_{a \in H} \chi(a) = 0.$$

It follows that for any two distinct characters χ_1 and χ_2, $\chi_1\overline{\chi_2}$is a non-trivial character and hence

$$\langle \chi_1, \chi_2 \rangle = \sum_{a \in H} \chi_1(a)\overline{\chi_2(a)} = \sum_{a \in H} \big(\chi_1\overline{\chi_2}\big)(a) = 0.$$

Furthermore, for any $\chi \in \hat{H}$

$$\langle \chi, \chi \rangle = \sum_{a \in H} \chi(a)\chi(-a) = \sum_{a \in H} \chi(0) = |H|.$$

In particular, there can be at most $|H|$ many distinct characters. In the case of $\mathbb{Z}_n$ it is easy to exhibit n distinct characters $\chi_1, \ldots, \chi_n$ as above. More generally it is easy to see that if $H = H_1 \oplus H_2$ and χ_i are characters of H_i then $\chi(h_1, h_2) = \chi_1(h_1)\chi_2(h_2)$ is a character of H and that this map from $\hat{H}_1 \times \hat{H}_2 \to \hat{H}$ is injective. In particular, for cardinality reasons, it follows that $\hat{H} \cong \hat{H}_1 \times \hat{H}_2$.

Using the classification theorem for finite abelian groups, and the above observation, we see that, when H is abelian, $H \cong \hat{H}$ (although this isomorphism is not natural) and so, by dimensional considerations $\hat{H}$ forms a basis for the $|H|$-dimensional space $\mathcal{F}(C)$. □

Whilst $H \cong \hat{H}$, we treat the two spaces differently in that we normally consider H with the 'counting measure', and consider $\hat{H}$ with a 'probability measure', which is just given by normalising our sums by the size of the group. For example, we can define another type of inner product

$$\mathop{\mathbb{E}}_{a\in H} f(a)\overline{g(a)} = \frac{1}{|H|}\sum_{a\in H} f(a)\overline{g(a)}.$$

This perhaps makes sense since in a broader setting this dualising operation turns discrete groups into compact groups, and vice versa.

In our particular case of interest, $H = \mathbb{Z}_n^2$, it can be checked that the discrete fourier transform of f takes the form

$$\hat{f}(\boldsymbol{x}) = \sum_{\boldsymbol{b}\in\mathbb{Z}_n^2} f(\boldsymbol{b})\omega^{-b_1x_1-b_2x_2}.$$

In the following proposition we collect some basic properties of the fourier transform.

Proposition 8.6. *Let H be an abelian group and let $f, g \in \mathcal{F}(H)$.*

(a) $\sum_{a\in H} f(a) = 0 \iff \hat{f}(0) = 0$;

(b) $\langle f, g\rangle = \frac{1}{|H|}\langle \hat{f}, \hat{g}\rangle$;

(c) As a special case of the above with $f = g$ we obtain Parseval's identity

$$\sum_{a\in H} |f(a)|^2 = \frac{1}{|H|}\sum_{a\in H} |\hat{f}(a)|^2;$$

(d) the inverse formula

$$f(a) = \frac{1}{|H|}\sum_{b\in H} \hat{f}(b)\chi_b(a);$$

In the specific case $H = \mathbb{Z}_n^2$ we also have the useful shift property

(e) If A is a non-singular 2×2 matrix over $\mathbb{Z}_n$, $\boldsymbol{b}\in\mathbb{Z}_n^2$ and $g(\boldsymbol{x}) = f(A\boldsymbol{x}+\boldsymbol{b})$, then

$$\hat{g}(\boldsymbol{y}) = \omega^{-\langle A^{-1}\boldsymbol{b},\boldsymbol{y}\rangle}\hat{f}((A^{-1})^T\boldsymbol{y}).$$

Proof. For statement (a) we note that

$$\hat{f}(0) = \langle f, \chi_0\rangle = \sum_{a\in H} f(a)\overline{\chi_0(a)} = \sum_{a\in H} f(a).$$

For (b), known as the Plancherel formula, we first define the *character table* C of H, which is a matrix whose rows and columns are indexed by H such that $C_{a,b} = \chi_a(b)$. We note that it follows from Proposition 8.5 that $CC^* = |H|I$. Indeed

$$(CC^*)_{a,b} = \sum_c \chi_a(c)\overline{\chi_b(c)} = |H|1_{a=b},$$

where we used the orthonormality of $\hat{H}$ and the fact that

$$\sum_{a\in H}\chi(a)\overline{\chi(a)}=\sum_{a\in H}\chi(a)\chi(a)^{-1}=\sum_{a\in H}\chi(a)\chi(a^{-1})=\sum_{a\in H}\chi(1)=|H|.$$

Since, by definition, $\hat{f}=fC$, we see that

$$\langle\hat{f},\hat{g}\rangle=\hat{f}\cdot\hat{g}^*=fCC^*g^*=|H|f\cdot g^*=|H|\langle f,g\rangle.$$

Then (c) follows by setting $f=g$.

For (d) we note that for any $b\in H$

$$||\chi_b||_2^2=\langle\chi_b,\chi_b\rangle=\sum_{a\in H}\chi_b(a)\overline{\chi_b(a)}=\sum_{a\in H}\chi_b(a)\chi_b(-a)=\sum_{a\in H}\chi_b(0)=|H|.$$

Hence, since the characters form an orthogonal basis of $\mathcal{F}(H)$

$$f=\sum_{b\in H}\frac{\langle f,\chi_b\rangle}{||\chi_b||_2^2}\chi_b=\frac{1}{|H|}\sum_{b\in H}\hat{f}(b)\chi_b.$$

Finally for (e) we see that

$$\begin{aligned}\hat{g}(\boldsymbol{y})&=\sum_{\boldsymbol{z}\in H}g(\boldsymbol{z})\chi_z(-\boldsymbol{y})\\&=\sum_{\boldsymbol{z}\in H}f(A\boldsymbol{z}+\boldsymbol{b})\omega^{\langle\boldsymbol{z},\boldsymbol{y}\rangle}\\&=\sum_{\boldsymbol{x}\in H}f(\boldsymbol{x})\omega^{\langle A^{-1}(\boldsymbol{x}-\boldsymbol{b}),\boldsymbol{y}\rangle}\\&=\omega^{-\langle A^{-1}\boldsymbol{b}),\boldsymbol{y}\rangle}\sum_{\boldsymbol{x}\in H}f(\boldsymbol{x})\omega^{\langle A^{-1}\boldsymbol{x},\boldsymbol{y}\rangle}\\&=\omega^{-\langle A^{-1}\boldsymbol{b}),\boldsymbol{y}\rangle}\sum_{\boldsymbol{x}\in H}f(\boldsymbol{x})\omega^{\langle\boldsymbol{x},(A^{-1})^T\boldsymbol{y}\rangle}\\&=\omega^{-\langle A^{-1}\boldsymbol{b},\boldsymbol{y}\rangle}\hat{f}((A^{-1})^T\boldsymbol{y}).\end{aligned}$$

□

8.2 Proof of Theorem 8.4

Our aim is to express (8.1) in terms of the fourier coefficients of f. The condition $\sum_x f(x)=0$ can be rewritten as $\hat{f}(0,0)=0$ by (a). Then, using Parseval's identity (c) to see that

$$\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2}f(\boldsymbol{x})^2=||f||_2^2=\frac{1}{n^2}||\hat{f}||_2^2,$$

and furthermore by the Plancherel formula (b) and the shift property (e) we see that

$$\begin{aligned}\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} f(\boldsymbol{x})f(T_1\boldsymbol{x}) &= \sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} \overline{f(\boldsymbol{x})}f(T_1\boldsymbol{x}) \\ &= \frac{1}{n^2}\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} \overline{\hat{f}(\boldsymbol{x})}\hat{f}(T_1\boldsymbol{x}) \\ &= \frac{1}{n^2}\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} \overline{\hat{f}(\boldsymbol{x})}\hat{f}((T_1^{-1})^T\boldsymbol{x}) \\ &= \frac{1}{n^2}\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} \overline{\hat{f}(\boldsymbol{x})}\hat{f}(T_2^{-1}\boldsymbol{x})\end{aligned}$$

since $T_1^{-1} = \begin{pmatrix}1 & -2\\ 0 & 1\end{pmatrix} = (T_2^{-1})^T$, and

$$\begin{aligned}\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} f(\boldsymbol{x})f(T_1\boldsymbol{x}+\boldsymbol{e_1}) &= \sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} \overline{f(\boldsymbol{x})}f(T_1\boldsymbol{x}+\boldsymbol{e_1}) \\ &= \frac{1}{n^2}\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} \overline{\hat{f}(\boldsymbol{x})}\hat{f}(T_1\boldsymbol{x}) \\ &= \frac{1}{n^2}\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} \overline{\hat{f}(\boldsymbol{x})}\omega^{-\langle T_1^{-1}\boldsymbol{e_1},\boldsymbol{x}\rangle}\hat{f}(T_2^{-1}\boldsymbol{x}) \\ &:= \frac{1}{n^2}\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} \omega^{-x_1}\overline{\hat{f}(\boldsymbol{x})}\hat{f}(\hat{T}_1\boldsymbol{x}).\end{aligned}$$

A similar calulation for the other two terms leads to the following inequality which is equivalent to (8.1)

$$\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} \overline{\hat{f}(\boldsymbol{x})}\Big(\hat{f}(T_2^{-1}\boldsymbol{x})\big(1+\omega^{-x_1}\big) + \hat{f}(T_1^{-1}\boldsymbol{x})\big(1+\omega^{-x_2}\big)\Big) \le \frac{5}{\sqrt{2}}\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2}|\hat{f}(x)|^2.$$

In fact, we can prove such a bound for an arbitrary function $F\colon \mathbb{Z}_n^2 \to \mathbb{C}$ with $F(0,0) = 0$.

Lemma 8.7. *For every $F\colon \mathbb{Z}_n^2 \to \mathbb{C}$ with $F(0,0) = 0$,*

$$\left|\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} \overline{F(\boldsymbol{x})}\Big(F(T_2^{-1}\boldsymbol{x})\big(1+\omega^{-x_1}\big) + F(T_1^{-1}\boldsymbol{x})\big(1+\omega^{-x_2}\big)\Big)\right| \le \frac{5}{\sqrt{2}}\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2}|F(\boldsymbol{x})|^2.$$

It is clear then, the Theorem 8.4 follows from Lemma 8.7. Replacing F with $G = |F|$, using the triangle inequality and the identity $|1+\omega^{-t}| = 2\left|\cos\frac{\pi t}{n}\right|$ we see that it suffices to prove

$$\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} 2G(\boldsymbol{x})\Big(G(T_2^{-1}\boldsymbol{x})\left|\cos\frac{\pi x_1}{n}\right| + G(T_1^{-1}\boldsymbol{x})\left|\cos\frac{\pi x_2}{n}\right|\Big) \le \frac{5}{\sqrt{2}}\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} G(\boldsymbol{x})^2, \tag{8.2}$$

for any real function $G\colon \mathbb{Z}_n^2 \to \mathbb{R}$.

One way to approach the problem that the 'product' terms on the left are difficult to sum, would be to try to replace them with square terms (which are what appear on the right) using

the elementary inequality $2\alpha\beta \le \alpha^2 + \beta^2$. This would then lead to, for any $\boldsymbol{x}$

$$\begin{aligned}
&\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} 2G(\boldsymbol{x})\Big(G(T_2^{-1}\boldsymbol{x})\left|\cos\frac{\pi x_1}{n}\right| + G(T_1^{-1}\boldsymbol{x})\left|\cos\frac{\pi x_2}{n}\right|\Big)\\
&\le \sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} 2G(\boldsymbol{x})^2 + G(T_2^{-1}\boldsymbol{x})^2\left|\cos\frac{\pi x_1}{n}\right| + G(T_1^{-1}\boldsymbol{x})^2\left|\cos\frac{\pi x_2}{n}\right|\\
&\le \sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} 4G(\boldsymbol{x})^2,
\end{aligned}$$

using that fact that T_2^{-1} and T_1^{-1} are invertible and that $|\cos x| \le 1$. However this is not quite good enough, in particular for small values of $\boldsymbol{x}$ where the cosines are close to 1.

However, more generally we can use the inequality $2\alpha\beta \le \gamma\alpha + \gamma^{-1}\beta$ which holds for any positive γ. Using this inequality for a fixed γ would also not be useful, however we can apply it with a cleverly chosen $\gamma : \left(\mathbb{Z}_n^2\right)^2 \to \mathbb{R}$ to the individual terms

$$2G(\boldsymbol{x})G(\boldsymbol{y}) \le \gamma(\boldsymbol{x},\boldsymbol{y})G(\boldsymbol{x})^2 + \gamma(\boldsymbol{x},\boldsymbol{y})^{-1}G(\boldsymbol{y})^2.$$

If we choose γ so that $\gamma(\boldsymbol{x},\boldsymbol{y})\gamma(\boldsymbol{y},\boldsymbol{x}) = 1$ then we can write this as

$$2G(\boldsymbol{x})G(\boldsymbol{y}) \le \gamma(\boldsymbol{x},\boldsymbol{y})G(\boldsymbol{x})^2 + \gamma(\boldsymbol{y},\boldsymbol{x})G(\boldsymbol{y})^2.$$

The final little trick is to notice that $(T_1\boldsymbol{x})_2 = x_2$ for all $\boldsymbol{x}$ and so

$$\begin{aligned}
\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} \gamma(T_1^{-1}\boldsymbol{x},\boldsymbol{x})G(T_1^{-1}\boldsymbol{x})^2\left|\cos\frac{\pi x_2}{n}\right| &= \sum_{T_1\boldsymbol{x}=\boldsymbol{y}\in\mathbb{Z}_n^2} \gamma(\boldsymbol{y},T_1\boldsymbol{y})G(\boldsymbol{y})^2\left|\cos\frac{\pi(T_1\boldsymbol{y})_2}{n}\right|\\
&= \sum_{\boldsymbol{y}\in\mathbb{Z}_n^2} \gamma(\boldsymbol{y},T_1\boldsymbol{y})G(\boldsymbol{y})^2\left|\cos\frac{\pi y_2}{n}\right|
\end{aligned}$$

and a similar equality holds for T_2.

