An introduction to expander graphs

E. Kowalski

ETH Zürich
Version of January 8, 2024
kowalski@math.ethz.ch

… à l'expansion de mon cœur refoulé s'ouvrirent aussitôt des espaces infinis.

M. Proust, **À l'ombre des jeunes filles en fleurs**
(deuxième partie, **Noms de Pays : le Pays**)

He walked as if on air, and the whole soul had obviously expanded,
like a bath sponge placed in water.

P.G. Wodehouse, **Joy in the Morning**
(chapter 16)

Contents

Preface

The goal of this book is to give an introduction to *expander graphs* and their applications. It is therefore related to the books of Lubotzky [78] (and his Colloquium Lectures [79]), of Sarnak [101], and of Tao [109], and to the detailed survey of Hoory, Linial and Wigderson [54]. Each of these is a wonderful source of information, but we hope that some readers will also find interest in special features of this new text. I hope in particular that the discussion of the basic formalism of graphs and of expansion in graphs, which is more detailed than is usual, will be helpful for many mathematicians. Of course, others might find it just pedantic; for these readers, maybe the variety of constructions of expander graphs, and of applications (some of which have not been discussed in previous books) will be a redeeming feature.

The first version of these notes were prepared in parallel with a course that I taught at ETH Zürich during the Fall Semester 2011. I thank all the people who attended the course for their remarks and interest, and corrections, in particular E. Baur, P-O. Dehaye, O. Dinai, T. Holenstein, B. Löffel, L. Soldo and P. Ziegler. Also, many thanks to R. Pink for discussing various points concerning non-concentration inequalities and especially for helping in the proof of the special case needed to finish the proof of Helfgott's Theorem in Chapter 6.

The text was continued (including both corrections and changes and the addition of some material) for a short course at TU Graz during the Spring Semester 2013. Many thanks to R. Tichy and C. Elsholtz for the invitation to give this course.

The final version arose also from teaching various minicourses in Neuchâtel ("Expanders everywhere!"), Lyon ("Colloque Jeunes Chercheurs en Théorie des Nombres"), and during the Ventotene 2015 conference "Manifolds and groups". I thank the respective organizers (A. Valette and A. Khukro for the first; L. Berger, M. Carrizosa, W. Nizioł, E. Royer and S. Rozensztajn for the second; S. Francaviglia, R. Frigerio, A. Iozzi, K. Juschenko, G. Mondello and M. Sageev for the last) for inviting me, and especially A. Iozzi for the last one, which was especially enjoyable in view of its setting. The last step was teaching a course again in the Spring 2016 at ETH Zürich; thanks to B. Löffel and to J. Volec for their help and corrections at that time.

I also thank M. Burger, J. Ellenberg, C. Hall and A. Valette for many discussions related to expanders (and their applications) over the years, and finally I thank N. Bourbaki for his kind invitation to talk in his seminar about expanders and sieve.

Finally, this work was partially supported supported by the DFG-SNF lead agency program grant 200021L_153647.

Zürich, October 2017.

For the reimpression in 2021, some minor corrections have been made. Thanks to C. Ballantine, A. Isakovic, E. Fuchs, A. Tran and M. Litman for sending some of them.

Zürich, March 2021.

CHAPTER 1

Introduction and motivation

This short chapter is highly informal, and the reader should not worry if parts are not immediately understood, or are somewhat ambiguous: we will come back with fully rigorous definitions of all terms later.

Our goal is to introduce *expander graphs*. The outline is roughly the following: (1) we will explain the definition, or precisely give three definitions and show that they are equivalent; (2) we will then give different proofs of the existence of expanders (which is by no means obvious from the definition!), first the original one (based on probabilistic methods) and then three others (with more or less details); (3) we will then present some of the remarkably varied and surprising applications of expanders, with a focus in "pure" mathematics (this part is to some extent a survey, since explaining from scratch the context in all cases would require too much space).

We begin with a brief informal outline of the definition of expanders, and some of their applications. Hopefully, the reader will be convinced that it is a story worth knowing something about, and turn to the rest of the book...

To start with, graphs seem very intuitive mathematical objects. For the moment, we consider them in this manner, while in the next chapter we will give a formal definition. So we view a graph as a set V of vertices, and a set E of edges joining certain pairs (x, y) of vertices, and we allow the possibility of having multiple edges between x and y, as well as loops joining a vertex x to itself. We visualize graphs geometrically, and think of the edges as ways to go from one vertex to another. For our purpose, these edges are considered to be unoriented. One can then speak of "which vertices can be linked to a given vertex x", or of the distance between two vertices x and y as the length of the shortest sequence of edges starting from x and ending at y.

Graphs enter naturally in many concrete problems as models for real-life objects, possibly using different conventions (e.g., oriented edges). Here are a few examples:

- **[Transport network]** In a given geographical area (a town, a country, or even the earth) one often visualizes the transport possibilities within this area (possibly restricted to certain means of transportation, such as trains, tramways, subways, planes, roads) as a graph. For instance, Figure 1.1 represents the well-known tramway network of Zürich in 2012. This graph has no loop but it has many multiple edges since a number of lines travel in parallel in certain areas.
- **[The brain]** Viewing neurons as vertices and synapses as edges, the brain of an animal is – in a rather rough first approximation – also a graph. To the author's knowledge (gathered from the internet...), the only species for which this graph has been determined in its entirety is the *nematode Caenorhabditis Elegans* (a worm of size approximately 1 millimeter, see [120]); this contains 302 neurons and about 8000 synapses. Determining this graph was done by White, Southgate, Thomson, Brenner [118] in 1986 (and corrected in 2011 by Varshney, Chen, Paniagua, Hall, Chklovskii [117], from which paper the figure is taken).

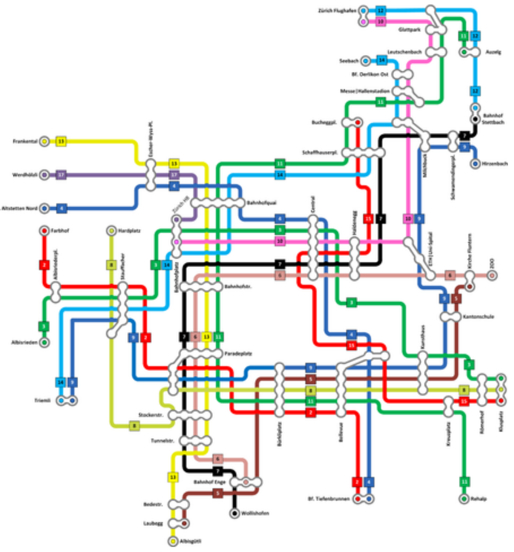


FIGURE 1.1. Zürich tramway network (Wikipedia, Author: mateusch, license Creative Commons Attribution-Share Alike 3.0 Unported.)

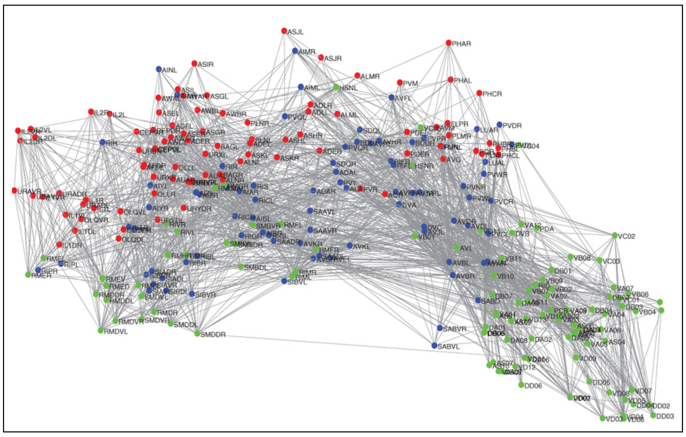


FIGURE 1.2. The nervous system of *Caenorhabditis Elegans*

- **[Relationship graphs]** Given a set of individuals and a relation between them (such as "X is a relative of Y", or "X knows Y", or "X has written a joint paper with Y"), one can draw the corresponding graph. Its *connectedness* properties are often of interest: this leads, for instance, to the well-known Erdős number of a mathematician, which is the distance to the vertex "Paul Erdős" on the collaboration graph joining mathematicians that have a joint paper. Genealogical trees form another example of this type, although the relation "X is a child of Y" is most naturally considered as an oriented edge.

Expander graphs, the subject of these notes, are certain families of graphs, becoming larger and larger, which have the following two competing properties: (1) they are fairly sparse (in terms of number of edges, relative to the number of vertices); (2) yet they are highly connected, and in fact highly "robust", in some sense.

There are different ways of formalizing these ideas. We assume given a family of finite graphs Γ_n with vertex sets V_n such that the size $|V_n|$ goes to infinity, and we first formalize the condition of sparsity by asking that the *degree* of Γ_n be bounded by some constant $v \geqslant 1$ for all n, i.e., for any n, any vertex $x \in V_n$ has at most v distinct neighbors in V_n.

If we think of graphs as objects that might be realized physically (as a communication network with physical links between vertices), with a certain cost associated with each physical edge, this assumption means that increasing the number of vertices (by taking a larger graph Γ_n from our family) will increase the cost linearly with respect to the increase in the number of vertices, since the number of edges of Γ_n is at most $v|V_n|$. Clearly, this sparsity is important if one constructs a tramway network...

The second condition satisfied by expander graphs generalizes the property of *connectedness*, which would simply mean that one can go from any vertex to any other in the graph, by at least some path. One natural strengthening is to ask that such a path is always rather short, which means that the maximal distance in the graph between two points is much smaller than the number of vertices. However, this is not sufficient to define an expander, because a small diameter does not prevent the existence of a "bottleneck" in a graph: even though the graph is connected, there might well exist a rather small subset B of edges such that the graph obtained by removing B (and all edges with at least one extremity in B) is disconnected. To avoid this, one wishes that any subset $V \subset V_n$ of vertices should have many connections with its complement $W = V_n - V$, i.e., there should be many edges linking vertices v and w with $v \in V$ and $w \in W$. Even more precisely, *expanders* are determined by the condition that, for some constant $c > 0$, independent of n, the number of such edges should be at least $c \min(|V|, |W|)$ for all (non-empty) subsets $V \subset V_n$, and for all n.

This definition of sparse, highly connected, robust, families of graphs is obviously quite strong. It is by no means obvious that they exist at all! In fact, as we will see, most elementary explicit families of graphs that one might write down do not satisfy the required condition. Nevertheless, expander families do exist, and in fact exist in great abundance (this was first shown using probabilistic methods.) Moreover, it is maybe even more surprising that they turn out to appear in many different areas of mathematics, and lead to extremely remarkable results in unexpected directions. For these applications, the existence question for expanders often re-appears in a different way, because one typically has little or no possibility to choose the graphs involved, and one must prove that the ones which do appear are, indeed, expanders. This explains why we present different approaches in Chapter 4:

- The original probabilistic approach;
- The recent construction of Ramanujan graphs by Marcus, Spielman and Srivastava (these are, in a certain precise sense, "best possible" expanders);
- The proof using Kazhdan's Property (T) that Cayley graphs of finite quotients of $\mathrm{SL}_3(\mathbf{Z})$ form a family of expanders;
- And finally the proof that (certain families of) Cayley graphs of $\mathrm{SL}_2(\mathbf{Z}/p\mathbf{Z})$ are expanders, when p runs over primes, a beautiful result that is the combination of the work of Helfgott and Bourgain-Gamburd.

Most of these constructions have already appeared in textbooks: for Property (T), see for instance the book of Bekka, de la Harpe and Valette [7, §6.1], or the accounts of Lubotzky [78, §3.3, §4.4] and Sarnak [101, §3.3]; an elementary proof of expansion of certain (very special) Cayley graphs of $\mathrm{SL}_2(\mathbf{Z}/p\mathbf{Z})$ is given by Davidoff, Sarnak and Valette [31]; and finally Tao [109] provides a general proof of a vast generalization of the last approach. In fact, the last result is clearly the most delicate, and our account is mostly contained in a separate chapter. We hope that this will provide a good introduction to the ideas surrounding this area, which has been spectacularly successful (and important) in recent years.

Here are some easily-described applications of expanders that should already give of hint of their surprising ubiquity:

- **[Barzdin–Kolmogorov; Embedding graphs in space]** One of the first mention of expanders, together with a proof of existence, is found in a paper of Pinsker [95] from 1973, and indeed until a few years ago, most references quoted this as the first appearance of expander graphs (see, for instance, the survey [54]). However, as pointed out by L. Guth, an earlier work of Barzdin and Kolmogorov [5] contains a definition of a class of (directed) graphs which is extremely close to that of expanders, and a similar probabilistic proof of their existence. The motivation of this work is a very nice result which is well worth describing in some detail: roughly speaking, the starting point is the fact that any finite graph can be realized in $\mathbf{R}^3$ (with points as vertices and smooth curves as edges), and the question raised by Barzdin and Kolmogorov is: "How small a volume does one need to realize a graph Γ in $\mathbf{R}^3$ as above, if we view the vertices and edges as having a fixed thickness?" By this, they mean that the vertices must be placed at points located at least at distance 1 from each other, and any non-adjacent edges must also be separated at least by such a distance. Barzdin and Kolmogorov first show, constructively,[1] that one can always do this for a 3-regular graph in a volume about $n^{3/2}$ (in fact, in a sphere of radius approximately $\sqrt{n}$), and they next show that this result is best possible; for this last goal, they define expander graphs (or a variant thereof) and show that a 4-regular expander cannot be realized in a volume less than $n^{3/2}$; finally, by proving that "random" graphs are expanders, they conclude that their upper bound is indeed best possible. We will prove a variant of this fact in Section 5.1; see [5] for the original paper, and the recent work of Gromov and Guth [48, §1.1] for a modern re-interpretation and many generalizations.
- **[The brain as expander]** There is some speculation that the brain, as graph, has good expansion qualities. Indeed, Barzdin and Kolmogorov mention in their paper that part of their motivation was related to thoughts about the graph structure of the neurons in the human brain, and recently L. Valiant [115] has proposed algorithmic models of computation for the brain based on graphs, and requiring good connectedness properties. He states:

  > In [5,14] it is shown that algorithms for the four random access tasks described above can be performed on the neuroidal model with realistic values of the numerical parameters. (...) In order that they be able to do this certain graph theoretic connectivity properties are required of the network. The property of expansion [15], that any set of a certain number of neurons have between them substantially more neighbors than their own number, is an archetypal such property. (This property, widely studied in computer science, was apparently first discussed in a neuroscience setting [16].) The vicinal algorithms for the four tasks considered here need some such connectivity properties. In each case random graphs with appropriate realistic parameters have it, but pure randomness is not necessarily essential.

[1] Indeed, this explicit construction seems to have been the best-known part of the paper, in some parts of the computer-science community, where constructing physical networks in as little space as possible is rather important...

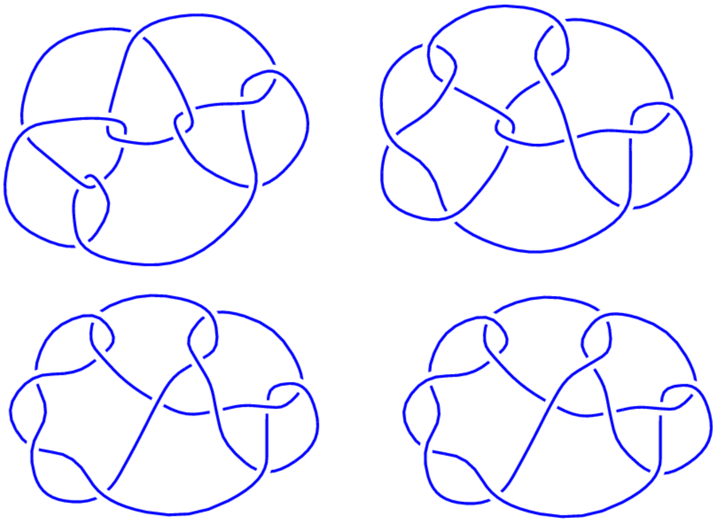


FIGURE 1.3. Some knots

The reference [15] is the survey [54] and [16] refers to the paper of Barzdin–Kolmogorov.

- **[Gromov–Guth; Knot distorsion]** A remarkable application of expander graphs to the construction of "complicated" knots was uncovered by Gromov and Guth [48, §4]. Again, the statement is easily understood, even if the connection with expanders is rather hard to see at first! We see a knot as the image of a smooth map $k : [0,1] \longrightarrow \mathbf{R}^3$, which is injective on $[0,1[$ and satisfies $k(0) = k(1)$. Gromov introduced an invariant, called the *distorsion*, to measure the complexity of a knot: it compares the distance between points of k when computed by following the knot itself, and by seeing the points as being in $\mathbf{R}^3$, and is given formally by

$$\mathrm{dist}(k) = \sup_{0 \leqslant s \neq t \leqslant 1} \frac{d_k(k(s), k(t))}{\|k(s) - k(t)\|},$$

where the norm on the denominator is the euclidean distance, while the distance in numerator is the intrinsic distance on the image of k, i.e., the infimum of the length of a curve $\gamma : [0,1] \longrightarrow \mathbf{R}^3$ such that the image of γ is contained in the image of k, and such that $\gamma(0) = x$, $\gamma(1) = y$.