This allows us to bound

$$\begin{aligned}
&\sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} 2G(\boldsymbol{x})\Big(G(T_2^{-1}\boldsymbol{x})\left|\cos\frac{\pi x_1}{n}\right| + G(T_1^{-1}\boldsymbol{x})\left|\cos\frac{\pi x_2}{n}\right|\Big)\\
&\le \sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} \left|\cos\frac{\pi x_1}{n}\right|\Big(\gamma(\boldsymbol{x},T_2^{-1}\boldsymbol{x})G(\boldsymbol{x})^2 + \gamma(T_2^{-1}\boldsymbol{x},\boldsymbol{x})G(T_2^{-1}\boldsymbol{x})^2\Big)\\
&\quad + \left|\cos\frac{\pi x_2}{n}\right|\Big(\gamma(\boldsymbol{x},T_1^{-1}\boldsymbol{x})G(\boldsymbol{x})^2 + \gamma(T_1^{-1}\boldsymbol{x},\boldsymbol{x})G(T_1^{-1}\boldsymbol{x})^2\Big)\\
&\le \sum_{\boldsymbol{x}\in\mathbb{Z}_n^2} G(\boldsymbol{x})^2\Big(\left|\cos\frac{\pi x_1}{n}\right|\Big(\gamma(\boldsymbol{x},T_2^{-1}\boldsymbol{x}) + \gamma(\boldsymbol{x},T_2\boldsymbol{x})\Big) + \left|\cos\frac{\pi x_2}{n}\right|\Big(\gamma(\boldsymbol{x},T_1^{-1}\boldsymbol{x}) + \gamma(\boldsymbol{x},T_1\boldsymbol{x})\Big)\Big).
\end{aligned}$$

Hence, we just need to choose an appropriate function $\gamma : \left(\mathbb{Z}_n^2\right)^2 \to \mathbb{R}$ so that for all $\boldsymbol{x}$

$$\left|\cos\frac{\pi x_1}{n}\right|\Big(\gamma(\boldsymbol{x},T_2^{-1}\boldsymbol{x}) + \gamma(\boldsymbol{x},T_2\boldsymbol{x})\Big) + \left|\cos\frac{\pi x_2}{n}\right|\Big(\gamma(\boldsymbol{x},T_1^{-1}\boldsymbol{x}) + \gamma(\boldsymbol{x},T_1\boldsymbol{x})\Big) \le \frac{\sqrt{5}}{2}. \tag{8.3}$$

The idea behind the function γ is relatively simple, we will set some values of γ to be $\alpha = \frac{5}{4} > 1$ (and so the inverse values to be $\frac{4}{5}$) and the rest to be 1. We will see that when $\boldsymbol{x}$

is far from the origin, $\left|\cos \frac{\pi x_i}{n}\right|$ is small enough that (8.3) will follow trivially from the fact that $\gamma \leq \frac{5}{4}$.

In order to define 'close to the origin' let us define the *diamond* $\left\{(x,y)\colon a(x)+a(y) < \frac{n}{2}\right\}$ where $a(x) = \min\{x, n-x\}$. Outside of the diamond, since $\left|\cos \frac{\pi z}{n}\right|$ is decreasing with $|z|$ for $|z| \leq \frac{n}{2}$, we have that $\left|\cos \frac{\pi x_1}{n}\right| + \left|\cos \frac{\pi x_2}{n}\right|$ is maximised on the boundary of the diamond. For example, in the first quadrant where $x_1, x_2 \geq 0$ we have on the boundary of the diamond that $x_2 = \frac{n}{2} - x_1$ and so $\cos \frac{\pi x_2}{n} = \sin \frac{\pi x_1}{n}$. It follows that

$$\left|\cos \frac{\pi x_1}{n}\right| + \left|\cos \frac{\pi x_2}{n}\right| = \cos \frac{\pi x_1}{n} + \cos \frac{\pi x_2}{n} = \cos \frac{\pi x_1}{n} + \sin \frac{\pi x_1}{n} \leq \sqrt{2}$$

and so the contribution to (8.3) from such a point is at most $2\frac{5}{4}\sqrt{2} = \frac{5}{\sqrt{2}}$.

Inside the diamond we have to be a little more careful, since the cosines could be close to one, we want to choose γ in such a way that it cannot be that all of

$$\gamma(\boldsymbol{x}, T_2^{-1}\boldsymbol{x}), \gamma(\boldsymbol{x}, T_2\boldsymbol{x}), \gamma(\boldsymbol{x}, T_1^{-1}\boldsymbol{x}) \text{ and } \gamma(\boldsymbol{x}, T_1\boldsymbol{x})$$

are large. In particular, if we can bound their sum by $\frac{5}{\sqrt{2}}$ for each $\boldsymbol{x}$, then (8.3) will follows, since $|\cos x| \leq 1$.

In order then to motivate the definition of our function then, it will be informative to consider how the four points $T_1\boldsymbol{x}, T_1^{-1}\boldsymbol{x}, T_2\boldsymbol{x}, T_2^{-1}\boldsymbol{x}$ can look for points inside the diamond.

We define a partial order $<$ on the torus $\mathbb{Z}_n^2$, which in some way measures distance from the origin, and set

$$\gamma(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} \frac{5}{4} & \text{if } \boldsymbol{x} > \boldsymbol{y} \\ \frac{4}{5} & \text{if } \boldsymbol{y} > \boldsymbol{x} \\ 1 & \text{otherwise.} \end{cases}$$

and we will do so in such a way that for every $\boldsymbol{x} \in \mathbb{Z}_n^2$ either

(a) Three of the four points $T_1\boldsymbol{x}, T_1^{-1}\boldsymbol{x}, T_2\boldsymbol{x}, T_2^{-1}\boldsymbol{x}$ are $> \boldsymbol{x}$ and one is $< \boldsymbol{x}$; or

(b) Two of the four points $T_1\boldsymbol{x}, T_1^{-1}\boldsymbol{x}, T_2\boldsymbol{x}, T_2^{-1}\boldsymbol{x}$ are $> \boldsymbol{x}$ and two are incomarable.

Then in case (a) the LHS of (8.3) is at most $3\frac{4}{5} + \frac{5}{4} = \frac{73}{20}$ and in case (b) the LHS is at most $2\frac{4}{5} + 2 = \frac{18}{5}$. An sharp reader will have noticed this is not quite as good as the claimed bound $\frac{5}{\sqrt{2}}$, however it is sufficient to bound the LHS away from 4, and a careful analysis of the same essential strategy is enough to prove the stronger bound.

So, let us finally define this partial order:

$\boldsymbol{x} > \boldsymbol{y}$ iff $a(x_1) \geq a(y_1)$ and $a(x_2) \geq a(y_2)$ and at least one of the inequalities is strict.

It remains to verify that property (a) or (b) hold, which can be reduced to a series of case checks. Indeed, we are comparing (x_1, x_2) with

$$(x_1 + 2x_2, x_2), (x_1 - 2x_2, x_2), (x_1, x_2 + 2x_1), (x_1, x_2 - 2x_1)$$

Let us first assume that $a(x_1) > a(x_2)$ and so by symmetry we may assume that $x_1 > x_2 \geq 0$ and $x_1 + x_2 < \frac{n}{2}$.

It follows that $a(x_1 - 2x_2) < a(x_1)$ and hence $(x_1 - 2x_2, x_2) < (x_1, x_2)$, however the other three points are $> \boldsymbol{x}$ since

$$a(x_1 + 2x_2) > x_1 \text{ and } a(x_2 \pm 2x_1) > x_2.$$

For example either $a(x_1 + 2x_2) = x_1 + 2x_2 \geq x_1$ or $a(x_1 + 2x_2) = n - x_1 - 2x_2 > \frac{n}{2} - x_2 > x_1$. The other cases follow in a similar manner.

Finally if $a(x_1) = a(x_2)$, and so again by symmetry we may assume that $\frac{n}{4} > x_1 = x_2 \geq 0$, then $a(x_1 - 2x_2) = a(x_1) = a(x_2) = a(x_2 - 2x_1)$ and so the points

$$(x_1 - 2x_2, x_2) \qquad \text{and} \qquad (x_1, x_2 - 2x_1)$$

are incomparable with $\boldsymbol{x}$, whereas it is again an easy check to see that $a(x_1 + 2x_2), a(x_2 + 2x_1) > a(x_1) = a(x_2)$ and so

$$(x_1 + 2x_2, x_2) \qquad \text{and} \qquad (x_1, x_2 + 2x_1)$$

are $> \boldsymbol{x}$.

9 The zig-zag product

In this section we will introduce a new kind of graph product, the *zig-zag product* and show that the zig-zag product of two expanders is still a quiet good expander, which will lead to an iterative construction of an explicit family of expanders. We will also give a recent application of the zig-zag product to complexity theory to show that Symmetric Logspace $SL = L$ Logspace. Logspace is the class of decision problems that can be solved using a deterministic Turing machine using a logarithmic amount of writable memory space, whereas Symmetric Logspace, which was originally described in terms of symmetric Turing machines (which is a bit more complex to introduce), can be more easily described as the class of decision problems which are log-space reducible to USTCON (undirected s-t-connectivity) which is the problem of determining whether there exists a path between two vertices in a graph.

9.1 The zig-zag product

In a slightly non-standard definition, we define the *kth power* of a graph $G = (V, E)$, written as G^k, is the graph with vertex set V and where the number of edges between a pair of vertices in $V(G^k)$ is the number of walks of length k between those two vertices in G. Often, G^k is used to denote instead the graph obtained by joining every two vertices in G at distance at most k. The reason for this slightly uncovential definition is that it interacts nicely with our spectral definitions, in particular it is easy to see that $A(G^k) = A(G)^k$, and so if G is an (n, d, α)-graph, then G^k is an (n, d^k, α^k)-graph.

The zig-zag product, which we write as $\textcircled{z}$, is an asymmetric binary relation on graphs. The zig-zag product of an (n, m)-graph and an (m, d)-graph is an (mn, d^2)-graph. Our main aim will be to show that the zig-zag product of two expanders is an expander. This is already true for various other types of graph products, for example graph powers. The usefulness of the zigzag product is that the degree of the product is a function only of the second graph, and so can be controlled.

Explicitly, let G be an (n, m, α)-graph and let H be and (m, d, β)-graph. for each vertex $v \in V(G)$ we will fix some enumeration $e_v^1, \ldots, e_v^m$ of the edges incident with v, and we will view the vertex set of H as $[m]$. The vertex set of $G \textcircled{z} H$ is given by the cartesian product $V(G) \times V(H)$. It will be useful to think of this vetex set as being given by replacing each vertex $v \in G$ with a cloud of m vertices $(v, 1), \ldots, (v, m)$, one for each edge incident with v.

To describe the edges of $G \textcircled{z} H$, it will be easier to first describe different product graph, $G \textcircled{r} H$ the *replacement product*, which is also on $V(G) \times V(H)$. The edges of $G \textcircled{r} H$ are given by the union of the original edges of G, where the edge e_v^i which is also e_w^j goes from (v, i) to (w, j), together with a copy of H on each cloud. The edges of $G \textcircled{z} H$ now come from walks of length three in $G \textcircled{r} H$, which 'zig-zag' between the copies of H, in that they take one step inside a copy of H, a second step between clouds, and then a third step inside a copy of H.

Formally, we have that a pair $\big((v, i), (w, j)\big) \in E(G \textcircled{z} H)$ if there are $k, \ell \in [m]$ such that $(i, k), (\ell, j) \in E(H)$ and $e_v^k = e_w^\ell$.

We note that the replacement product of graphs had been relatively well studied before, often

used to blow up a graph into a sparses one, with lower degrees, whilst maintaining connectivity properties. Before we analyse explicitly the expansion of the zigzag product, let us give a heuristic explanation, in terms of the entropy of the random walk

9.2 Entropy analysis

Let us return to the perspective suggested earlier in the notes, which considers the stepwise increase in entropy during a random walk. Our intuition should be that if G and H are good expanders, then so is $G \textcircled{z} H$, and so the entropy of the random walk on $G \textcircled{z} H$ should grow significantly with each step. Why should this be the case?

Well, we can view each step in the random walk on $G \textcircled{z} H$ as being made up of a deterministic step between two clouds, coming between two random steps within clouds. Note that the two two random steps within the clouds are entirely independent of each other.

We can think about the distribution p as being composed of two marginal distributions p_G and p_H, the projection of p onto $V(G)$ or $V(H)$. Now, if p_H is far from uniform, then p must be far from uniform on some clouds, and so in the first random step inside the cloud the entropy of p will increase by virtue of H's expansion. The other two steps will not harm this increase (since entropy never decreases when multiplying by a stochastic matrix).

The more interesting case then is where within most clouds the p is close to uniform, although the distribution is not uniformly spread between clouds. In this case the first random step cannot increase the entropy, since the restrictions to each cloud will remain close to uniform. However, if the distribution on most clouds is near uniform, then the (deterministic) middle step of the random walk is like a random step on G, and so by the expansion of G the entropy of p_G must increase.

However, this middle step is simply a permutation on $V(G \textcircled{z} H)$, and so, whilst the entropy of p_G increases, the entropy of the whole distribution must remain unchanged, and so the entropy of p_H must decrease. In particular, p_H is then far from uniform, and so there must be a significant number of clouds on which the distribution is far from uniform, and so in the second random step we are back in the first case, where we get an increase in the entropy of the distribution.

So, in other words, the key point here is that the middle step is simultaneously a permutation (which doesn't effect the entropy of p), but also an operation whose G-marginal is a random step on G.

9.3 Expansion of the zigzag product

Theorem 9.1 (The Zig-Zag Theorem, Reingold-Vadhan-Widgerson)**.** *Let G be an (n, m, α)-graph and let H be and (m, d, β)-graph. Then $G \textcircled{z} H$ is an $(nm, d^2, \phi(\alpha, \beta))$-graph where ϕ satisfies the following:*

(1) If $\alpha < 1$ and $\beta < 1$, then $\phi(\alpha, \beta) < 1$;

(2) $\phi(\alpha, \beta) \leq \alpha + \beta$;

(3) $\phi(\alpha, \beta) \leq 1 - (1 - \beta^2)\frac{1-\alpha}{2}$.