A question of Gromov was to construct knots k such that $\mathrm{dist}(k)$ is large, and in fact such that it is large even if k is deformed arbitrarily. Formally, this means that one considers the "intrinsic" distorsion of k as defined by

$$\mathrm{idist}(k) = \inf_{k'} \mathrm{dist}(k'),$$

where k' runs over all knots that can be obtained from k by "smooth" deformation. As Gromov and Guth explain, it is very hard to construct knots with $\mathrm{idist}(k)$ arbitrarily large. The first examples were found by Pardon [93], and they are explicit but very special. Gromov and Guth show that sequences of knots k_n constructed using special manifolds which are related to certain types of expander graphs always have a large distorsion, and in fact that if k_n comes from a family (Γ_n) of expander graphs, we have

$$\mathrm{idist}(k_n) \geqslant c|\Gamma_n|,$$

for some constant $c > 0$, so that the distorsion grows linearly with the number of vertices of the graphs; it is, as Gromov and Guth show, as fast as the distorsion of these knots can grow.

It is important to notice in this example, in contrast with the first one, that in order to obtain the desired conclusion, it is not enough to know that expander graphs merely *exist*: the construction of knots is based on quite special sequences of graphs, and these must be expanders. This feature is shared by the next two examples, and motivates much of the recent work on expander graphs, and in particular the results discussed in Chapter 6.

We will say more about this example, with a sketch of the strategy of the proof, in Example 5.4.7.

- [**Sieve in discrete groups**] Classical sieve methods are concerned with establishing multiplicative properties of integers arising from additive constructions, typically to attempt to find prime numbers in sequences such as polynomial values $P(n)$ (e.g., are there infinitely many primes of the form n^2+1 for some integer n?) or shifted primes $p+k$ (e.g., $k = 2$ corresponds to the twin primes problem). Starting from works of Bourgain, Gamburd and Sarnak [13], these problems have been extended to very different settings, where for instance one considers the multiplicative properties of integers obtained as values $P(a_{1,1}, \ldots, a_{n,n})$ for some fixed polynomial P of n^2 variables evaluated at the coordinates $(a_{i,j})$ of a matrix in $\mathrm{SL}_n(\mathbf{Z})$, or in some subgroup of $\mathrm{SL}_n(\mathbf{Z})$. Implementing the classical sieve ideas for such problems turns out to be unavoidably related to proving that certain families of finite graphs, which are so-called Cayley graphs of congruence quotients of the underlying group, are expanders. Here the results of Bourgain–Gamburd, and their generalizations, are therefore necessarily of the highest importance. Moreover, an impressive recent series of works of Bourgain and Kontorovich (see, e.g., the survey [66] of Kontorovich and their recent papers) goes even beyond the "elementary" consequences of expansion.

  There are other sieve methods, of a rather different nature, which can lead to applications that do not involve arithmetic at all. One example is the *large sieve* in discrete groups, which we will discuss in Section 5.3, proving a special case of a beautiful result of Lubotzky and Meiri [81] concerning the "sparsity" of proper powers (elements of the form g^m for some $m \geqslant 2$) in finitely generated linear groups with sufficiently expanding quotients.

- [**Arithmetic properties in families**] Readers with some interest and knowledge of arithmetic geometry are probably fond of many results (or problems) of the following kind: for each algebraic number t, we have an associated "arithmetic object" A_t, defined over some finite extension k_t of the coefficient field $\mathbf{Q}(t)$. We believe that constructing A_t should involve solving some polynomial equations of large degree (that depend on t), and so k_t should, in general, be a proper extension of $\mathbf{Q}(t)$, and even a large degree extension. But miracles can happen, polynomials can have surprising factorizations, and sometimes it may turn out that $k_t = \mathbf{Q}(t)$. But how often can such miracles *really* happen? For instance, take the arithmetic object A_t to be simply one of the 1009-th roots of the algebraic number $1 - t^{1009}$. To say that $k_t = \mathbf{Q}(t)$ in that case means that $1 - t^{1009}$ is a 1009-th power of an element of $\mathbf{Q}(t)$, i.e., that there is an element

$\beta \in \mathbf{Q}(t)$ such that $t^{1009} + \beta^{1009} = 1$. If we only consider $t \in \mathbf{Q}$, the miraculous values of t are those for which Fermat's equation for the prime 1009 has a solution! We know there is none other than $t = 0$ and $t = 1$ – but it cannot be claimed that this knowledge was easily gained.

For certain specific types of families of arithmetic objects, related to certain Galois representations, Ellenberg, Hall and myself [34] proved the rather strong finiteness property that, for any given d, there are only finitely many t with $\mathbf{Q}(t)$ of degree $\leqslant d$ such that $k_t = \mathbf{Q}(t)$. For instance, if ℓ is a large enough prime number, the set of algebraic numbers t of degree $\leqslant 2$ such that the Galois actions on the ℓ-torsion points of the elliptic curves

$$y^2 = x(x-1)(x-t), \qquad y^2 = x(x-1)(x-2t)$$

are isomorphic is *finite*. And this all comes from expansion properties of certain graphs... Remarkably, one absolutely needs here the latest developments concerning expansion in Cayley graphs of finite linear groups – see Examples 5.5.6 and 5.5.9.

We will discuss these applications (with varying degree of detail) in Chapter 5; some of them, and other examples, are also discussed in the survey of Lubotzky [79].

And to conclude this introduction, it is worth pointing out that some very lively activity is ongoing concerning *higher-dimensional* analogues of expander graphs. We simply refer to Lubotzky's survey [80], which explains some of the motivations and problems in finding the most suitable definition for these objects. Strikingly, for certain natural-looking definitions, even the analogue of the simple fact that a complete graph is a good expander is a very deep result! In a few years, one may hope that there will exist another book, with similarly wide-ranging examples and applications, concerning these new objects!

Prerequisites and notation

For the most part of this book, we only require basic linear algebra (including finite-dimensional Hilbert spaces), basic algebra (elementary group theory and some properties of finite fields), calculus and some elementary probability. Although a number of concepts are presented with terminology influenced by functional analysis (L^2-spaces in particular), the setting is most often that of finite-dimensional spaces. However, we will use more advanced concepts and results in some of the surveys of applications in Chapter 5. We always give references, and in the places where we give complete arguments (and not sketches of proofs), we try to keep these prerequisites at a minimum level.

We will use the following notation:

(1) For a set X, $|X| \in [0, +\infty]$ denotes its cardinal, with $|X| = \infty$ if X is infinite. There is no distinction in this text between the various infinite cardinals.

(2) For a subset Y of a set X, we denote by $\mathbf{1}_Y$ the characteristic function of Y. If Z is also a subset of X, we also denote by $Y - Z$ the set of $y \in Y$ such that $y \notin Z$.

(3) Given a group G, we denote by $[G, G]$ the *commutator group* of G, which is generated by all commutators $[g, h] = ghg^{-1}h^{-1}$ (note that not all elements of $[G, G]$ are themselves commutators, see for instance [72, Example 4.4.5 (1)]). The subgroup $[G, G]$ is normal in G, and the quotient group $G/[G, G]$ is abelian; it is called the *abelianization* of G. If $[G, G] = G$, then G is called *perfect*.

(4) For a group G and an element $x \in G$, we denote by $C_G(x)$ the centralizer of x in G, i.e., the subgroup of elements $y \in G$ such that $xy = yx$. For a subgroup H of G, we denote by $N_G(H)$ its normalizer, the subgroup of all $x \in G$ such that $xHx^{-1} = H$.

(5) The symmetric group on k letters is denoted $\mathfrak{S}_k$.

(6) We denote by $\mathbf{F}_p$ the finite field $\mathbf{Z}/p\mathbf{Z}$, for p prime, and more generally by $\mathbf{F}_q$ a finite field with q elements, where $q = p^n$, $n \geqslant 1$, is a power of p. We will recall the basic facts that we need when we require them.

(7) When considering a normed vector space E, we usually denote the norm by $\|v\|$, and sometimes write $\|v\|_E$, when more than one space (or norm) are considered simultaneously. When considering a Hilbert space H, we denote by $\langle \cdot, \cdot \rangle$ the inner product. We use the convention that the inner product is linear in the first variable, and conjugate-linear in the other, i.e., we have

$$\langle \alpha v, w \rangle = \alpha \langle v, w \rangle, \qquad \langle v, \alpha w \rangle = \bar{\alpha} \langle v, w \rangle,$$

for two vectors v, w and a scalar $\alpha \in \mathbf{C}$.

(8) If H is a Hilbert space, then an endomorphism u of H is said to be *positive* if $\langle u(x), x \rangle \geqslant 0$ for all $x \in H$. (This corresponds to what are often called *positive semi-definite* matrices).

(9) A *unitary representation* of a group G is a homomorphism $\varrho \colon G \to \mathrm{U}(E)$, where $\mathrm{U}(E)$ is the group of unitary transformations of a Hilbert space E. If the group

G has a topology compatible with the group structure, we also ask that the map

$$\begin{cases} G \times E \to E \\ (g, x) \mapsto \varrho(g)x \end{cases}$$

be continuous.