The first bound is just the qualitative statement that the zig-zag product of two expanders is an expander. The two quantitative bounds (2) and (3) are crucial for applications. The first is useful when α and β are small, and so their sum is still small, whereas the latter is useful when α and β are large and the first is no longer effective.

We note that Theorem 9.1 (3) was previously known to hold for the replacement product, although (2) need not. We also note that, if G can be m-coloured (and we take this colouring as the ordering of the edges adjacent at each vertex in the definition of $G \textcircled{z} H$), then $G \textcircled{z} H$ is in fact a *lift* of H^2.

Since the proof of Theorem 9.1 is reasonably involved we will just give proofs of a slighty weaker bounds, which are still strong enough for the construction given in the next sections.

Proof of Theorem 9.1. We will prove (1) and a slightly weaker form of (2) in a similar fashion, following the heuristic sketch we gave in the previous section. Explicitly we will show that

$$\phi(\alpha, \beta) \leq \alpha + \beta + \beta^2.$$

So, we will attempt to bound the spectral gap in $G \textcircled{z} H$ by considering the random walk on $G \textcircled{z} H$. Each step in this walk can be split into three parts

(i) A random step inside a cloud;

(ii) A *deterministic* step between clouds;

(iii) Another random step inside a cloud.

We see that we can write the transition matrix of this walk as follows: If we let $\hat{B} = \hat{A}(H)$, then the random steps in (i) and (iii) are done on n disjoint copies of H, and so the transition matrices in these steps are $\tilde{B} = \hat{B} \times I_n$ (The Kroenecker product, giving a block diagonal matrix). In the deterministic step (ii) we move from a vertex (v, k) to the unique vertex (w, ℓ) for which $e_v^k = e_w^\ell$. Consequently, the transition matrix in this step is given by the permutation matrix P of an involution given by

$$P_{(v,k),(w,\ell)} = \begin{cases} 1 & \text{if } e_v^k = e_w^\ell, \\ 0 & \text{otherwise.} \end{cases}$$

Hence we can write the transition matrix of the random walk on $G \textcircled{z} H$ as $Z = \hat{B} P \hat{B}$. Then, as always, our claim is equivalent to bounding the Rayleigh quotient of vectors $\boldsymbol{f} \perp \boldsymbol{u}$ with Z.

Our hope will be to prove this by decomposing the function $f : V(G) \times [m] \to \mathbb{R}$ in way which reflects the product structure of $G \textcircled{z} H$. Let us define a function h on $V(G) \times [m]$ which is given by the average of $\boldsymbol{f}$ over each cloud. That is

$$h(v, i) = \frac{1}{m} \sum_{j \in [m]} f(v, j).$$

We then define $\boldsymbol{g} = \boldsymbol{f} - \boldsymbol{h}$, so that g sums up to zero on every cloud. Our proof will then in some way formalise the intuition from the previous section. We are splitting the vector $\boldsymbol{f}$ up into a part $\boldsymbol{g}$ which represents the deviation from the uniform of $\boldsymbol{f}$ on each cloud (and shrinks under the random walk step in the clouds) and a part $\boldsymbol{h}$ which represents the probability of being in each cloud, and hence the marginal distribution on $V(G)$, and shrinks under the permutation step.

Then we can expand

$$|\boldsymbol{f}Z\boldsymbol{f}^T| = |(\boldsymbol{g}+\boldsymbol{h})\tilde{B}P\tilde{B}(\boldsymbol{g}+\boldsymbol{h})^T| \leq |\boldsymbol{g}\tilde{B}P\tilde{B}\boldsymbol{g}^T| + 2|\boldsymbol{g}\tilde{B}P\tilde{B}\boldsymbol{h}^T| + |\boldsymbol{h}\tilde{B}P\tilde{B}\boldsymbol{h}^T|.$$

Now, since $\hat{B}$ is a block diagonal matrix, each block of which is $\hat{B}$ and so has $\boldsymbol{u}$ as an eigenvector with eigenvalue one, it follows that $\tilde{B}\boldsymbol{h} = \boldsymbol{h}$. Hence

$$|\boldsymbol{f}Z\boldsymbol{f}^T| \leq |\boldsymbol{g}\tilde{B}P\tilde{B}\boldsymbol{g}^T| + 2|\boldsymbol{g}\tilde{B}P\boldsymbol{h}^T| + |\boldsymbol{h}P\boldsymbol{h}^T| = |\langle \boldsymbol{g}\tilde{B}P, \boldsymbol{g}\tilde{B}\rangle| + 2|\langle \boldsymbol{g}\tilde{B}, \boldsymbol{h}\rangle| + |\boldsymbol{h}P\boldsymbol{h}^T|.$$

We note that P is a permutation matrix, and so a contraction. Furthermore, since H is an (m, d, β)-graph, $||\boldsymbol{v}\hat{B}||_2 \leq \beta||\boldsymbol{v}||_2$ for any $\boldsymbol{v} \perp \boldsymbol{u}$ and so, since g sums to zero on every cloud

$$||\boldsymbol{g}\tilde{B}||_2 \leq \beta||\boldsymbol{g}||_2.$$

Hence, using Cauchy-Schwartz, it follows that

$$\begin{aligned}
|\boldsymbol{f}Z\boldsymbol{f}^T| &\leq |\langle \boldsymbol{g}\tilde{B}P, \boldsymbol{g}\tilde{B}\rangle| + 2|\langle \boldsymbol{g}\tilde{B}, \boldsymbol{h}\rangle| + |\boldsymbol{h}P\boldsymbol{h}^T| \\
&\leq ||\boldsymbol{g}\tilde{B}P||_2||\boldsymbol{g}\tilde{B}||_2 + ||\boldsymbol{g}\tilde{B}||_2||\boldsymbol{h}||_2 + |\boldsymbol{h}P\boldsymbol{h}^T| \\
&\leq ||\boldsymbol{g}\tilde{B}||_2||\boldsymbol{g}\tilde{B}||_2 + ||\boldsymbol{g}\tilde{B}||_2||\boldsymbol{h}||_2 + |\boldsymbol{h}P\boldsymbol{h}^T| \\
&\leq \beta^2||\boldsymbol{g}||_2^2 + 2\beta||\boldsymbol{g}||_2||\boldsymbol{h}||_2 + |\boldsymbol{h}P\boldsymbol{h}^T|.
\end{aligned}$$

Finally, for the last term, if we define the function $w\colon V(G) \to \mathbb{R}$ by $w(i) = \sqrt{m}h(v, i)$ then we have that $||\boldsymbol{w}||_2^2 = ||\boldsymbol{h}||_2^2$ and the definition of P leads to

$$\boldsymbol{h}P\boldsymbol{h}^T = \boldsymbol{w}\hat{A}\boldsymbol{w}^T,$$

where $\hat{A} = \hat{A}(G)$ is the transition matrix of the random walk on G. However, since $\boldsymbol{f} \perp \boldsymbol{u}$ by assumption, it follows that $\boldsymbol{h} \perp \boldsymbol{u}$ and so $\boldsymbol{w} \perp \boldsymbol{u}$. In particular,

$$\boldsymbol{h}P\boldsymbol{h}^T = \boldsymbol{w}\hat{A}\boldsymbol{w}^T \leq \alpha||\boldsymbol{w}||_2^2 = \alpha||\boldsymbol{h}||_2^2.$$

Hence we can conclude that

$$|\boldsymbol{f}Z\boldsymbol{f}^T| \leq \beta^2||\boldsymbol{g}||_2^2 + 2\beta||\boldsymbol{g}||_2||\boldsymbol{h}||_2 + \alpha||\boldsymbol{h}||_2^2,$$

and since $||\boldsymbol{g}||_2^2, ||\boldsymbol{h}||_2^2 \leq ||\boldsymbol{f}||_2^2$, it is immediate that

$$\begin{aligned}
|\boldsymbol{f}Z\boldsymbol{f}^T| &\leq \beta^2||\boldsymbol{g}||_2^2 + 2\beta||\boldsymbol{g}||_2||\boldsymbol{h}||_2 + \alpha||\boldsymbol{h}||_2^2 \\
&\leq ||\boldsymbol{f}||_2^2(\alpha + \beta + \beta^2).
\end{aligned} \tag{9.1}$$

In fact, an even better bound can be given by noting that $\langle \boldsymbol{h}, \boldsymbol{g}\rangle = 0$ and so $||\boldsymbol{f}||_2^2 = ||\boldsymbol{g}||_2^2 + ||\boldsymbol{h}||_2^2$, and so we're really trying to minimise the quadratic form $\begin{pmatrix} \alpha & \beta \\ \beta & \beta^2 \end{pmatrix}$. However, this still

isn't sufficient to prove (2), for which we need to more carefully consider the relationship between $\boldsymbol{g}$ and $\boldsymbol{h}$.

We note that (9.1) is then sufficient to prove (1) when $||\boldsymbol{g}||$ is sufficiently small. Indeed, if $||\boldsymbol{g}||_2^2 \leq \frac{1-\alpha}{3\beta}||\boldsymbol{f}||_2$ then

$$\begin{aligned}
|\boldsymbol{f}Z\boldsymbol{f}^T| &\leq \beta^2||\boldsymbol{g}||_2^2 + 2\beta||\boldsymbol{g}||_2||\boldsymbol{h}||_2 + \alpha||\boldsymbol{h}||_2^2 \\
&\leq \beta^2\left(\frac{1-\alpha}{3\beta}\right)^2||\boldsymbol{f}||_2^2 + 2\beta\frac{1-\alpha}{3\beta}||\boldsymbol{f}||_2 + \alpha||\boldsymbol{f}||_2^2 \\
&\leq ||\boldsymbol{f}||_2^2\Big(\alpha + \frac{2}{3}(1-\alpha) + \frac{1}{9}(1-\alpha)^2\Big) \\
&= ||\boldsymbol{f}||_2^2(\frac{7}{9} + \frac{\alpha^2}{9} - \frac{8}{9}\alpha\Big) \\
&< ||\boldsymbol{f}||_2^2
\end{aligned}$$

since $\alpha < 1$.

On the other hand, if $||\boldsymbol{g}||_2$ is large, in that $||\boldsymbol{g}||_2^2 > \frac{1-\alpha}{3\beta}||\boldsymbol{f}||_2$, then we go back to the observation that

$$\boldsymbol{f}Z\boldsymbol{f}^T = \langle \boldsymbol{f}\tilde{B}P\tilde{B}, \boldsymbol{f}\rangle = \langle (\boldsymbol{g}+\boldsymbol{h})\tilde{B}P, (\boldsymbol{g}+\boldsymbol{h})\tilde{B}\rangle = \langle (\boldsymbol{g}\tilde{B}+\boldsymbol{h})P, \boldsymbol{g}\tilde{B}+\boldsymbol{h}\rangle$$

and use the fact that P is a contraction and $\langle \boldsymbol{g}\tilde{B}, \boldsymbol{h}\rangle = \langle \boldsymbol{g}, \boldsymbol{h}\tilde{B}\rangle = \langle \boldsymbol{g}, \boldsymbol{h}\rangle = 0$ to see that

$$\begin{aligned}
|\boldsymbol{f}M\boldsymbol{f}^T| &= |\langle (\boldsymbol{g}\tilde{B}+\boldsymbol{h})P, \boldsymbol{g}\tilde{B}+\boldsymbol{h}\rangle| \\
&\leq ||\boldsymbol{g}\tilde{B}+\boldsymbol{h}||_2^2 \\
&= ||\boldsymbol{g}\tilde{B}||_2^2 + ||\boldsymbol{h}||_2^2 \\
&\leq \beta^2||\boldsymbol{g}||_2^2 + ||\boldsymbol{f}||_2^2 - ||\boldsymbol{g}||_2^2 \\
&\leq ||\boldsymbol{f}||_2^2\left(1 - (1-\beta^2)\left(\frac{1-\alpha}{3\beta}\right)^2\right) \\
&< ||\boldsymbol{f}||_2^2.
\end{aligned}$$

Instead of (3) we will prove the slightly weaker bound

$$\phi(\alpha, \beta) \leq 1 - (1-\alpha)(1-\beta)^2.$$

In order to do so, rather than trying to bound Rayleigh quotient of an arbitrary vector $\boldsymbol{f}$ by decomposing $\boldsymbol{f}$ in a sensible manner, we will instead try to bound the operator norm of Z by decomposing Z in a sensible manner.

To do so, we will need the following useful lemma.

Lemma 9.2. *Let $\hat{A}$ be the transition matrix of an (n, d, λ)-graph and let J be a matrix with every entry equal to $\frac{1}{n}$. Then $\hat{A} = (1-\lambda)J + \lambda C$ where the operator norm $||C||$ of C satisfies*

$$\sup_{||v||_2=1} ||\boldsymbol{v}C||_2 := ||C|| \leq 1.$$

Proof. Rearranging gives us that $C = \frac{\hat{A}-(1-\lambda)J}{\lambda}$. Let $\boldsymbol{u}$ be the uniform distribution, then $\boldsymbol{u}$ is an eigenvector of both $\hat{A}$ and J with eigenvalue 1, and hence also of C. Given any $\boldsymbol{v} \perp \boldsymbol{u}$, it follows $\boldsymbol{v}\hat{A}$ and $\boldsymbol{v}J$ are also $\perp \boldsymbol{u}$, and hence so is $\boldsymbol{v}C$.

Hence, it suffices to bound the norm of $\boldsymbol{v}C$ for all $\boldsymbol{v} \perp \boldsymbol{u}$. For this, we note that $\boldsymbol{v}J = 0$ and $||\boldsymbol{v}\hat{A}||_2 \leq \lambda ||\boldsymbol{v}||_2$ and so

$$||\boldsymbol{v}C||_2 = \left\| \frac{\boldsymbol{v}\hat{A} - (1-\lambda)\boldsymbol{v}J}{\lambda} \right\|_2 = \left\| \frac{\boldsymbol{v}\hat{A}}{\lambda} \right\|_2 \leq ||\boldsymbol{v}||_2.$$

□

We can think of the above lemma as saying that a step in the random walk given by $\hat{A}$ can be viewed as being a convex combination of a truly random step, together with a contraction.