(10) A unitary representation $\varrho\colon G \to \mathrm{U}(E)$ is said to be irreducible if $E \neq 0$ and if there is no closed subspace $F \subset E$ with $F \neq 0$ and $F \neq E$ that is stable under all linear maps $\varrho(g)$. A closed subspace $F \subset E$ that is stable under these linear maps is called a *subrepresentation* of ϱ.

(11) If X is a set and $f\colon X \to \mathbf{C}$ and $g\colon X \to [0, +\infty[$ are functions on X, we write $f \ll g$, or $f = O(g)$, if there exists a real number $C \geqslant 0$ such that $|f(x)| \leqslant Cg(x)$ for all $x \in X$. We also say that C is the "implied constant", and we will usually specify how it may depend on other parameters. If f and g are both non-negative, we write $f \asymp g$ if $f \ll g$ and $g \ll f$.

(12) If X is a topological space, $x_0 \in X$ and f, g are complex-valued functions on X, we write $f \sim g$ as $x \to x_0$ if f/g is defined for x in a neighborood of x_0 and if $\lim_{x \to x_0} f(x)/g(x) = 1$.

(13) In Section 5.3, we will use the Prime Number Theorem, that states that the number $\pi(x)$ of prime numbers $p \leqslant x$ satisfies

$$\pi(x) \sim \frac{x}{\log x} \tag{1.1}$$

as $x \to +\infty$.

CHAPTER 2

Graphs

2.1. Graphs

We consider graphs of a certain specific type, for reasons that will be clear (and that we will point out explicitly): unoriented graphs, where loops based at a vertex and multiple edges are permitted. The definition needs to be chosen carefully to give a fully rigorous expression to this idea, but there is more than one way to do it, so one should see this definition as specifying a specific "encoding" of the intuitive notion that we want to use, and not as the only way to define graphs. We will mention a few other options at some points.

Definition 2.1.1 (Graph). A *graph* Γ is given by a triple (V, E, ep) where V and E are arbitrary sets, called respectively the set of vertices of Γ and the set of edges of Γ, and

$$\mathrm{ep} : E \longrightarrow V^{(2)}$$

is an arbitrary map, called the *endpoint map*, where $V^{(2)}$ denotes the set of subsets $e \subset V$ of cardinality either 1 or 2.

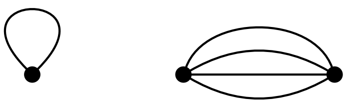
If $\alpha \in E$ is an edge of Γ, the elements of $\mathrm{ep}(\alpha)$ are called *extremities* of α. If $\alpha \neq \beta$ are distinct edges of Γ, they are called *adjacent* at a vertex $x \in V$ if $x \in \mathrm{ep}(\alpha) \cap \mathrm{ep}(\beta)$ is a common extremity.

Given a vertex $x \in V$, the number of edges α such that x is an extremity, i.e., such that $x \in \mathrm{ep}(\alpha)$, is called the *degree* or *valency* of x, denoted $\mathrm{val}(x)$. If the valency is the same, say equal to $d \geqslant 0$, at all vertices, the graph is called *regular*, or *d*-regular.

A graph is *finite* when both V and E are finite; it is *countable* if both V and E are countable.

This definition is the same as, for instance, in the textbook of Bondy and Murty [9] (who call ep the *incidence map*), and in the graph theory sections of the book of Ceccherini-Silberstein, Scarabotti and Tolli [26, Ch. 8].

Remark 2.1.2. (1) The intuition should be clear, as the terminology indicates: to express a graph (say, one drawn on paper) in this form, one takes as set of edges the "physical" ones, and one defines $\mathrm{ep}(\alpha)$ to be the set of extremities of such an edge. This allows *loops*, which are edges where $\mathrm{ep}(\alpha) = \{x\}$ is a singleton (the loop is then based at x, of course), as well as multiple edges with the same endpoints, say $\alpha_1 \neq \alpha_2$ with $\mathrm{ep}(\alpha_1) = \mathrm{ep}(\alpha_2) = \{x, y\}$.

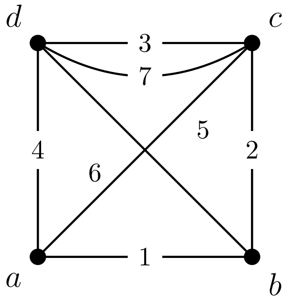


Conversely, to "draw" a graph Γ coded as a triple (V, E, ep), we can draw the points of V, then for each $\alpha \in E$, we look at $\mathrm{ep}(\alpha)$ and draw either (1) a loop from x to x if $\mathrm{ep}(\alpha) = \{x\}$ is a single element, or (2) an arc (without orientation) from x to y if $\mathrm{ep}(\alpha) = \{x, y\}$ with $x \neq y$.

For instance, consider the graph with $V = \{a, b, c, d\}$, $E = \{1, 2, 3, 4, 5, 6, 7\}$, and

$$\mathrm{ep}(1) = \{a, b\}, \quad \mathrm{ep}(2) = \{b, c\}, \quad \mathrm{ep}(3) = \{c, d\}, \quad \mathrm{ep}(4) = \{a, d\},$$
$$\mathrm{ep}(5) = \{a, c\}, \quad \mathrm{ep}(6) = \{b, d\}, \quad \mathrm{ep}(7) = \{c, d\},$$

and check that it can be represented as



As in this figure, it is not always possible to draw the edges without some overlap (graphs for which it is possible are called *planar*; see for instance [9, Ch. 10] for a discussion of properties and characterizations of planar graphs). However, for any finite graph, it is possible to "draw" it in $\mathbf{R}^3$ without overlap. This should be fairly clear intuitively, and the reader should attempt to see what is involved in a rigorous proof. (Basically, $\mathbf{R}^3$ minus a finite number of smooth compact curves, seen as images of maps $\gamma : [0, 1] \longrightarrow \mathbf{R}^3$, is path-connected.)

(2) If Γ has no loops (which means that every set of endpoints $\mathrm{ep}(\alpha)$ contains two elements) and no multiple edges (so that ep is an injection of E into the set of subsets of order 2 in V), the graph is called *simple*. In that case, the set of edges can also be identified with a subset $R \subset V \times V$ such that $(x, y) \in R$ if and only if $(y, x) \in R$ (expressing the fact that edges are not oriented) and such that $(x, x) \notin R$ for all $x \in V$ (expressing the absence of loops). This is a more common way of "coding" simple graphs.

This point of view is sufficient for many purposes, and it is used in many common textbooks, such as that of Diestel [33]. However, for our purposes, it is sometimes very important to allow multiple edges and loops (see, for instance, Example 2.3.2 (5), where we use a graph with a single vertex and many loops to identify the Cayley graph of a free group). We also prefer not to impose a specific relation between V and E, except for the existence of the endpoint map, because it allows us (for instance) to keep track of possible structures in the set of edges between two vertices.

We will sometimes omit mention of ep when considering a simple graph, viewing the edges as a set of subsets of V with two elements.

(3) We sometimes write $\mathrm{ep}(\alpha) = \{x, y\}$, without specifying that $x \neq y$: this is a convenient manner of designating the endpoints of an edge without distinguishing between loops and proper edges. If we want simply to state that two vertices x and y are joined by (at least) an edge, we will also write $x \sim y$.

(4) By convention, for a graph Γ, we write $|\Gamma| = |V|$: the "size" of Γ is identified with the number of vertices. We also sometimes write $x \in \Gamma$ to mean $x \in V$. Along the same lines, we will sometimes write V_Γ for the set of vertices (resp. E_Γ for the set of edges) of a graph Γ (or even just V and E when there is no possible ambiguity).

(5) Serre [103, I.2] defines a graph as a tuple (V, E, o, e, i) where V is the set of vertices, E is a set of (oriented) edges, $o \colon E \to V$ and $e \colon E \to V$ are two maps (giving the origin and end of an edge), and $i \colon E \to E$ is an involution without fixed points (i.e., $i(i(x)) = x$ and $i(x) \neq x$ for all $x \in E$) such that $o(i(x)) = e(x)$ and $e(i(x)) = o(x)$.

Intuitively, the oriented edges always come in pairs with opposite orientation, and i is the map that sends an edge from x to y to the opposite edge from y to x. This definition might seem unintuitive at first, but it has advantages when it comes to "functorial" constructions in particular (this is crucial in Serre's book [103]). The graph in the sense of Definition 2.1.1 associated to such a tuple (V, E, o, e, i) is $(V, E/i, \mathrm{ep})$, where E/i is the set of pairs $\{x, i(x)\}$ for $x \in E$, and $\mathrm{ep}(\{x, i(x)\}) = \{o(x), e(x)\}$. We will make occasional remarks later on about this definition (see also [10, Ch. I, §2]).

Exercise 2.1.3 (Number of edges vs. number of vertices). Show that if $\Gamma = (V, E, \mathrm{ep})$ is a finite graph, we have

$$\sum_{x\in V} \mathrm{val}(x) = 2|E_2| + |E_1|$$

where $E_i = \{\alpha \in E \mid \alpha \text{ has } i \text{ extremities}\}$, i.e., $|E_1|$ is the number of loops and $|E_2|$ the number of edges joining distinct vertices.

In order to encode a finite graph, one can also use its *adjacency matrix*:

Definition 2.1.4 (Adjacency matrix). Let Γ be a finite graph. The *adjacency matrix* $A_\Gamma = (a(x, y))$ is the matrix with rows and columns indexed by V_Γ and with $a(x, y)$ equal to the number of edges with extremities (x, y), formally

$$a(x, y) = |\{\alpha \in E_\Gamma \mid \mathrm{ep}(\alpha) = \{x, y\}\}|.$$

Note that the adjacency matrix is always symmetric (in the sense that $a(x, y) = a(y, x)$), which reflects our use of *unoriented* edges. It is easy to go in the opposite direction: given any symmetric "matrix" $A = (a_{x,y})$ with rows and columns indexed by a finite set V and non-negative integral coefficients $a_{x,y}$, one defines a finite graph with adjacency matrix A by taking V as set of vertices and

$$E = \{(x, y, i) \in V \times V \times \mathbf{Z} \mid a_{x,y} \neq 0 \text{ and } 1 \leqslant i \leqslant a_{x,y}\}$$

with

$$\mathrm{ep}(x, y, i) = \{x, y\}$$

for all $(x, y, i) \in E$ (for instance: there are other choices).

Exercise 2.1.5. Devise at least one other way of formalizing the "same" class of graphs as in our definition.

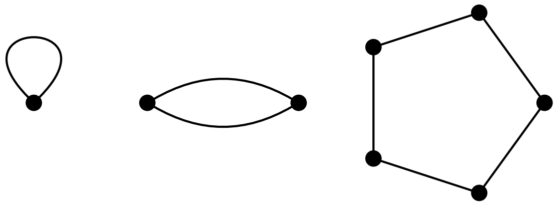
Example 2.1.6. Here are some elementary examples of "coding" for various families of graphs using Definition 2.1.1. The examples will be used many times in this chapter and the next in order to illustrate some basic concepts.

(1) [Cycle] Let $m \geqslant 1$ be an integer. The m-cycle C_m is the graph with vertices $V_m = \mathbf{Z}/m\mathbf{Z}$, edges $E_m = \mathbf{Z}/m\mathbf{Z}$, and endpoint map given by

$$\mathrm{ep}(i) = \{i, i + 1\}$$

for $i \in \mathbf{Z}/m\mathbf{Z}$. In other words, except when $m = 1$ (in which case the cycle is a single loop based at 0), there are two edges adjacent to any given $i \in V_m$: the edges coded by $i - 1$, and the one coded by i itself.

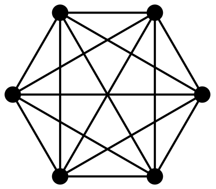
Here are the graphs for $m = 1$, $m = 2$ and $m = 5$:



(2) [Path] Let $m \geqslant 0$ be an integer. The *path of length m*, denoted P_m, is the graph with vertex set $V_m = \{0, \ldots, m\}$ and edge set $E_m = \{1, \ldots, m\}$, where $\mathrm{ep}(i) = \{i-1, i\}$ for $1 \leqslant i \leqslant m$. A path of length 0 is a graph with a single vertex and no edges. Here is the path of length 4:

We often say, somewhat abusively, that the vertices 0 and m are the *extremities* of the path.

(3) [Complete graph] Let again $m \geqslant 1$ be an integer. The *complete graph K_m* with m vertices has also $V_m = \{1, \ldots, m\}$ but now $E_m = \{(x, y) \in V_m \mid x < y\}$, with $\mathrm{ep}((x, y)) = \{x, y\}$. In other words, each pair of distinct vertices is joined by (exactly) one edge. Here is the complete graph K_6:

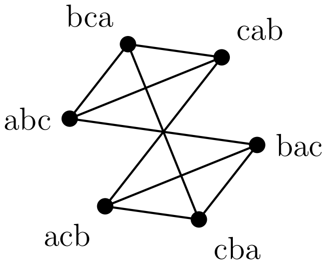


All graphs in these first examples are simple graphs, *except* for the cycles C_1 and C_2. Most of them are regular: C_1 is 1-regular, C_m is 2-regular for $m \geqslant 2$; P_0 is 0-regular, P_1 is 1-regular (but P_k is not regular for $k \geqslant 2$); K_m is $(m-1)$-regular for all $m \geqslant 1$.

(4) [A Cayley graph] Our last sequence of examples is less obvious, but it illustrates the type of graphs that will be one of the main focus of these notes, starting from the end of Chapter 3: *Cayley graphs associated to finite groups* (see Section 2.3 for the general definition).

However, we do not need to mention groups immediately in this example. Following Diaconis and Saloff-Coste [32], we fix $n \geqslant 3$ and take as vertex set V_n all the possible arrangements of a deck D_n of n cards (so there are $n!$ elements in V_n). Then we define G_n as the simple graph where the vertex set is V_n and the edges correspond to either exchanging the top two cards (connecting, say, $(a, b, c, d) \in V_4$, to (b, a, c, d)), or bringing the bottom card to the top, or conversely (connecting, say $(a, b, c, d) \in V_4$ to (d, a, b, c) – bottom to top – and (a, b, c, d) to (b, c, d, a) – top to bottom.)

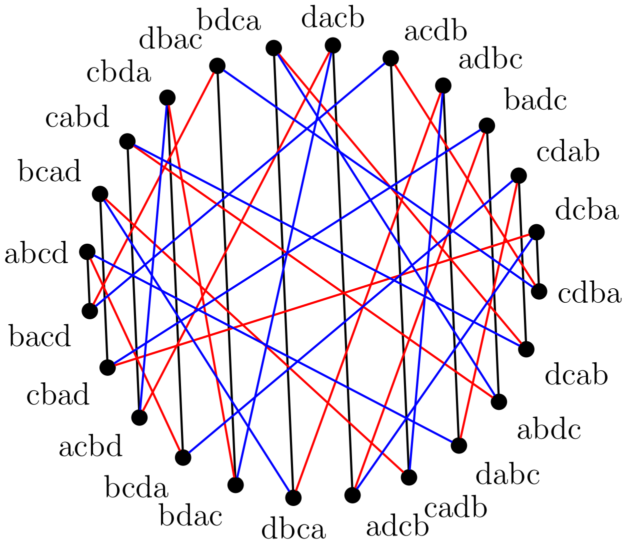
Thus, by definition, G_n is a 3-regular graph for each $n \geqslant 3$, with $n!$ vertices. Here is an illustration of G_3, with the deck $D_3 = \{a, b, c\}$, and in Figure 2.1 one of G_4, with deck $D_4 = \{a, b, c, d\}$ (it is by far the most complicated graph we will draw...).



Exercise 2.1.7. Transcribe these examples using Serre's definition (Remark 2.1.2 (5)); write down their adjacency matrices, and express them in your own coding (Exercise 2.1.5).

Here are, furthermore, some elementary constructions with graphs (others, such as the universal cover, or path graphs, will occur later).

FIGURE 2.1. The graph G_4



EXAMPLE 2.1.8. (1) [Disjoint union] Let (Γ_i) be an arbitrary family of graphs, with $\Gamma_i = (V_i, E_i, \mathrm{ep}_i)$. The *disjoint union* Γ of the graphs Γ_i has vertex set the disjoint union of the V_i's, edge set the disjoint union of the E_i's, and for an edge α, that belongs to some E_i, we put $\mathrm{ep}(\alpha) = \mathrm{ep}_i(\alpha)$. The intuitive meaning is clear: we are just viewing the collection of drawings representing the various graphs Γ_i as a single bigger graph.

(2) [Removing vertices] Let $\Gamma = (V, E, \mathrm{ep})$ be a graph and let X be a subset of V. We define a graph $\Gamma - X = (V', E', \mathrm{ep}')$ with vertex set $V - X$, edges given by the edges of Γ with no extremity in X (i.e., $E' = \{\alpha \in E \mid \mathrm{ep}(\alpha) \cap X = \emptyset\}$), and ep' is the restriction to E' of ep. If $X = \{v_0\}$ has a single element, we often write $\Gamma - v_0$ instead of $\Gamma - \{v_0\}$. We say that $\Gamma - X$ is obtained by removing from Γ the vertices in X.

(3) The easiest way to define certain graphs is by specifying the vertex set and the sets of edges between any two vertices. In other words, we begin with a set V and with sets $E(x, y)$ for each $(x, y) \in V \times V$, which are disjoint as (x, y) varies and satisfy $E(x, y) = E(y, x)$ for all (x, y) (since the edges are not oriented). Then we define E to be the disjoint union of these sets, and the endpoint map so that $\mathrm{ep}(\alpha) = \{x, y\}$ if the edge α belongs to the subset $E(x, y)$ of E.