So, returning to the analysis of the zig-zag product, recall that $Z = \tilde{B}P\tilde{B}$, where $\tilde{B}$ is Kroenecker product of the transition matrix $\hat{B}$ for H and I_n. By Lemma 9.2, since H is an (m, d, β)-graph, we can write

$$\hat{B} = (1-\beta)J + \beta E$$

where J is the all $\frac{1}{m}$ matrix and E has operator norm at most one. It follows that

$$\tilde{B} = (1-\beta)\tilde{J} + \beta\tilde{E}$$

where the tildes correspond as before to Kroenecker products. It is easy to see that the operator norm of $\tilde{E}$ is still at most one.

Then we can expand as before

$$\begin{aligned} Z &= \tilde{B}P\tilde{B} \\ &= ((1-\beta)\tilde{J} + \beta\tilde{E})P((1-\beta)\tilde{J} + \beta\tilde{E}) \\ &= (1-\beta)^2\tilde{J}P\tilde{J} + \beta^2\tilde{E}P\tilde{E} + \beta(1-\beta)\big(\tilde{J}P\tilde{E} + \tilde{E}P\tilde{J}\big) \\ &:= (1-\beta)^2\tilde{J}P\tilde{J} + F \end{aligned}$$

where we absorb the three last terms into F. We first claim that $||F|| \leq 2\beta - \beta^2$.

Indeed,

$$||F|| = \left\| \beta^2\tilde{E}P\tilde{E} + \beta(1-\beta)\big(\tilde{J}P\tilde{E} + \tilde{E}P\tilde{J}\big) \right\| \leq \beta^2 + 2\beta(1-\beta) = 2\beta - \beta^2,$$

where we used the fact that P as a permutation matrix, $\tilde{E}$ by assumption and J by observation have operator norm ≤ 1.

Finally we claim that $\tilde{J}P\tilde{J} = \hat{A} \times J$, where $\hat{A}$ is the transition matrix of G. Indeed, if we think about them both as transition matrices acting on $V(G) \times V(H)$. For the first, if we start at some point (v, a) then we first choose a' uniformly from $V(H)$, then transition to the vertex (w, b') such that the edge labelled b' at w is the edge labelled a' at v, and then choose b uniformly from $V(H)$ and go to (w, b). The second corresponds to the following set of transitions - starting at (v, a) we choose a random neighbour w of v in G, we choose b uniformly from $V(H)$ and we go to (w, b). However, by the definition of the replacement product, these two processes are the same.

Hence we see that

$$Z = (1-\beta)^2\hat{A} \times J + (2\beta - \beta^2)F$$

However, the eigenvalues of $X \times Y$ are given by the products $\lambda_i \mu_j$ where λ_i and μ_j are the eigenvalues of X and Y (with multiplicity). Since J has spectrum $1, 0, \ldots$ it follows that the $\lambda(\hat{A} \times J) = \lambda(\hat{A}) = \alpha$ and hence, since all the matrices involved have $\boldsymbol{u}$ as their largest eigenvector,

$$\lambda(Z) \leq (1-\beta)^2 \lambda(\hat{A}) + (2\beta - \beta^2)\lambda(F) = (1-\beta)^2\alpha + 2\beta - \beta^2 = 1 - (1-\beta)^2(1-\alpha).$$

□

9.4 Construction of an expander family using the zig-zag product

We start by taking some small expander H, which will be a $(d^4, d, \frac{1}{5})$-graph for some constant d. The existence of a such a graph is not hard to show using for example Theorem 7.5, and since d is constant, algorithmically we can find the graph in constant time by a brute force search.

We then inductively define a sequence of graph G_n as follows

$$G_1 = H^2, \qquad G_{n+1} = (G_n)^2 \textcircled{z} H \text{ for } n \geq 1.$$

Proposition 9.3. *G_n is a $(d^{4n}, d^2, \frac{1}{2})$-graph for all n.*

Proof. For $n = 1$ this is clear since H^2 is even a $(d^4, d^2, \frac{1}{25})$-graph. We proceed by induction.

Note that $(G_n)^2$ is a (d^{4n}, d^4)-graph and H is a (d^4, d)-graph, and so $(G_n)^2 \textcircled{z} H$ is well-defined and is a (d^{4n+1}, d^2)-graph. Furthermore, by Theorem 9.1 (2) (and in fact, even by the weaker version we proved) $(G_n)^2 \textcircled{z} H$ is an (d^{4n+1}, d^2, γ)-graph, where

$$\gamma \leq \left(\frac{1}{2}\right)^2 + \frac{1}{5} + \frac{1}{25} = \frac{1}{2}.$$

□

We note however that this construction is only mildly explicit. However, by interleaving the construction with a further graph product, tensoring G_n with itself, one can obtain a family which grow much faster and which is then strongly explicit.

9.5 An application to complexity theory : $SL = L$

Roughly, without going into too much detail, we are interested in problems that can be solved with low 'space complexity', without storing too much information. In particular, we are interested in whether the problem USTCON of determining whether there is a path between two given vertices s and t of a graph G can be solved in this manner.

One reason why this problem is interesting is that it can be shown that a large class of problems, known as SL, have a log-space reduction (a reduction using only logarithmic space) to USTCON. Hence, showing that USTCON can be solved in logarithmic space implies that this 'heirachy' collapses, and all these problems can be solved in logarithmic space.

There are many well-known and efficient algorithms in terms of their time-complexity for determining graph connectivity, which run in linear time, such as DFS algorithms, but these also can require linear space to run (for example maintaining a stack of vertices). In fact, an old result of Savitch shows that there is a $\log^2$ space algorithm which solves this problem.

More recently an Aleluinas, Karp, Lipton, Lovász and Rackoff gave a probabilistic logspace algorithm that solves USTCON. The algorithmic is particularly simple, we perform a random walk on G of a polynomial length p starting at s, and see if it ever reaches t.

In fact, not only is the algorithm simple, but the analysis is also relatively simple. Indeed, the algorithm clearly only uses logarithmic space - all we need to remember is the current position of the walk, how far we've walked and the goal vertex t, which can all be encoded in logarithmically many bits. Note also that this algorithm is still reasonably time efficient as well, running in polynomial time,

So, we need to show that for an appropriate choice of p the probability that the random walk doesn't meet t is sufficiently small (in fact, vanishingly small). However, standard arguments for random walks imply that in a connected graph $G = (V, E)$, the expected time it takes a random walk starting at a vertex s to hit any vertex t is at most $2|E||V| \leq n^3$. In particular, by Markov's inequality, the probability that the random walk doesn't hit t in the first $2n^3$ steps is at most $\frac{1}{2}$. In particular, if we repeat this process n many times, resulting in a walk of length $p = O(n^5)$, then by the Markov property the probability that we don't hit t in each segment of the walk is at most $\frac{1}{2}$ independently, and hence the probability that we never hit t is at most 2^{-n}.

Alternatively, one can analyse this algorithm by first replacing each vertex v in the graph with a cycle of length $d(v)$, to make the graph regular whilst preserving connectivity, and using the easy to show fact that in any connected 3-regular graph the second largest eigengvalue satisfies

$$\lambda_2 \leq 3 - \Omega\left(\frac{1}{n\text{diam}(G)}\right) \leq 3 - \Omega\left(\frac{1}{n^2}\right).$$

Our analysis of the random walk in G in Theorem 4.3 then implies that a random walk of length $O(n^3)$ in G results in a distribution exponentially close to uniform, and hence visits each vertex with reasonably large probability. Repeating such a walk polynomial many times will then end up visiting each vertex whp.

A natural approach then is to try to *derandomise* this algorithm. In particular, the hope would be to generate deterministically a walk which must explore all the vertices in a connected graph. Various people tried to implement this idea without success, until Reingold came up with an ingenious solution based on the zig-zag product.

As in the previous paragraph, we can assume that G is D-regular for some constant D (we can even add self loops to raise this regularity if we wish). If it were the case that G were an expander, then it would have a logarithmic diameter. Hence, since the degrees are bounded, we could simply enumerate all of the logarithmically long paths starting at s and check if one of them arrives at t. Since there are only polynomially many such paths, we can keep track of this process in logarithmic space. Hence, for expander graphs, the logspace alogorithm is trivial.

However, it's not clear how we can find a log-space reduction of the problem to the same problem over a graph G which is a bounded degree expander. The idea is to use the zig-zag

product to increase expansion, without increasing the vertex degree too dramatically.

As in the previous argument, we can assume that G is an (n, D, α)-graph where $\alpha \leq 1 - \Omega\left(\frac{1}{n^2}\right)$, and let us suppose that $D = d^{16}$ and that we have to hand a $(d^{16}, d, \frac{1}{2})$-graph H. We inductively construct a sequence of graphs G_i such that

$$G_1 = G, \qquad \text{and} \qquad G_{i+1} = (G_i \circledS H)^8.$$

Note that, as long as G_i is an (m, d^{16})-graph for some m, then $G_i \circledS H$ is an (md^{16}, d^2)-graph, and so G_{i+1} is an (md^{16}, d^{16})-graph.

We terminate this sequence after $k = O(\log n)$ steps. Since, in each step the size of G increases by some constant factor $D = d^{16}$, it follows that G_k is an (nD^k, D)-graph. We would like to show

- Neighbourhood queries for G_k can be answered in logarithmic space;
- G_k is an (nD^k, D, γ)-graph for some constant $\gamma < 1$.

The first claim is not at all obvious, since we're only working in logarithmic space, we can't keep track of each graph G_i, and so we have to evaluate the whole recursion 'locally' for each query. But, it turns out that each of the steps can be performed with only an additional constant amount of space. The proof of this however involves the construction of a clever data structure which we will not go into.

To prove the second claim, we use Theorem 9.1 (3)

$$\phi(\alpha, \beta) \leq 1 - (1 - \beta^2)\frac{1 - \alpha}{2}.$$

To begin with we have that G_1 has an expansion ratio $\alpha \leq 1 - \Omega\left(\frac{1}{n^2}\right)$, and we will see that this roughly squares in each iteration, and hence reaches a constant after logarithmically many iterations.

If we write λ_i and μ_i for the normalised generalised second eigenvalue of G_i and $G_i \circledS H$, respectively. Then, Theorem 9.1 (3) gives

$$\mu_i \leq 1 - \left(1 - \frac{1}{4}\right)\frac{1 - \lambda_i}{2} = 1 - \frac{3(1 - \lambda_i)}{8},$$

and so

$$\lambda_{i+1} = \mu_i^8 \leq \left(1 - \frac{3(1 - \lambda_i)}{8}\right)^8.$$

If $\lambda_i \leq \frac{1}{2}$, then $\lambda_{i+1} \leq \left(\frac{13}{16}\right)^8 < \frac{1}{2}$, whereas if $x \in \left[\frac{1}{2}, 1\right]$ then it can be checked that

$$\left(1 - \frac{3(1 - x)}{8}\right)^4 \leq x,$$

and so $\lambda_{i+1} \leq \max\left\{\frac{1}{2}, \lambda_i^2\right\}$. Hence

$$\lambda_k \leq \max\left\{\frac{1}{2}, \left(1 - \Omega\left(\frac{1}{n^2}\right)\right)^{2^k}\right\} \leq \max\left\{\frac{1}{2}, \exp\left(-\Omega\left(\frac{2^k}{n^2}\right)\right)\right\} \leq \frac{1}{2}$$

If $k = O(\log n)$ for a large enough leading constant.

10 Lossless conductors and expanders

In the course we have mostly focused on edge-expansion, and in particular the Cheeger constant $h(G)$ and the edge isoperimetric parameter

$$\Phi_E(G,k) = \min\{\partial_E(S) \colon S \subseteq V, |S| = k\}.$$

In general it seems harder to control the vertex isoperimetric parameter

$$\Phi_V(G,k) = \min\{\partial_V(S) \colon S \subseteq V, |S| = k\}.$$

We saw earlier that in 'most' (n,d)-graphs linear size sets expand almost optimally, and so for any δ, $\Phi_V(G,\epsilon n) \geq d - 2 - \delta$ for sufficiently small ϵ. However our constructions of 'good' edge-expanders, or at least those with optimal spectral expansion, i.e. the Ramanujan graphs, do not lead to such optimal vertex expansion.

Indeed, we showed on the example sheet that an (n,d,α)-graph satisfies

$$\Phi'_V(G,\rho n) = \min_{S \subseteq V} \frac{|\Gamma(S)|}{|S|} \geq \frac{1}{\rho(1-\alpha^2)+\alpha^2}$$

and since, by Theorem 6.2, $\alpha = \frac{\lambda_2}{d} \geq \frac{2+o_d(1)}{\sqrt{d}}$, it follows that for Ramanujan graphs we get the following bound

$$\Phi'_V(G,\rho n) \geq (1 + o_{\rho,d}(1))\frac{d}{4}.$$

In fact, Kahale gave a better bound which improves this bound for Ramanujan graphs to

$$\Phi'_V(G,\rho n) \geq (1 + o_{\rho,d}(1))\frac{d}{2},$$

and gave some constructions, which make small local adaptations to a Ramanujan graph which reduce the vertex expansion without significantly increasing λ_2, to show that this is in fact tight. In particular, even with our constructions of optimal edge-expanders, we are not guaranteed to get a better lower bound on $\Phi_V(G,\epsilon n)$ than around $\frac{d}{2}$.

However, it turns out that in a variety of applications it would be useful to have (n,d)-graphs which are 'good' vertex expanders, in the sense that small sets expand by a factor of γd for some $\gamma > \frac{1}{2}$, for example in the construction of expander based linear codes.

It is still a major open problem to construct such families of graphs, but in this section we will present a construction that takes a step in this direction, building at least bipartite graphs where the 'left hand side' has good expansion, that is, for any $\delta > 0$ we can construct a family of bipartite graphs which are d-left-regular and such that every subset of the left hand side of small linear size expands by a factor of $(1-\delta)d$.