Having selected with Definition 2.1.1 a specific type of coding for graphs has, at least, the advantage that it prompts us to give quickly a definition of what it means for two graphs to be "the same": in the examples above, we selected a specific set of vertices, but these could obviously be replaced by any set with the same number of elements, provided the edges are also "transported" to refer to this new set. Similarly, the specific sets of edges are just convenient labelings, and other sets could be equally suitable. This leads to the definition of isomorphism of graphs or, more generally, to *maps* (or *morphisms*) of graphs:

DEFINITION 2.1.9 (Maps of graphs). Let Γ_1 and Γ_2 be graphs. A *morphism*, or *graph map*, from Γ_1 to Γ_2 is a pair (f, f_*) where

$$f : V_{\Gamma_1} \longrightarrow V_{\Gamma_2}$$

is a map between the vertex sets and

$$f_* : E_{\Gamma_1} \longrightarrow E_{\Gamma_2}$$

is a map between the edges, such that

$$\mathrm{ep}(f_*(\alpha)) = f(\mathrm{ep}(\alpha)) \tag{2.1}$$

for all $\alpha \in E_{\Gamma_1}$. In other words: an edge α between x and y is sent to an edge $f_*(\alpha)$ with extremities $f(x)$ and $f(y)$. We most often simply write f for such a map, using f_* for the edge map.

If the graphs are simple, then the companion edge-map f_* is uniquely specified by f itself: in that case, whenever there is an edge e between x and y, it is unique, and there must also be an edge between $f(x)$ and $f(y)$, which determines $f_*(e)$. However, in the presence of multiple edges, we must specify where each individual edge between x and y goes.

The following definitions and facts are again easy and fairly formal, but are extremely important:

Definition 2.1.10. (1) Let Γ be a graph. The *identity map* $\Gamma \to \Gamma$ of Γ is the pair $(\mathrm{Id}_V, \mathrm{Id}_E)$, and is denoted Id_Γ.

(2) For any graphs Γ_1, Γ_2, Γ_3 and maps

$$\Gamma_1 \xrightarrow{(f,f_*)} \Gamma_2 \xrightarrow{(g,g_*)} \Gamma_3,$$

the *composite map* is defined by the pair $(g \circ f, g_* \circ f_*)$. We simply write $g \circ f$ for this map.

(3) The following properties hold:

$$h \circ (g \circ f) = (h \circ g) \circ f$$

for any three maps that can be composed, and if $f : \Gamma_1 \longrightarrow \Gamma_2$, we have

$$f \circ \mathrm{Id}_{\Gamma_1} = f, \qquad \mathrm{Id}_{\Gamma_2} \circ f = f.$$

Remark 2.1.11. In the language of categories, this states that there is a category of graphs where objects are graphs, and morphisms are graph maps. We will not use this language very extensively, but we will use remarks (like this one) to indicate the interpretation of certain facts in these terms.

Exercise 2.1.12. Using the coding of graphs you obtained in Exercise 2.1.5, define what is a graph map, and check that these maps correspond to those defined above. Do the same with Serre's definition of Remark 2.1.2 (5). (In the language of categories, you should be able to find an equivalence of categories between "your" category of graphs and the one we defined.)

Example 2.1.13. Let $\Gamma = (V, E, \mathrm{ep})$ be an arbitrary graph. We can associate to it a simple graph $\Gamma^s = (V^s, E^s)$ as follows: we use the same set of vertices V, but remove all loops and all multiple edges from E. This means $V^s = V$, and

$$E^s = \{\{x, y\} \subset V \mid x \neq y \text{ and there exists } \alpha \in E \text{ with } \mathrm{ep}(\alpha) = \{x, y\}.\}, \tag{2.2}$$

If Γ has no loops, there is a canonical[1] map $\Gamma \to \Gamma^s$, which is the identity on the vertices and which maps an edge to the unique edge in Γ^s with the same extremities. There is also always a graph map $\Gamma^s \to \Gamma$ which is the identity on vertices, but it is not canonical, since it must map an edge in Γ^s to some, arbitrarily chosen, edge in Γ between the same extremities, so that it is not unique if multiple edges exist. If Γ has a loop, there is no map $\Gamma \to \Gamma^s$ at all.

[1] "Canonical" here roughly means that it is defined without making any choices.

More definitions:

Definition 2.1.14 (Isomorphism, automorphism, embedding). (1) A graph map $f : \Gamma_1 \to \Gamma_2$ is an *isomorphism* with inverse g if and only if $f \circ g = \mathrm{Id}_{\Gamma_2}$ and $g \circ f = \mathrm{Id}_{\Gamma_1}$. If $\Gamma = \Gamma_1 = \Gamma_2$, then f is called an automorphism of Γ.

(2) The inverse of an isomorphism is unique and is denoted f^{-1}. In fact, a morphism (f, f_*) is an isomorphism if and only if f and f_* are both bijections, and then $(f, f_*)^{-1} = (f^{-1}, f_*^{-1})$. In particular, the inverse of (f, f_*) is also an isomorphism. Moreover, the composite of two isomorphisms is also an isomorphism; hence the set of automorphisms of Γ, with the composition law, is a group, which is denoted $\mathrm{Aut}(\Gamma)$.

(3) An *embedding* $\Gamma_1 \hookrightarrow \Gamma_2$ is a graph map (f, f_*) such that f and f_* are both injective. If Γ_1 and Γ_2 and both simple, it suffices that the vertex map $f\colon V_1 \to V_2$ is injective.

Remark 2.1.15. These are fairly dry and formal definitions. Their meaning is quite clear: to say that two graphs are isomorphic through f means exactly that their vertices are in bijection, and that for any two vertices on the first graph, the edge map gives a bijection between the edges that join them on both graphs. This corresponds to the intuitive idea of changing the labeling of the vertices and edges while respecting the graph structure.

Similarly, to say that (f, f_*) is an embedding means that the vertices of Γ_1 can be identified with a subset of those of Γ_2, and that the edges between any two vertices in Γ_1 are then a subset of those in Γ_2 (where there could be more edges, of course.)

Example 2.1.16. (1) If Γ^s is the simple graph associated to a graph Γ, as in Example 2.1.13, the maps $\Gamma^s \to \Gamma$ described in this example are all embeddings of Γ^s in Γ.

Moreover, if Γ is itself a simple graph, there is a canonical isomorphism $\Gamma \xrightarrow{\sim} \Gamma^s$ (although the sets of edges E and E^s defined in (2.2) might not be identical) given by the identity on vertices and $f_*(\alpha) = \mathrm{ep}(\alpha) \in E^s$ for $\alpha \in E$.

(2) The path P_k, for $k \geqslant 1$, has a non-trivial automorphism f (in fact an involution, i.e., we have $f \circ f = \mathrm{Id}$) which is intuitively given by "reversing the path", and can be defined formally by

$$f(i) = m - i, \qquad f_*(j) = m + 1 - j$$

for any vertex $i \in V_m = \{0, \ldots, m\}$ and edge $j \in E_m = \{1, \ldots, m\}$. To check the definition of a graph map, note that

$$\mathrm{ep}(f_*(j)) = \mathrm{ep}(m + 1 - j) = \{m - j, m - j - 1\} = f(\{j, j + 1\}) = f(\mathrm{ep}(j)),$$

and since f and f_* are both involutions, (f, f_*) is an isomorphism equal to its own inverse. (For $m = 0$, the definition "works" but it is the identity of the graph with a single vertex and no edges...)

(3) [Subgraphs] Let $\Gamma = (V, E, \mathrm{ep})$ be a graph. For any subset $V' \subset V$ of vertices, and any subset $E' \subset E$ of edges with extremities lying in V' (i.e., such that $\mathrm{ep}(\alpha) \subset V'$ for any $\alpha \in E'$), the pair of inclusions $(V' \hookrightarrow V, E' \hookrightarrow E)$ is an embedding of the graph (V', E', ep) inside (V, E, ep). We then say that (V', E', ep) is a *subgraph* of Γ.

If E' is the set of *all* edges with extremities in V', i.e., if E' is defined to be

$$E' = \{\alpha \in E \mid \mathrm{ep}(\alpha) \subset V'\},$$

we say that (V', E', ep) is a *full subgraph* of Γ. Such subgraphs are therefore in one-to-one correspondence with subsets of V.

(4) Let k and m be non-negative integers with $m \leqslant k$. For any integer $n \geqslant 0$ such that $n + m \leqslant k$, we have a graph map

$$f_n \colon P_m \to P_k$$

defined by $f(i) = i + n$ for $0 \leqslant i \leqslant m$ and $f_*(i) = i + n$ for $1 \leqslant i \leqslant m$. Intuitively, f_n runs over the part of the path of length k from m to $m + n$.

Embeddings or other graph maps can frequently be used to define invariants and distinguish special families of graphs. Here is an important example:

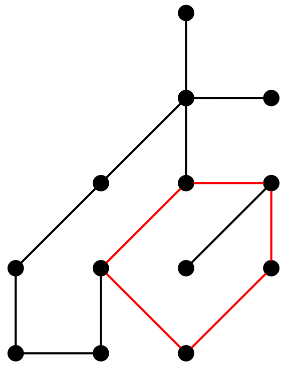
Definition 2.1.17 (Girth). Let $\Gamma = (V, E, \mathrm{ep})$ be a graph.

(1) For $m \geqslant 1$, a *cycle of length* m in Γ is an embedding $C_m \to \Gamma$.

(2) The *girth* of Γ is the smallest integer $m \geqslant 1$ such that there exists at least one cycle of length m in Γ, or $+\infty$ if no cycle exists at all in Γ. We denote this integer $\mathrm{girth}(\Gamma)$.

Example 2.1.18. The girth of the cycle C_m itself is equal to m. Moreover, Γ has girth 1 if and only if Γ has at least one loop, and it has girth 2 if and only if it has no loop, but there are two distinct vertices which are joined by at least two edges. Similarly, having girth 3 means there are no loops, no multiple edges, but there exists a *triangle* in Γ, i.e., three distinct vertices x_1, x_2, x_3 and three edges α_1, α_2 and α_3 with α_1 joining x_1 and x_2, α_2 joining x_2 and x_3 and finally α_3 joining x_1 and x_3. (This is also equivalent to being a simple graph with an embedding of $K_3 = C_3$). For instance, the girth of K_m is infinite for $m = 1$ or 2, and is equal to 3 for $m \geqslant 3$. The reader is invited to check all these assertions...

Here is an example of graph with girth 5.



Exercise 2.1.19. (1) Let Γ be a finite d-regular graph with girth $g \geqslant 3$. Prove that

$$|\Gamma| \geqslant d(d-1)^{\lfloor (g-3)/2 \rfloor}.$$

(2) Show that the girth of a finite graph d-regular graph Γ with $d \geqslant 3$ is $\ll \log(|\Gamma|)$, where the implied constant depends only on d.

Example 2.1.20 (Trees and forests). Graphs with infinite girth have a name:

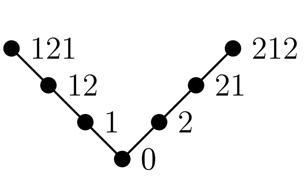
Definition 2.1.21 (Forests (and trees)). A graph Γ with infinite girth (i.e., there is no embedding $C_m \longrightarrow \Gamma$, for any $m \geqslant 1$) is called a *forest*. Anticipating the definition of connected graphs, a connected forest is called a *tree*.

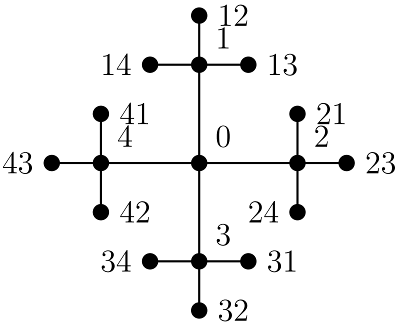
In particular, forests (and trees) are simple graphs. An example is the path P_k of length $k \geqslant 1$. Here are some more interesting examples. Fix some integers $d \geqslant 2$ and $k \geqslant 1$. The *finite rooted tree of degree d and depth k*, denoted $T_{d,k}$, is a simple graph defined by taking V to be the set of all words of length $\leqslant k$ (including the empty word, of length 0, which is called the "root" vertex of the tree) in the alphabet $A = \{1, \ldots, d\}$ with no letter repeated twice in a row, i.e.

$$V = \bigcup_{0 \leqslant j \leqslant k} \{(s_1, \ldots, s_j) \in A^j \mid s_i \neq s_{i+1} \text{ for } 1 \leqslant i \leqslant j-1\},$$

with edges between "neighboring" words, where w_1 is a neighbor of w_2 if w_2 can be obtained from w_1 either by adding a letter on the right (chosen among the $d-1$ letters distinct from the rightmost letter of w_1), or by removing the last letter.

Here are pictures of $T_{2,3}$ and $T_{4,2}$, with the vertices labeled with the corresponding words, which should clarify the matter. (Note that $T_{d,k}$ is *not* d-regular.)





One can extend this construction to infinite depth: the *d-regular tree* T_d, for $d \geqslant 2$, is the infinite graph with vertices given by all words of length $\geqslant 0$, without repeated letter, in the alphabet $\{1, \ldots, d\}$, and with edges described in the same way using neighboring words.

Exercise 2.1.22. Show that the number of vertices and edges of the finite tree $T_{d,k}$ are given by

$$|T_{d,k}| = d\frac{(d-1)^k - 1}{d-2} + 1, \quad |E_{d,k}| = |T_{d,k}| - 1 = d\frac{(d-1)^k - 1}{d-2}$$

if $d \geqslant 3$, and $|T_{2,k}| = 2k+1$, $|E_{2,k}| = 2k$.

One can also try to distinguish special graphs using (surjective) maps *to* another fixed one. Here is a classical notion that can be interpreted in this manner:

Definition 2.1.23 (Bipartite graph). A graph Γ is *bipartite* if there exists a partition $V_\Gamma = V_0 \cup V_1$ of the vertex set in two disjoint subsets, so that any edge has one extremity in V_0, and one in V_1, i.e., such that

$$\mathrm{ep}(\alpha) \cap V_0 \neq \emptyset, \qquad \mathrm{ep}(\alpha) \cap V_1 \neq \emptyset$$

for each $\alpha \in E_\Gamma$. One sometimes says that V_0 is the set of "inputs" and V_1 the set of "outputs".

Example 2.1.24. (1) A graph with a single vertex x and no edges is bipartite according to this definition, with $V_0 = \{x\}$ and $V_1 = \emptyset$. On the other hand, as soon as a bipartite graph Γ contains an edge, the subsets V_0 and V_1 are non-empty.

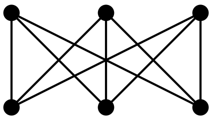
(2) The complete bipartite graph $K_{m,n}$ with $m \geqslant 1$ inputs and $n \geqslant 1$ outputs is the simple bipartite graph defined by the vertices

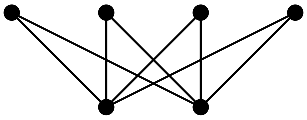
$$V_0 = \mathbf{Z}/m\mathbf{Z}, \qquad V_1 = \mathbf{Z}/n\mathbf{Z}, \qquad V = V_0 \cup V_1$$

(a disjoint union) and edges

$$E = \{\{x_0, x_1\} \subset V \mid x_0 \in V_0, \quad x_1 \in V_1\}.$$

Here are pictures of $K_{3,3}$ and $K_{2,4}$:

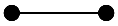




The reader can check, for instance, that the girth of $K_{m,n}$ is equal to 4 for m, $n \geqslant 2$, while it is infinite for $m = 1$ or $n = 1$.

We now have an easy proposition:

PROPOSITION 2.1.25 (Bipartiteness criterion). *A graph $\Gamma = (V, E, \mathrm{ep})$ is bipartite if and only if there exists a graph map $\Gamma \longrightarrow P_1$, where P_1 is the path of length 1:*



PROOF. We denote by $\{0, 1\}$ the two vertices of P_1 and by α_0 its unique edge. If Γ is bipartite, with a partition $V = V_0 \cup V_1$ in inputs and outputs, we can define rather obviously a graph map $f : \Gamma \to P_1$ by

$$f(x) = \begin{cases} 0 & \text{if } x \in V_0 \\ 1 & \text{if } x \in V_1, \end{cases}$$

and $f_*(\alpha) = \alpha_0$ for all $\alpha \in E_\Gamma$. This is indeed a graph map because for any $\alpha \in E_\Gamma$, we have

$$\mathrm{ep}(f_*(\alpha)) = \{0, 1\} = f(\mathrm{ep}(\alpha))$$

since, by definition, any edge α has one extremity in V_0 and one in V_1.

Conversely, let (f, f_*) be a graph map from an arbitrary graph Γ to P_1. Defining $V_0 = f^{-1}(0)$, $V_1 = f^{-1}(1)$, we have of course a disjoint union $V = V_0 \cup V_1$. Then we consider an arbitrary edge $\alpha \in E_\Gamma$. Its image $f_*(\alpha)$ has no choice but to be equal to the unique edge α_0, hence

$$\{0, 1\} = \mathrm{ep}(f_*(\alpha)) = f(\mathrm{ep}(\alpha)),$$

which is only possible if the extremities of α are two distinct elements, one in V_0 and the other in V_1. This means exactly that the partition $V = V_0 \cup V_1$ makes Γ into a bipartite graph. □

2.2. Metric, diameter, and so on

Our edges have, for the moment, not been really used, except as abstract elements. Of course, an edge is intuitively supposed to represent a way of going from one extremity to another. And if one goes from x to an adjacent vertex (or neighbor) y, there is no reason to stop there. Going further on longer adventures along the edges of a graph will lead us to the topic of expansion. But first, we explain how to measure how far we can go:

DEFINITION 2.2.1 (Paths and distance on a graph). Let $\Gamma = (V, E, \mathrm{ep})$ be a graph.