As part of this we will introduce a variety of useful and interesting notions coming from the area of *randomness enhancing objects* such as conductors and extractors.

10.1 Conductors and lossless expanders

A key idea in this section will be to consider the notion of entropy introduced earlier in the course. Previously we saw how the random walk on a graph G in some sense 'tranforms' distributions

on $V(G)$ and considered the effect this has on the entropy of the distributions. In this section we will be considering a bipartite analogue of this - given a bipartite graph $G = (L, R, E)$ and a distribution $\boldsymbol{p}$ on L, a single step in the random walk on G will transform $\boldsymbol{p}$ to a distribution $\boldsymbol{q}$ on R. In particular we will be interested in bounding the entropy of $\boldsymbol{q}$ from below in terms of the entropy of $\boldsymbol{p}$, and in which graphs can we do so effectively.

Recall the notion of *min-entropy*

$$H_\infty(\boldsymbol{p}) = -\log(||\boldsymbol{p}||_\infty)$$

we defined earlier in the course, which we can think of as a measure of distance from $\boldsymbol{p}$ to a uniform distribution. Note that if $\boldsymbol{p}$ is a distribution on $\{0,1\}^n$ then $0 \leq H_\infty(\boldsymbol{p}) \leq \log n$. We define a *$k$-source* to be a distribution with min-entropy at least k.

Bounding the min-entropy from below if quite a strong condition on $\boldsymbol{p}$, $H_\infty(\boldsymbol{p}) \geq k$ if and only if no point in the space has probability more than 2^{-k} (whereas if we have a similar bound on the Shannon entropy then it is only true that this holds 'on average'). So, it will be useful to weaken the notion of having large min-entropy to that of being 'close' (in total variational distance) to a distribution with high min-entropy. To this end we say a distribution $\boldsymbol{p}$ is a *(k, ϵ)-source* if there is some k-source $\boldsymbol{q}$ such that $||\boldsymbol{p} - \boldsymbol{q}||_1 \leq 2\epsilon$, in which case we say that $\boldsymbol{p}$ and $\boldsymbol{q}$ are *ϵ-close*.

As is standard in the area, and to facilitate our analysis of the entropy, we will use bit-strings to label the vertices and edges of the graph G. For ease of notation we will assume that $|L| = N = 2^n$, $|R| = M = 2^m$ and that each vertex in L has degree $D = 2^d$. Then we can view a graph $G = (L, R, E)$ as a function

$$E\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$$

where the vertex $x \in L$ is connected to the $D = 2^d$ vertices $E(x, \cdot) \in R$.

We say that a D-left-regular graph $G = (L, R, E)$ is a *(K_{max}, ϵ)-lossless expander* if every set $S \subset L$ of size $|S| = K \leq K_{\max}$ has at least $(1 - \epsilon)DK$ neighbours in R. In other words, sufficiently small vertex sets on the left have almost the maximal possible vertex expansion. Since G is D-left-regular, an alternative view is that most of the neighbours of the set S have a unique neighbour in S. Naturally we must have that $K_{\max}$ is somewhat smaller than $\frac{M}{D}$.

From the point of view of the graph as a function we see that the following is a stronger condition: A function $E\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a *(k_{max}, ϵ)-lossless conductor* if for any $k \leq k_{\max}$ and any k-source $\boldsymbol{p}$ over $\{0,1\}^n$, the distribution $E(\boldsymbol{p}, \boldsymbol{u_d})$ is a $(k + d, \epsilon)$-source, where $\boldsymbol{u_d}$ is the uniform distribution over $\{0,1\}^d$. Indeed, it is simple to show that if G is a graph such that the function E is a $(k_{\max}, \epsilon)$-lossless conductor then G is a $(K_{\max}, \epsilon)$-lossless expander, where $K_{\max} = 2^{k_{\max}}$.

If we view the process described above as taking some distribution $\boldsymbol{p}$ and 'injecting' some randomness into it via the choice of a random edge in $\{0,1\}^d$, then the idea behind a lossless conductor is that *none* of the additional entropy added in this step should be lost, up to a small ℓ_1 pertubation.

Hence the existence of the required expanders will be guaranteed by the following theorem.

Theorem 10.1. *For any $\epsilon > 0$ and $m \leq n$ there exists an explicit family of $\left(m - d - \log\left(\frac{1}{\epsilon}\right) - O(1), \epsilon\right)$-lossless conductors, where $d = O\left(n - m + \log\left(\frac{1}{\epsilon}\right)\right)$.*

Leading to the following corollary.

Corollary 10.2. *For any $\epsilon > 0$ and $M \leq N$ there exists an explicit family of D-left-regular bipartite graphs that are $\left(\Omega\left(\frac{\epsilon M}{D}\right), \epsilon\right)$-lossless expanders, where $D \leq \left(\frac{N}{\epsilon M}\right)^c$ for some constant c.*

Note that, in the (useful) case where $\frac{M}{N}$ and ϵ are bounded below by a constant, the degree D will also be constant.

In order to prove Theorem 10.1 we will need to introduce a number of randomness enhancing objects. The first is a slight generalisation of a lossless conductor.

A function $E\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a (k_{max}, a, ϵ)-*conductor* if for any $k \leq k_{\max}$ and any k-source $\boldsymbol{p}$ over $\{0,1\}^n$, the distribution $E(\boldsymbol{p}, \boldsymbol{u_d})$ is a $(k + a, \epsilon)$-source. The idea here is that the conductor will 'conduct' at least a bits of entropy from the d random bits coming from $\boldsymbol{u_d}$, as long as the the entropy of the input $\boldsymbol{p}$ is not too large.

Obviously the above cannot hold when $k_{\max}$ is too large in terms of m, since the output can be at most an m-source. In the optimal case we say a function $E\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is an (a, ϵ)-*extracting conductor* if it is an $(m - a, a, \epsilon)$-conductor. The interesting thing here is that not only does this function 'conduct' a bits of entropy from $\boldsymbol{u_d}$, but once the entropy of the input $\boldsymbol{p}$ gets large enough we find that the output has to be ϵ-close to uniform! Functions with this second property are known in the literature as extractors, the idea being that they can 'extract' a uniformly random output from a weakly random source $\boldsymbol{p}$ together with a small random seed $\boldsymbol{u_d}$.

We say a pair of functions $\langle E, C\rangle\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m \times \{0,1\}^b$ is an (k_{max}, a, ϵ)-*buffer conductor* if $E\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is an (a, ϵ)-extracting conductor and $\langle E, C\rangle$ is a $(k_{\max}, \epsilon)$-lossless conductor. The idea here is to ensure that none of the added entropy from $\boldsymbol{u_d}$ is lost - whatever entropy is not gained in the first function can be saved completely by the second, which we view as an overflow buffer or bucket.

Finally, a pair of functions $\langle E, C\rangle\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m \times \{0,1\}^b$ where $n + d = m + b$ is an (k_{max}, a, ϵ)-*permutation conductor* if $E\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is an $(k_{\max}, a, \epsilon)$-extracting conductor and $\langle E, C\rangle$ is a permutation of $\{0,1\}^{n+d}$ (i.e., it is a bijection). We note that this is in fact a special case of a buffer conductor, where the fact that $\langle E, C\rangle$ is a $(k_{\max}, \epsilon)$-lossless conductor follows since $\langle E, C\rangle$ is a permutation.

10.2 The construction

10.3 The zig-zag product for bipartite graphs

We will construct our family of lossless expanders using a generalisation of the zig-zag product for conductors. However, in order to define this product it will be instructive to adapt slightly the zig-zag product to bipartite graphs.

Let H be a d-regular bipartite graph with s vertices in its two partition classes L_H and R_H and let G be an s-regular bipartite graph with n vertices in its two partition classes L_G and R_G.

The zig-zag product $G \textcircled{z} H$ (in a slight abuse of notation) is a d^2 regular bipartite graph with sn vertices in each partition class, which are given by $L_G \times L_H$ and $R_G \times R_H$. Whilst these are not vertices in our graph, it will be useful to imagine the vertex set as being a subset of $L_G \times V(H) \cup R_G \times V(H)$, that is a copy of H for each vertex of G.

The edges emanating from a vertex $(x, y) \in L_G \times L_H$ in the left partition class are labelled by pairs in $[d] \times [d]$ in the following way: We imagine that the edges emanating from each vertex in G and H have been labelled by $[s]$ and $[d]$ (independently for both endpoints). The other endpoint (u, v) of the edge labelled (a, b) at (x, y) is determined as follows:

- Let v'' be the neighbour of y in H labelled a. We think of this as taking a left to right step in the 'local copy of H';
- We let u be the neighbour of x labelled v'' in G and let v' be the label of the edge to x from u. We think of this as taking a left right step along an edge of G between the copies of H;
- Finally let v be the neighbour of v' in H labelled b. Again we think of this as taking a left to right step in the 'local copy of H'.

It is perhaps reasonable to hope that, as in the previous section, if G and H are good edge-expanders, then also $G \textcircled{z} H$ will be. Furthermore, as before the degree of $G \textcircled{z} H$ is controlled solely by the degrees in H, making this perhaps a good construction to use for building bounded degree expanders. However, whilst the degree of $G \textcircled{z} H$ is d^2, it is easy to see that its vertex expansion cannot exceed d, by considering the neighbourhood of a single copy of H on the left.

In order to explain how we might hope to improve this, ignoring for now the question of bipartiteness, let us first note that we can view a single step in a random walk on an expander graph as a type of permutation conductor. Indeed, suppose that G is an (N, M, α)-graph, where $N = 2^n$ and $M = 2^m$ and we have some arbitrary labelling over $\{0,1\}^m$ of the edges at each vertex. Then a step in a random walk in G, starting at a vertex x, chooses some uniform label $e \in \{0,1\}^m$ and moves to the other endpoint of the edge labelled e at x.

Another way to think about this is as the function $\langle E, C \rangle \colon \{0,1\}^n \times \{0,1\}^m \to \{0,1\}^n \times \{0,1\}^m$ which takes a pair (x, e) to the pair (x', e') where x' is the other endpoint of the edge labelled e and e' is the label of this edge at x'. This is sometimes called the *rotation map* of G. In this way if we have a distribution $\boldsymbol{p}$ on the vertices of G, $\boldsymbol{u}$ is the unifrom distribution on $\{0,1\}^m$ and $E(\boldsymbol{p}, \boldsymbol{u}) = (\boldsymbol{q}, \boldsymbol{r})$, then $\boldsymbol{q}$ is the distribution on $V(G)$ after one step in the random walk.

It is clear that this map is a permutation, however the expansion of the underlying graph guarantees that E is a conductor, which can be shown using the equivalence between min-entropy and Rényi entropy

A bit of thought shows that, if we have an (N, M)-graph G and an (M, D)-graph H, where $D = 2^d$, we can view a single step in the random walk on the zig-zag product $G \textcircled{z} H$ as coming from a combination of three permutation conductors, two arise from the rotation map $\langle E_H, C_H \rangle$ of H and one from the rotation map $\langle E_G, C_G \rangle$ G.

Indeed, if we start with some vertex $(v, a) \in V(G) \times V(H) = V(G \textcircled{z} H)$ then the random

walk first chooses uniformly some $e \in \{0,1\}^d$ and considers the endpoints a' of the edge labelled e at a in H, it then considers the endpoint w of the edge labelled a' at v in G, and the label a'' of this same edge at w. It finally chooses uniformly some $e; \in \{0,1\}^d$ and considers the endpoints b of the edge labelled e' at a'' in H, and moves to the vertex (w, b).

So, if we start with some distribution $(\boldsymbol{x_1}, \boldsymbol{x_2})$ on $V(G) \times V(H)$ then the distribution after one step in the random walk can be caluclated as follows: Let $\boldsymbol{r_1}$ and $\boldsymbol{r_2}$ be uniformly distributed on $\{0,1\}^d$

- We let $(\boldsymbol{y_2}, \boldsymbol{z_2}) = \langle E_H, C_H \rangle (\boldsymbol{x_2}, \boldsymbol{r_1})$;
- We let $(\boldsymbol{y_1}, \boldsymbol{z_1}) = \langle E_G, C_G \rangle (\boldsymbol{x_1}, \boldsymbol{y_2})$;
- We let $(\boldsymbol{y_3}, \boldsymbol{z_3}) = \langle E_H, C_H \rangle (\boldsymbol{z_1}, \boldsymbol{r_2})$;

and the final distribution is given by $(\boldsymbol{y_1}, \boldsymbol{y_2})$. Here we can see that, even though each of the permutation conductors are 'lossless', some of the outputs, namely $\boldsymbol{z_2}$ and $\boldsymbol{z_3}$ are not used, and so some of the entropy inputted by the random seeds $\boldsymbol{r_1}$ and $\boldsymbol{r_2}$ might not be 'conducted' through to the output $(\boldsymbol{y_1}, \boldsymbol{y_2})$.

Indeed, when we start with a distribution $\boldsymbol{x_2}$ which is uniform on H, then the first random step cannot conduct any extra entropy to $\boldsymbol{y_2}$, and so all of this injected randomness is lost in $\boldsymbol{z_2}$.

Broadly the idea will be to try to 'save' this randomness for later by using the fact that the permutation conductor is a buffer conductor, and to avoid losing the extra randomness in $\boldsymbol{z_2}, \boldsymbol{z_3}$ by replacing our second step in H with a lossless conductor. This idea will be formalised in the notion of a zig-zag product for conductors.

10.4 The zig-zag product for conductors

The zig-zag product for conductors will have to be carried out with quite carefully chosen parameters in order to give the desired constant degree lossless expanders that we wish for in Corollary 10.2. Explicitly one can show using probabilistic methods the existence of the following objects.

Lemma 10.3.

1. *A (k, ϵ)-lossless conductor $E\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ exists for all $m > k + d + \log \frac{1}{\epsilon}$.*
2. *An (a, ϵ)-extracting conductor $E\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ exists for all $d > a + 2\log \frac{1}{\epsilon}$.*
3. *An (n, a, ϵ)-buffer conductor $E\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m \times \{0,1\}^b$ exists for all $d > a + 2\log \frac{1}{\epsilon}$ and $m + b > n + d + \log \frac{1}{\epsilon}$.*

Further, it can be checked that the rotation map of a large expanding graph leads to a permutation conductor. Very roughly, suppose we have an (N, D, α)-graph, where $N = 2^n$ and $D = 2^d$ then the rotation map will give an $(n - O(a), a, \epsilon)$-extractor where $d = O\left(a + \log \frac{1}{\epsilon}\right)$.

Our construction then uses three parts $\langle E_1, C_1\rangle, \langle E_2, C_2\rangle$ and E_3 (which in some sense take the role of the three different steps in the zig-zag product).

a) An (k_1, a_1, ϵ)-permutation conductor $\langle E_1, C_1\rangle \colon \{0,1\}^{n_1} \times \{0,1\}^{d_1} \to \{0,1\}^{m_1} \times \{0,1\}^{b_1}$;

b) An (n_2, a_2, ϵ)-buffer conductor $\langle E_2, C_2\rangle \colon \{0,1\}^{n_2} \times \{0,1\}^{d_2} \to \{0,1\}^{d_1} \times \{0,1\}^{b_2}$;

c) An (k_2, ϵ)-lossless conductor $E_3 \colon \{0,1\}^{b_1+b_2} \times \{0,1\}^{d_3} \to \{0,1\}^{m_3}$.

By combining them in an appropriate manner, we will produce a conductor $E \colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ where $n = n_1 + n_2$, $d = d_2 + d_3$ and $m = m_1 + m_3$.

How do we compute the output of E? Well, given a pair $(x, r) \in \{0,1\}^n \times \{0,1\}^d$ let us split x into x_1 and x_2 of length n_1 and n_2 and r into r_2 and r_3 of length d_2 and d_3. Then the output $E(x_1x_2, r_2r_3)$ is given by y_1y_3 which we compute in three steps:

i) We first compute $\langle E_2, C_2\rangle(x_2, r_2) = (y_2, z_2)$;

ii) Next we compute $\langle E_1, C_1\rangle(x_1, y_2) = (y_1, z_1)$;

iii) Finally we compute $y_3 = E_3(z_1z_2, r_3)$.

Let us illustrate why we might hope this to work via an informal 'bookkeeping' of the entropy involved. Suppose that our input $X = (X_1, X_2)$ is a k-source, for some $k \leq n_1$.

Since $\langle E_1, C_1\rangle$ and $\langle E_2, C_2\rangle$ are both 'lossless', the triple (Y_1, Z_1, Z_2) will contain the original k bits of entropy, as well as the extra bits from the input of R_2. The next condition is non-trivial, we need Y_1 to contain sufficient entropy that the conditional entropy left in the buffers (Z_1, Z_2) is small, so that the lossless conductor E_3 is able to conduct all of the entropy from (Z_1, Z_2, R_3) into Y_3.

It will turn out, perhaps rather magically, that we can split into two cases, we may either assume that for every $x_1 \in \mathrm{supp}(X_1)$ the confiditional entropy of $(X_2|X_1 = x_1)$ is small, or for every $x_1 \in \mathrm{supp}(X_1)$ the confiditional entropy of $(X_2|X_1 = x_1)$ is large.

In the first case, $\langle E_2, C_2\rangle$ will conduct the entropy of X_2 into Y_2, and so Z_1 cannot have very large entropy (at most the bits from R_2). Also, if we choose b_2 to be small than Z_2 cannot have very large entropy. However, since as mentioned (Y_1, Z_1, Z_2) has large entropy, it follows that Y_1 has large entropy.

In the second case, $\langle E_2, C_2\rangle$ extracts an almost uniform Y_2 and so $\langle E_1, C_1\rangle$ either conducts some extra entropy into Y_1, or the entropy of X_1 is already so large that this isn't possible. If we choose n_2 small enough that the entropy of X_1 is already reasonably large, then in both cases Y_1 will have high entropy.

Hence, the entropy in the buffers (Z_1, Z_2) will be small enough that we can use E_3 to losslessly conduct all the entropy from (Z_1, Z_2, R_3) into Y_3.

However the above analysis contains lots of delicate interdependencies in terms of the choices of our parameters. To avoid keeping track of all these dependencies, let us 'skip to the end'

and make a single, prescient, choice for our parameters, which we will show are suitable for the above construction. To that end, let $a = 1000 \log \frac{1}{\epsilon}$ and $d = 2a$, and assume that

a) $\langle E_1, C_1 \rangle \colon \{0,1\}^{n-20a} \times \{0,1\}^{14a} \to \{0,1\}^{n-20a} \times \{0,1\}^{14a}$ is an $(n-30a, 6a, \epsilon)$-permutation conductor;

b) $\langle E_2, C_2 \rangle \colon \{0,1\}^{20a} \times \{0,1\}^{a} \to \{0,1\}^{14a} \times \{0,1\}^{21a}$ is a $(20a, 0, \epsilon)$-buffer conductor;

c) $E_3 \colon \{0,1\}^{35a} \times \{0,1\}^{a} \to \{0,1\}^{19a}$ is a $(17a, \epsilon)$-lossless conductor,

where we see that these fall within the achievable ranges from Lemma 10.3.

We aim to show that the resulting conductor, after applying the zig-zag product to the above building blocks, $E \colon \{0,1\}^n \times \{0,1\}^{2a} \to \{0,1\}^{n-a}$ is an $(n - 10a, 4\epsilon)$-lossless conductor. This would be sufficient to prove Theorem 10.1 for arbitrarily large n and for a specific choice of m.

10.5 Proof of Theorem 10.1

So, we want to track the change in entropy from the input (X_1, X_2) to the output $E(X_1, X_2) = (Y_1, Y_3)$. We need to show that if $H_\infty(X_1, X_2) = k \leq n - 10a$ then (Y_1, Y_2) is a $(k + 2a, 4\epsilon)$-source. For ease of discussion we will simply ignore all the small ℓ_1-errors in the outputs of our conductors (essentially assuming for simplicity that $\epsilon = 0$). What would happen if we kept track carefully is that these errors would accumulate additively in the end to give the final error for the conductor E. Under this simplification we wish to show that $H_\infty(Y_1, Y_2) \geq k + 2a$.

To this end, let us note a basic property of min-entropy (and in particular, joint min-entropy) that will be useful for us.

Let X and Y be random variables, then for any $z \in \mathrm{supp}(X)$ it is clear that

$$\begin{aligned} H_\infty(X,Y) &= -\log \max_{x,y} \mathbb{P}(X = x, Y = y) \leq -\log \max_{y} \mathbb{P}(Y = y | X = z)\mathbb{P}(X = z) \\ &= H_\infty(Y|X = z) - \log \mathbb{P}(X = z). \end{aligned} \tag{10.1}$$

As a corollary we get the following

Lemma 10.4. *Suppose that $H_\infty(X,Y) \geq a$ and $H_\infty(Y|X = z) \leq b$ for all $z \in supp(X)$, then $H_\infty(X) \geq a - b$.*

We will also need a small technical lemma, which allows us to partition a joint distribution according to conditional min-entropy.

Lemma 10.5. *Let (X_1, X_2) be a probability distribution on a finite product space. Given $0 < \epsilon \leq \frac{1}{2}$ and a, there exists a probability distribution (Y_1, Y_2) on the same space such that:*

- *The distributions (X_1, X_2) and (Y_1, Y_2) are ϵ-close;*
- *The distribution (Y_1, Y_2) is a convex combination of two other distributions $(\hat{Y}_1, \hat{Y}_2)$ and $(\check{Y}_1, \check{Y}_2)$, each having min-entropy at least $H_\infty(X_1, X_2) - \log\left(\frac{1}{\epsilon}\right)$;*

- *For all $x \in supp(\hat{Y}_1)$ we have $H_\infty(\hat{Y}_2|\hat{Y}_1 = x) \geq a$;*
- *For all $x \in supp(\check{Y}_1)$ we have $H_\infty(\check{Y}_2|\check{Y}_1 = x) < a$;*

Proof. We first split $\mathrm{supp}(X_1)$ according to $H_\infty(X_2|X_1 = x)$ so that

$$\hat{S} = \{x \colon H_\infty(X_2|X_1 = x) \geq a\} \qquad \text{and} \qquad \check{S} = \{x \colon H_\infty(X_2|X_1 = x) < a\}.$$

We then define $(\hat{Y}_1, \hat{Y}_2)$ and $(\check{Y}_1, \check{Y}_2)$ so that $\hat{Y}_1$ and $\check{Y}_1$ have disjoint supports $\hat{S}$ and $\check{S}$ respectively, according to the following:

$$\begin{aligned}\mathbb{P}\big((\hat{Y}_1, \hat{Y}_2) = (x_1, x_2)\big) &= \mathbb{P}\big((X_1, X_2) = (x_1, x_2)|X_1 \in \hat{S}\big);\\ \mathbb{P}\big((\check{Y}_1, \check{Y}_2) = (x_1, x_2)\big) &= \mathbb{P}\big((X_1, X_2) = (x_1, x_2)|X_1 \in \check{S}\big).\end{aligned}$$

Let $p = \mathbb{P}(X_1 \in \hat{S})$. Then we see that $(X_1, X_2) = p(\hat{Y}_1, \hat{Y}_2) + (1-p)(\check{Y}_1, \check{Y}_2)$, where $(\hat{Y}_1, \hat{Y}_2)$ and $(\check{Y}_1, \check{Y}_2)$ have disjoint supports. In particular,

$$H_\infty(\hat{Y}_1, \hat{Y}_2) \geq H_\infty(X_1, X_2) - \log\frac{1}{p} \qquad \text{and} \qquad H_\infty(\check{Y}_1, \check{Y}_2) \geq H_\infty(X_1, X_2) - \log\frac{1}{1-p}.$$

Hence if $p, 1-p \geq \epsilon$ then we can take $(Y_1, Y_2) = (X_1, X_2)$.

On the other hand, if $p < \epsilon$ then, since $\epsilon < \frac{1}{2}$, we have that $1 - p \geq \epsilon$, and hence $(\check{Y}_1, \check{Y}_2)$ is a suitable choice for (Y_1, Y_2), as long as we can show that it is ϵ-close to (X_1, X_2). However this is straighforward since

$$\sum_{x_1 \in \hat{S}, x_2} \Big|\mathbb{P}\big((X_1, X_2) = (x_1, x_2)\big) - \mathbb{P}\big((\check{Y}_1, \check{Y}_2) = (x_1, x_2)\big)\Big| = \mathbb{P}(X_1 \in \hat{S}) = p < \epsilon$$

and

$$\sum_{x_1 \in \check{S}, x_2} \Big|\mathbb{P}\big((\check{Y}_1, \check{Y}_2) = (x_1, x_2)\big) - \mathbb{P}\big((X_1, X_2) = (x_1, x_2)\big)\Big| = \left(\frac{1}{1-p} - 1\right)(1-p) = p < \epsilon.$$

The case $p > 1 - \epsilon$ is similar. □

Firstly we note that, since $\langle E_2, C_2\rangle$ is a $(20a, 0, \epsilon)$-buffer conductor and $\langle E_1, C_1\rangle$ is a permutation conductor, no entropy is lost when we apply $\langle E_2, C_2\rangle$ and $\langle E_1, C_1\rangle$. In particular we have (ignoring the ℓ_1 errors) that

$$k + a = H_\infty(X_1, X_2, R_2) = H_\infty(X_1, Y_2, Z_2) = H_\infty(Y_1, Z_1, Z_2).$$

We note that at this point, all we need to show is that Y_1 has a large enough entropy. Indeed, suppose we can show that

$$H_\infty(Y_1) \geq k - 15a. \tag{10.2}$$

Since E_3 is a $(17a, \epsilon)$-lossless conductor, we have that for any $y_1 \in \mathrm{supp}(Y_1)$

$$H_\infty(Y_3|Y_1 = y_1) \geq \min\{H_\infty(Z_1, Z_2|Y_1 = y_1) + a, 17a\} := h.$$

In other words, for all $y_3 \in \mathrm{supp}(Y_3)$ and $y_1 \in \mathrm{supp}(Y_1)$, $\mathbb{P}(Y_3 = y_3 | Y_1 = y_1) \leq 2^{-h}$ and so $\mathbb{P}(Y_3 = y_3, Y_1 = y_1) \leq 2^{-h}\mathbb{P}(Y_1 = y_1)$.

If $h = 17a$ then, since $H_\infty(Y_1) \geq k - 15a$, $\mathbb{P}(Y_1 = y_1) \leq 2^{-k+15a}$, and it follows that

$$\mathbb{P}(Y_3 = y_3, Y_1 = y_1) \leq 2^{-k-2a}$$

and hence $H_\infty(Y_1, Y_3) \geq k + 2a$.

Conversely, if $h = H_\infty(Z_1, Z_2 | Y_1 = y_1) + a$ then

$$\begin{aligned}
\mathbb{P}(Y_3 = y_3, Y_1 = y_1) &\leq 2^{-h}\mathbb{P}(Y_1 = y_1) \\
&= 2^{-a}2^{-H_\infty(Z_1,Z_2|Y_1=y_1)}\mathbb{P}(Y_1 = y_1) \\
&= 2^{-a}\max_{z_1,z_2}\mathbb{P}(Z_1 = z_1, Z_2 = z_2 | Y_1 = y_1)\mathbb{P}(Y_1 = y_1) \\
&= 2^{-a}\max_{z_1,z_2}\mathbb{P}(Z_1 = z_1, Z_2 = z_2, Y_1 = y_1) \\
&= 2^{-a}2^{-H_\infty(Y_1,Z_1,Z_2)} \\
&= 2^{-a}2^{-k-a},
\end{aligned}$$

and so $H_\infty(Y_1, Y_3) \geq k + 2a$ as before.

So, we wish to prove (10.2). By Lemma 10.5 we can assume we are in one of the following two cases:

Case 1: $H_\infty(X_1, X_2) \geq k - \log\frac{1}{\epsilon} \geq k - a$ and for all $x_1 \in \mathrm{supp}(X_1)$, $H_\infty(X_2 | X_1 = x_1) \geq 14a$.

In this case, since trivially $H_\infty(X_2 | X_1 = x_1) \leq 20a$ for all $x_1 \in \mathrm{supp}(X_1)$, by Lemma 10.4

$$H_\infty(X_1) \geq H_\infty(X_1, X_2) - 20a \geq k - 21a.$$

Furthermore, since $\langle E_2, C_2\rangle$ is a $(20a, 0, \epsilon)$-buffer conductor, conditioned on the value of X_1 the output (Y_2, Z_2) is such that Y_2 is ϵ-close to uniform, and so the joint distribution (X_1, Y_2) is ϵ-close to an independent pair (X_1, U_{14a}). Hence, since E_1 is a $(n - 30a, 6a, \epsilon)$-extracting conductor, $H_\infty(X_1) \geq k - 21a$ and $k - 21a \leq n - 31a$, it follows that

$$H_\infty(Y_1) \geq k - 21a + 6a = k - 15a.$$

Case 2: $H_\infty(X_1, X_2) \geq k - \log\frac{1}{\epsilon} \geq k - a$ and for all $x_1 \in \mathrm{supp}(X_1)$, $H_\infty(X_2 | X_1 = x_1) < 14a$.

First we note that, since $H_\infty(X_1, X_2) \geq k - a$, by (10.1) we have that for all $x_1 \in \mathrm{supp}(X_1)$

$$H_\infty(X_2 | X_1 = x_1) - \log\mathbb{P}(X_1 = x_1) \geq H_\infty(X_1, X_2) \geq k - a.$$

Now, since $\langle E_2, C_2\rangle$ is a $(20a, 0, \epsilon)$-buffer conductor and $H_\infty(X_2 | X_1 = x_1) < 14a$

$$H_\infty(Y_2 | X_1 = x_1) \geq H_\infty(X_2 | X_1 = x_1).$$

It follows that

$$\begin{aligned} H_\infty(X_1, Y_2) &= \min_{x_1} \{H_\infty(Y_2|X_1 = x_1) - \log \mathbb{P}(X_1 = x_1)\} \\ &\geq \min_{x_1} \{H_\infty(X_2|X_1 = x_1) - \log \mathbb{P}(X_1 = x_1)\} \\ &= H_\infty(X_1, X_2) \geq k - a. \end{aligned}$$

Then, since $\langle E_1, C_1 \rangle$ is a permutation, $H_\infty(Y_1, Z_1) \geq k - a$. However, trivially $H_\infty(Z_1|Y_1 = y_1) \leq 14a$ for every y_1 (since Z_1 takes values in $\{0,1\}^{14a}$) and so, by Lemma 10.4

$$H_\infty(Y_1) \geq k - 15a.$$

Hence (10.2) holds, and the proof is completed.

To end this section, we note that there is no known way to algebraically construct lossless expanders, and indeed it seems that the strong vertex expansion that such graphs posses are not implied by simple algebraic properties of graphs. Furthermore, our construction only provides explicit examples of bipartite graphs which expand in a single direction. Relaxing either of these conditions is an interesting open problem.

Question 10.6. *For any $\delta > 0$ and sufficiently large d, give an explicit construction of an arbitrarily large (n, d)-graph G with $\Phi_V(G, \epsilon n) \geq d - 2 - \delta$ where $\epsilon = \epsilon(\delta, d)$.*

11 Metric Embeddings

Finite metric spaces arise in many different context. For example, much of modern science deals with large data sets, which often come with or can be naturally equipped with, some distance metric. Another example comes from graphs, where a natural family of metric spaces arises from the graph metric by assigning some set of edge lengths to a graph, allowing one to treat the graph as a geometric object. In these ways and more, we can treat finite metric spaces from a geometric perspective, but also from a combinatorial and computational perspective.

The topic we'll consider in this chapter is the question of *embedding* a metric space X into another Y. We might not always be able to so in a way that preserves the metric structure of X precisely, but only *approximately* in some sense. However, if the structure of Y is much simpler then it can be useful to view X as being structurally similar to a subspace of Y. Perhaps the simplest examples, geometrically, of metric spaces are the euclidean spaces $\mathbb{R}^d$, and the question we'll be considering is how much do we need to change the structure of a finite metric space to embed it in Euclidean space. It will turn out that, perhaps unexpectedly, expander graphs arise naturally in this context as extremal examples - they require the most distortion over all spaces with the same number of points to be able to be embedded in Euclidean space.

11.1 Embedding metric spaces into Euclidean space

Let us introduce then formally the notions we'll be working with. A *semimetric space* is a pair (X, d), where X is a set of points and d is a *distance function* $d\colon X \times X \to \mathbb{R}^+$ which is symmetric, non-negative and satisfies the triangle inequality

$$d(x, y) \leq d(x, z) + d(y, z) \qquad \text{for all } x, y, z \in X.$$

(X, d) is then a *metric space* if $d(x, y) = 0$ if and only if $x = y$.

Suppose we have an embedding $f\colon X \to \mathbb{R}^n$ of the space (X, d) into the metric space $(\mathbb{R}^n, ||\cdot||_2)$, n-dimensional Euclidean space with the ℓ_2 distance. We define

$$\text{expansion}(f) = \max_{x_1, x_2 \in X} \frac{||f(x_1) - f(x_2)||_2}{d(x_1, x_2)},$$

$$\text{contraction}(f) = \max_{x_1, x_2 \in X} \frac{d(x_1, x_2)}{||f(x_1) - f(x_2)||_2},$$

$$\text{distortion}(f) = \text{expansion}(f) \cdot \text{contraction}(f).$$

Note that these definitions ensure the the distortion of f is unchanged when scaling f by a linear factor.

It is not hard to see that there are some metric spaces which cannot be embedded into Euclidean space without distortion, regardless of the dimension of the target space. Indeed, consider the metric of the star graph $K_{1,3}$ with center 4 and leaves $1, 2, 3$. We see that $d(4, 1) = d(4, 2) = d(4, 3) = 1$ and $d(x, y) = 2$ otherwise. In particular, by the fact that the triangle equality is strict in $\mathbb{R}^n$ except for colinear triples, it follows that each triple containing 4 lies on a single line, and so all four points must lie on a single line. However, this clearly leads to a contradiction.

11.2 Minimising the ℓ_2 distortion

Given a finite metric space (X, d) let us write $c_2(X, d)$ for the least possible distortion in any embedding of (X, d) into $(\mathbb{R}^n, ||\cdot||_2)$. Note that the minimum distortion can clearly be achieved in $\mathbb{R}^{|X|}$.

A well-known result of Bourgain shows that arbitrary metric spaces can be embedded into Euclidean space with only logarithmic distortion.

Theorem 11.1. *[Bourgain] Any n-point metric space (X, D) can be embedded into Euclidean space with distortion $O(\log n)$.*

Another interesting and powerful result shows that subspaces of Euclidean spaces can be embedded into a space of much smaller dimension without signifcant loss in distortion. We say a metric space (X, d) is an *ℓ_2-metric* if there is a distance preserving embedding $f\colon X \to \mathbb{R}^n$.

Theorem 11.2 (Johnson-Lindenstrauss)**.** *For any $\epsilon > 0$ and any n-point ℓ_2-metric (X, d) there exists an embedding $f\colon X \to \mathbb{R}^m$ of distortion $\leq 1 + \epsilon$ where $m = O\left(\frac{\log n}{\epsilon^2}\right)$.*

In fact, the proof of the above theorem is remarkably simple, one takes a random linear projection to a low dimensional subspace and the result follows from standard concentration results. By combining the two theoroms we see that for arbitrary spaces, logarithmic distortion can be achieved even in dimension $O\left((\log n)^2\right)$.

It turns out that there is in fact an efficient (polynomial time) algorithm which computes $c_2(X, d)$, using semi-definite duality, which gives us a simple way to prove lower bounds on $c_2(X, d)$. Let us show that we can reduce the problem of computing $c_2(X, d)$ to a semi-definite optimisation problem.

Theorem 11.3 (Linial-London-Rabinovich)**.** *There is a polynomial time algorithm that, given a metrix space (X, d), computes $c_2(X, d)$.*

Proof. Let $X = \{x_1, \ldots, x_n\}$ and suppose we have some embedding $f\colon X \to \mathbb{R}^n$. Since we can always scale f so that contraction$(f) = 1$, we may assume that distortion$(f) \leq \gamma$ if and only if

$$d(x_i, x_j)^2 \leq ||f(x_i) - f(x_j)||_2^2 \leq \gamma^2 d(x_i, x_j)^2 \qquad \text{for all} 1 \leq i < j \leq n. \tag{11.1}$$

Recall that a symmetric $n \times n$ matrix Z is said to be positive semi-definite if $\boldsymbol{v} Z \boldsymbol{v}^T \geq 0$ for all $\boldsymbol{v} \in \mathbb{R}^n$. It can be seen that this is equivalent to the following two conditions:

1. All eigenvalues of Z are non-negative;
2. $Z = WW^T$ for some matrix W.

Let use write $PSD = PSD_n$ for the collection of all $n \times n$ positive semi-definite matrices.

Let us consider the matrix U whose ith row is given by $f(x_i) = u_i$ and let $Z = UU^T \in PSD$. In this way (11.1) is equivalent to

$$d(x_i, x_j)^2 \leq z_{ii} + z_{jj} - 2z_{ij} \leq \gamma^2 d(x_i, x_j)^2 \qquad \text{for all} 1 \leq i < j \leq n. \tag{11.2}$$

since $||u_i - u_j||^2 = z_{ii} + z_{jj} - 2z_{ij}$. Hence, we see that $c_2(X,d) \leq \gamma$ if and only if there is a positive semi-definite matrix Z satisfying (11.2).

However, this is an optimisation problem which can be solved in polynomial time using the so-called ellipsoid algorithm. □

So, the algorithm above constructs an equivalent optimisation problem, and solves it using the ellipsoid algorithm. However, if we look at the dual problem we can extract a handy method for proving lower bounds on the distortion. However, to do so, we need to know how to dualise the constraint that $Z \in PSD$. We can do so using a simple but useful fact from linear algebra.

Claim 11.4. A matrix Z is positive semi-definite if and only if $\sum_{i,j} q_{i,j} z_{i,j} \geq 0$ for all positive semi-definite matrices Q.

Proof. To prove the sufficiency of the condition, given an arbitrary $v \in \mathbb{R}^n$ let us consider the matrix $Q \in PSD$ given by $q_{i,j} = v_i v_j$. Then

$$\boldsymbol{v} Z \boldsymbol{v}^T = \sum_{i,j} q_{i,j} z_{i,j} \geq 0,$$

and so $Z \in PSD$.

For the converse, we first note that any positive semi-definite matrix of rank 1 is of the form $q_{i,j} = v_i v_j$ for some $\boldsymbol{v} \in \mathbb{R}^n$, and so $\sum_{i,j} q_{i,j} z_{i,j} \geq 0$ since $Z \in PSD$. However, any positive semi-definite matrix can be written as UU^T for some matrix U with orthogonal rows. Hence, every $Q \in PSD$ can be written as the sum of rank 1 positive semi-definite matrices, implying the claim. □

Theorem 11.5. *[Linial-London-Rabinovich] For any finite metrix space (X,d)*

$$c_2(X,d) = \max_{P \in PSD, P\boldsymbol{u}^T = 1} \sqrt{\frac{\sum_{p_{i,j}>0} p_{i,j} d(x_i,x_j)^2}{-\sum_{p_{i,j}<0} p_{i,j} d(x_i,x_j)^2}}.$$

Proof. As we saw, $c_2(X,d)$ is the solution to the primal problem

$$\begin{aligned} \sum_{i,j} q_{i,j} z_{i,j} \geq 0 \qquad & \text{for all } Q \in PSD, \\ z_{ii} + z_{jj} - 2z_{ij} \geq d(x_i,x_j)^2 \qquad & \text{for all } ij, \\ \gamma d(x_i,x_j)^2 \geq z_{ii} + z_{jj} - 2z_{ij} \qquad & \text{for all } i,j. \end{aligned}$$

Hence, duality implies that for $\gamma < c_2(X,d)$, there must exist a non-negative combination of the constraints of the primal problem that yields a contradiction.

So, we are looking for a linear combination of the constraints that yields a contradiction. So we have a contradiction of the form

$$\sum_k a_k q^k_{i,j} z_{i,j} + \sum_{i,j} b_{i,j}(z_{ii} + z_{jj} - 2z_{ij}) + c_{i,j}(z_{ii} + z_{jj} - 2z_{ij}) \geq \sum_{i,j} b_{i,j} d(x_i,x_j)^2 + \gamma^2 c_{i,j} d(x_i,x_j)^2,$$

where we've chosen some set of matrices $Q^1, Q^2 \ldots \in PSD$ and we have $a_k, b_{i,j}, -c_{i,j} \geq 0$.

We first note that, since PSD forms a positive cone, we may assume that there is a single matrix $Q = Q^1$ and that $a_1 = 1$.

Then, in order to cancel the off-diagonal variables $z_{i,j}$ on the left hand side, we see we need to take

- $b_{i,j} = \frac{q_{i,j}}{2}$ and $c_{i,j} = 0$ if $q_{i,j} > 0$; and
- $b_{i,j} = 0$ and $c_{i,j} = -\frac{q_{i,j}}{2}$ if $q_{i,j} < 0$.

In order for the diagonal entries $z_{i,i}$ to be cancelled by this process we see that we need

$$0 = z_{i,i}\left(q_{i,i} + \sum_{q_{i,j}>0} \frac{q_{i,j}+q_{j,i}}{2} + \sum_{q_{i,j}<0} \frac{q_{i,j}+q_{j,i}}{2}\right) = z_{i,i}\sum_{i,j} q_{i,j}.$$

Hence, this can only be a contradiction if the row sums of Q are 0. Assuming this to be the case we see that the inequality we end up with is

$$0 \geq \sum_{q_{i,j}>0} q_{i,j} d(x_i, x_j)^2 + \gamma^2 \sum_{q_{i,j}<0} q_{i,j} d(x_i, x_j)^2,$$

which is a contradiction if and only if

$$\gamma \leq \sqrt{\frac{\sum_{q_{i,j}>0} q_{i,j} d(x_i, x_j)^2}{-\sum_{q_{i,j}<0} q_{i,j} d(x_i, x_j)^2}}.$$

□

As we will see, this theorem is useful as it allows us to get a lower bound on $c_2(X, d)$ by simply choosing an appropriate positive semi-definite matrix P and evaluating the function in the theorem.

11.3 Distortion bounds via semi-definite duality

Let us demonstrate the power of Theorem 11.5 by applying it to some graph metrics. Given a graph $G = (V, E)$ there is a natural metric space (V, d_G) given by the graph distance $d_G(x, y)$. Let us write $c_2(G) = c_2(V, d_G)$.

11.3.1 Embedding the hypercube into Euclidean space

Recall that Q^r is the r-dimensional hypercube and note that the graph metric $r := d_{Q^r}$ coincides with the Hamming metric. There is a natural embedding of Q^r into $\mathbb{R}^r$ given by identifying the vertices of Q^r with $\{0, 1\}^r$, and it is easy to check that the contraction of this map is $\sqrt{r}$ and the expansion is one. Hence, $c_2(Q^r) \leq \sqrt{r}$.

In order to show that this is best possible, we want to choose a clever positive semi-definite matrix P. Let us define

$$p_{x,y} = \begin{cases} -1 & \text{if } d(x,y) = 1 \\ r-1 & \text{if} x = y \\ 1 & \text{if} d(x,y) = r \\ 0 & \text{if} else \end{cases}$$

Then we see that $P\boldsymbol{u}^T = 0$ and it is easy to check that $P \in PSD$ (for example by determining its spectrum, which is the same of the hypercube).

However, we can check that

$$\sum_{p_{x,y}>0} q_{x,y} d(x,y)^2 = r^2 2^r \qquad \text{and} \qquad -\sum_{p_{x,y}<0} q_{x,y} d(x,y)^2 = r 2^r$$

and so, by Theorem 11.5, it follows that $c_2(Q^r) \geq \sqrt{r}$ and well. Hence $c_2(Q^r) = \sqrt{r}$.

11.3.2 Embedding expander graphs into Euclidean space

We note first a simple fact - for any graph G we have $c_2(G) = \text{diam}(G)$. Indeed, if we simply map the vertices of G arbitrarily to the vertices of a simplex in $\mathbb{R}^n$ then, since the pairwise distance of all vertices in the simplex is one, the expansion is 1 and the contraction is equal to the diameter of G.

In particular, since constant degree expander graphs have logarithmic diameter, we see that they can be embedded with logarithmic distortion. We will see that this is optimal up to a multiplicative constant.

To do so we will need the following simple lemma.

Lemma 11.6. *Let G be an (n,k)-graph with an even number of vertices and let H be the graph on the same vertex set, where two vertices are adjacent if their distance in G is at least $\log_k n$. Then H has a perfect matching.*

Proof. Since G is k-regular, it follows that there are at most $\frac{n}{2}$ vertices at distance at most $\log_k n - 1$ from any fixed vertex, and so the minimum degree of H is at least $\frac{n}{2}$. Hence, by Dirac's theorem H contains a Hamilton cycle, and so a perfect matching. □

Theorem 11.7 (Linial-London-Rabinovich)**.** *Let $k \geq 3$ and let $\epsilon > 0$. If G is an (n,k)-graph with $\lambda_2(G) \leq k - \epsilon$, the $c_2(G) = \Omega(\log n)$, where the implicit constant only depends on k and ϵ.*

Proof. Let H be the graph defined in the statement of Lemma 11.6 and let B be the adjacency matrix of the perfect matching M in H guaranateed by the lemma.

Let us consider the matrix $P = kI - A(G) + \epsilon(B - I)$. Since G is k-regular and B is 1-regular, it follows that $P\boldsymbol{u}^T = 0$. We wish to show that $P \in PSD$.

Given an arbitrary $\boldsymbol{x} \in \mathbb{R}^n$ we wish to show that $\boldsymbol{x}P\boldsymbol{x}^T \geq 0$. Note that, since $\boldsymbol{u}$ is an eigenvector of P with eigenvalue 0, we may assume that $\boldsymbol{x} \perp \boldsymbol{u}$. then we see that

$$\boldsymbol{x}(kI - A(G))\boldsymbol{x}^T \geq (k - \lambda_2)||\boldsymbol{x}||_2^2 \geq \epsilon||x||_2^2,$$

and

$$\boldsymbol{x}(B - I)\boldsymbol{x}^T = \sum_{(i,j)\in M} 2x_i x_j - \sum_i x_i^2 = -\sum_{(i,j)\in M} (x_i + x_j)^2 = -||\boldsymbol{x}||_2^2.$$

It follows that

$$\boldsymbol{x}P\boldsymbol{x}^T \geq \epsilon||x||_2^2 - \epsilon||x||_2^2 = 0,$$

as desired.

Furthermore, since P is negative off the diagonal for all edges of G, and positive off the diagonal for all edges of B (note an edge cannot be in B and G, since adjacent vertices in G are too close to be adjacent in B), we can calculate

$$-\sum_{p_{i,j}<0} d(i,j)^2 p_{i,j} \leq e(G) = kn;$$

and

$$\sum_{p_{i,j}>0} d(i,j)^2 p_{i,j} \geq \epsilon n(\log_k n)^2.$$

Hence, it follows from Theorem 11.5 that

$$c_2(G) \geq \sqrt{\frac{\epsilon n(\log_k n)^2}{kn}} = O(\log n),$$

as claimed. □

More generally we can view these two examples from the point of view of *Poincaré type inequalities* on graphs. Given a graph function $f\colon V(G) \to \mathbb{R}^n$, a Poincaré type inequality compares the average of terms $||f(u) - f(v)||_2$ considering over *all pairs of vertices* with just those considered over the edges of G.

Theorem 11.8. *Let $G = (V, E)$ be a k-regular graph with second eigenvalue λ_2. For every embedding $f\colon V \to \mathbb{R}^n$*

$$\mathbb{E}_{(u,v)\in V^2}||f(u) - f(v)||_2^2 \leq \frac{k}{k - \lambda_2}\mathbb{E}_{(u,v)\in E}||f(u) - f(v)||_2^2.$$

Sketch. We first note that it suffices to prove the theorem for $f\colon V \to \mathbb{R}$, since if the theorem holds co-ordinate wise for f then it clearly holds for f. We also note that, since both sides of the inequality of invariant under translation by a constant, we may assume that f has average 0. However then the inequality is just the variational definition (in terms of the Rayleigh quotient) of the second eigenvalue. □

The previous results for embedding the hypercube and expander graphs can be derived (at least up to a constant factor) from this inequality.

11.4 Algorithms for cut problems via embeddings

Let us consider the natural computational problem of determining the expansion ratio $h(G)$ of an n vertex graph G. This is just one of a family of related problems to do with finding edges cuts, for example finding balanced cuts, which often arise naturally as important parts of 'divide-and-conquer' type algorithms, where small cuts guarantee smaller interference between the solutions in different parts.

Whilst it has been known for a long time that determining $h(G)$ is co-NP hard, it is an interesting and open questions how well it can be approximated in polynomial time. A major breakthrough came from the work of Leighten and Rao who showed that there is a polynomial time algorithm which approximates $h(G)$ to within a factor of $O(\log n)$.

Here we will present a different proof of this fact, due to Linial, London and Rabinovich, which solves this problem by establishing a connection between cut problems in graphs and low-distortion embeddings in Euclidean space. We note that it has since been shown that a factor of $O(\sqrt{\log n})$ is achievable, and it is an open question whether in fact a constant factor approximation is.

We first note

$$\frac{2h(G)}{n} = \frac{2}{n} \min_{|S|\leq \frac{n}{2}} \frac{e(S,S^c)}{|S|} \geq \min \frac{e(S,S^c)}{|S||S^c|} \geq \frac{1}{n} \min_{|S|\leq \frac{n}{2}} \frac{e(S,S^c)}{|S|} = \frac{h(G)}{n},$$

and so, if we only care about approximating $h(G)$, it is equivalent to look at the problem of approximating $\min \frac{e(S,S^c)}{|S||S^c|} := \Psi(G)$.

It turns out that this questions of minimising $\frac{e(S,S^c)}{|S||S^c|}$ can be quite nicely restated in terms of some very natural semimetrics, the *cut metrics*. Given a subset $S \subseteq V(G)$ we can define a semimetric on V be letting $d_S(x,y) = 0$ if $x,y \in S$ or $x,y \in S^c$ and $d_S(x,y) = 1$ otherwise. In this case we see that

$$\frac{e(S,S^c)}{|S||S^c|} = \frac{\sum_{(i,j)\in E} d_S(i,j)}{\sum_{i,j} d(i,j)}.$$

Hence, $\Psi(G)$ is equivalent to minimising a certain quantity over cut metrics. A natural relaxation would then be to minimise this quantity over *all semimetrics* so let us define

$$LR(G) = \min_{d \text{ a semimetric}} \frac{\sum_{(i,j)\in E} d(i,j)}{\sum_{i,j} d(i,j)}.$$

It is then clear that $LR(G) \leq \min \frac{e(S,S^c)}{|S||S^c|} \leq \frac{2h(G)}{n}$, but we will also be able to show that $\frac{h(G)}{n} \leq \min \frac{e(S,S^c)}{|S||S^c|} \leq O(\log n) LR(G)$. Furthermore, since $LR(G)$ can be expressed as the solution of the following linear programming problem

$$\begin{aligned}
\text{minimise} \quad & \sum_{(i,j)\in E} d(i,j) \\
\text{subject to} \quad & \sum_{i,j} d(i,j) = 1 \\
d(i,k) \leq d(i,j) + d(j,k) & \text{ for all } i,j,k \\
d(i,j) \geq 0 & \text{ for all } i \neq j \\
d(i,i) = 0 & \text{ for all } i,
\end{aligned}$$

we can find a solution to $LR(G)$ in polynomial time.

However, there is another natural intermediary relaxation of $\Psi(G)$ given by

$$LR_1(G) = \min_{d \text{ an } \ell_1\text{-semimetric}} \frac{\sum_{(i,j)\in E} d(i,j)}{\sum_{i,j} d(i,j)},$$

where we an ℓ_1-semimetric is any semimetric which arises from mapping V into $\mathbb{R}^n$ under the ℓ_1 norm. We note that it is easy to show that every ℓ_2-metric is also an ℓ_1-(semi)metric. In fact, it turns out that this is not even a relaxation at all!

Indeed, if we view the set of semimetrics d on V as living in $\mathbb{R}^{\binom{n}{2}}$, then it is clear that the set of semimetrics forms a complex cone (they are closed under positive scalar multipliation and convex combinations). It is perhaps less clear that the ℓ_1-semimetrics also form a convex cone, but this turns out to be tree. In fact, it turns out that they coincide with the convex cone CUT generated by the cut metrics.

Lemma 11.9. *CUT coincides with the set of ℓ_1-semimetrics.*

Proof. Let $d \in \text{CUT}$ so that $d = \sum_{S\subseteq[n]} \alpha_S d_S$ with $\alpha_S \geq 0$ for all S. Let us take an embedding f of $[n]$ into $\mathbb{R}^{2^n}$ as follows

$$f(i)_S = \alpha_S \text{ if } i \in S \text{ and } 0 \text{ otherwise.}$$

It is easy to see that

$$d(i,j) = \sum_{S\colon i\in S, j\notin S \text{ or } i\notin S, j\in S} \alpha_S = ||f(i) - f(j)||_1.$$

Conversely, suppose that d is an ℓ_1-semimetric and let $f\colon [n] \to \mathbb{R}^n$ witness this. Since d is just the sum of the metrics given by the co-ordinates of f, and CUT is a cone, it suffices to show that each co-ordinate lies in the cut cone. However embeddings $f_i\colon [n] \to \mathbb{R}$ are particularly simple, and it is easy to see that they can be generated by cut metrics. □

Furthermore, since the cut cone is convex, the optimum in $LR_1(G)$ is obtained on an extremal ray, which are then clearly given by the cut metrics, and so it follows that $\Psi(G) = LR_1(G)$. However, since every ℓ_2-metric is an ℓ_1-semimetric, it follows from Theorem 11.1 (which also holds for semimetrics) that we can approximate the semimetric minimising $LR(G)$ by an ℓ_1-semimetric with distortion $O(\log n)$. It follows that $O(\log n)LR(G) \geq LR_1(G) = \Psi(G)$ and hence we can approximate $h(G)$ up to an $O(\log n)$ factor in polynomial time.

We note that the question of determining $\Psi(G)$ is very closely related to another well-known computational problem, that of determining the maximal *all-pairs multicommodity flow* in a graph, whose solution is in fact given by $LR(G)$.

This type of argument seems quite alluring from a computing perspective - any problem to do with minimising certain types of cuts $e(S, S^c)$ in a graph G be be transformed into a convex optimisation problem over the cut cone. There is a general theory of optimisation over some convex domain Ω which would then be very useful for solving such problems, but in order to apply it one has to be able to solve efficiently two basic problems for Ω:

- *Membership* - Given a point x to determine if $x \in \Omega$;
- *Separation* - Given a point $x \notin \Omega$ to find a hyperplane separating x from Ω.

Unfortunately, the cut cone is computationally quite bad. Even the membership problem, which is the simpler of the two, is $\boldsymbol{NP}$-hard. To put it another way, it is difficult to determine if a metric is an ℓ_1-metric.

However, this does suggest the following problem.

Question 11.10. *Is there a different cone which is a good approximation (ideally constant distortion) to the cut cone, but for which the membership and separation problems can be solved in polynomial time (in the dimension)?*

In fact, Linial and Goemans, who raised this question, even suggested a suitable candidate: We say that a metric space (X, d) is of *negative type* if $\sqrt{d}$ is an ℓ_2-metric. It can be shown that such metrics form a convex cone for which both the membership and separation problem can be solved efficiently. Also, every ℓ_1-metric is of negative type.

Question 11.11. *Can every metric of negative type be embedded into an ℓ_1-metric with bounded distortion?*