(1) A *path of length* $k \geqslant 0$ in Γ is a *graph map* $P_k \xrightarrow{\gamma} \Gamma$, i.e., an ordered sequence $(x_0, \ldots, x_k)$ of vertices of Γ, and an ordered sequence $(\alpha_1, \ldots, \alpha_k)$ of edges of Γ such that

$$\mathrm{ep}(\alpha_i) = \{x_{i-1}, x_i\}$$

for $1 \leqslant i \leqslant k$. If $k \geqslant 1$, the *extremities* of the path γ are the vertices $x = \gamma(0)$, $y = \gamma(k)$, where 0 and k denote the distinguished vertices of P_k which have a single adjacent vertex. One says that γ is a path from x to y, and one writes $\ell(\gamma) = k$ for its length.

(2) For any two vertices x, $y \in V$, the *distance on* Γ *between* x *and* y, denoted $d_\Gamma(x, y)$ is defined as the minimum length of a path between x and y, if such a path exists, or $+\infty$ otherwise, i.e.,

$$d_\Gamma(x, y) = \min\{\ell(\gamma) \mid \gamma \text{ is a path from between } x \text{ and } y\} \in \{0, 1, \dots\} \cup \{+\infty\}.$$

(3) The graph is *connected* if and only if $d_\Gamma(x, y)$ is finite for all x and $y \in V$, i.e., any two points can be joined by at least one path.

(4) A *geodesic* in Γ is a path γ such that the length of γ is equal to the distance in Γ between the extremities of γ.

Note that a path is allowed to "backtrack", since edges are unoriented, and that the vertices x_i might not be distinct. On the other hand, to compute the length, we need only look at paths that do not involve twice the same edge in succession.

Definition 2.2.2. Let Γ be a graph. A path $\gamma\colon P_k \to \Gamma$ of length $k \geqslant 0$ in Γ is *non-backtracking* if $\gamma_*(i) \neq \gamma_*(i + 1)$ for $1 \leqslant i \leqslant k - 1$, i.e., if the ordered sequence of edges corresponding to γ does not contain consecutively the same edge.

Remark 2.2.3. Note that this does not exclude that the same edge occurs multiple times, only that it should only do so after some intermediate edges.

There are two useful operations that can be performed on paths between various vertices, and which we define formally (their intuitive meaning is clear!):

- Given a path $\gamma\colon P_k \to \Gamma$ of length $k \geqslant 0$ from x to y, a non-negative integer $m \leqslant k$ and an integer $n \geqslant 0$ with $n + m \leqslant k$, the composition $\gamma \circ f_n$, where f_n is the graph map $f_n\colon P_m \to P_k$ of Example 2.1.16 (4), is a path of length m; it will be called the *restriction* of γ to the interval $\{m, \dots, m + n\}$.
- Suppose we are given paths γ_1 and γ_2 of lengths k_1 and k_2 respectively, such that γ_1 goes from x to z and γ_2 from z to y. We can then "concatenate" them to have a path from x to y. Formally, we define $\gamma_3\colon P_{k_1+k_2} \to \Gamma$ by

$$\gamma_3(i) = \begin{cases} \gamma_1(i) & \text{if } 0 \leqslant i \leqslant k_1, \\ \gamma_2(i - k_1) & \text{if } k_1 + 1 \leqslant i \leqslant k_1 + k_2, \end{cases}$$

for $0 \leqslant i \leqslant k_1 + k_2$ and

$$\gamma_{3,*}(i) = \begin{cases} \gamma_{1,*}(i) & \text{if } 1 \leqslant i \leqslant k_1, \\ \gamma_{2,*}(i - k_1) & \text{if } k_1 + 1 \leqslant i \leqslant k_2. \end{cases}$$

The assumption $\gamma_1(k_1) = \gamma_2(0)$ implies that γ_3 is indeed a graph map; since $\gamma_3(0) = x$ and $\gamma_3(k_1 + k_2) = y$, it is a path of length $k_1 + k_2$ from x to y, called the *concatenation* of γ_1 and γ_2.

The intuitive meaning of these definitions should be pretty clear. Their importance comes from connecting graphs with metric geometry:

Proposition 2.2.4. (1) *If* Γ *is a connected graph, the distance function* d_Γ *is a metric on* V, *i.e., it is non-negative and satisfies*

$$d_\Gamma(x, y) = d_\Gamma(y, x),$$
$$d_\Gamma(x, y) = 0 \text{ if and only if } x = y,$$
$$d_\Gamma(x, y) \leqslant d_\Gamma(x, z) + d_\Gamma(z, y)$$

for all vertices x, y, $z \in V$.

(2) *If we define an equivalence relation on* V *by*

$$x \sim y \iff d_\Gamma(x,y) < +\infty,$$

then the full subgraph of Γ *corresponding to an equivalence class* $V' \subset V$ *is a connected graph such that the distance* $d_{\Gamma'}$ *is the restriction of* d_Γ *to* $V' \times V'$*, and there are no edges with an extremity in* V' *and another outside* V'*. These subgraphs are called the* connected components *of* Γ*.*

Proof. Part (1) is intuitively clear, but we give details. The symmetry is because a path P_k can be reversed using the automorphism f of P_k (Example 2.1.16, (2); note in passing that this depends on the fact that the edges are unoriented). The map $\gamma \mapsto \gamma \circ f$ is then an involution (since f is an involution) between paths of length k from x to y and paths of length k from y to x, which implies $d_\Gamma(x,y) = d_\Gamma(y,x)$.

Further, $d_\Gamma(x,y) = 0$ if and only if there exists a path of length 0 from x to y; but a path $\gamma\colon P_0 \to \Gamma$ of length 0 has only one extremity, so that this holds if and only if $x = y$. Finally, the triangle inequality comes from the possibility of concatenating a path of length $k_1 = d_\Gamma(x,z)$ between x and z with one of length $k_2 = d_\Gamma(z,y)$ between z and y to obtain one of length $k_1 + k_2$ between x and y, as seen above, which shows that $d_\Gamma(x,y) \leqslant k_1 + k_2 = d_\Gamma(x,z) + d_\Gamma(z,y)$.

For (2), the fact that $\sim$ is an equivalence relation is elementary, and if V' is an equivalence class, we note that any edge $\alpha \in E$ has either all or no extremities in V': if $\mathrm{ep}(\alpha) = \{x,y\}$ with $x \in V'$, then the edge α shows (by definition) that $d_\Gamma(x,y) \leqslant 1$, so that $y \sim x$ is also in V'. Thus, if E' is the set of edges with an extremity in V', the graph (V', E', ep) is a full subgraph of Γ. Using a base vertex $x \in V'$, so that any $y \in V'$ is at finite distance to x, and the triangle inequality, we see that any two points of V' are at finite distance, i.e., (V', E', ep) is connected.

Moreover, since one can not connect elements of V' in Γ using edges others than those in E', we also see that the distance in Γ' is the restriction to $V' \times V'$ of d_Γ. □

Because of this construction, a number of classical invariants from metric geometry can be immediately "imported" into graph theory. We will consider in particular the *diameter*, and we recall the definition:

Definition 2.2.5 (Diameter of a graph). Let Γ be a graph. The diameter of Γ, denoted $\mathrm{diam}(\Gamma)$, is the largest distance between two vertices in Γ, i.e., we have

$$\mathrm{diam}(\Gamma) = \sup_{x,y \in V} d_\Gamma(x,y) \in \{0,1,2,\ldots\} \cup \{+\infty\}.$$

Example 2.2.6. If Γ is a finite connected, graph, its diameter will be finite. One of the key questions that the concept of expander graphs (hence, this book!) addresses is: given certain connected finite graphs, what can one say about their diameters? In particular, is this diameter relatively *small*, compared with the number of vertices?

We can immediately treat the obvious examples, among the graphs which were already described in Example 2.1.6:

- The path P_k has diameter k;
- The complete graph K_m has diameter 1 for $m \geqslant 2$ ($K_1 = P_0$ has diameter 0);
- The diameter of the complete bipartite graph $K_{m,n}$ is 2 if either m or n is $\geqslant 2$, while $\mathrm{diam}(K_{1,1}) = 1$;
- The diameter of the cycle C_m is given by

$$\mathrm{diam}(C_m) = \begin{cases} \frac{m}{2} & \text{if } m \text{ is even} \\ \frac{m-1}{2} & \text{if } m \text{ is odd.} \end{cases}$$

Checking rigorously these values is left to the reader as an exercise. For the graphs G_n of Example 2.1.6, (4), computing the diameter is not so easy. In Exercise 2.3.5, the reader will be invited to prove that $\mathrm{diam}(G_n) \asymp n^2$. Since $|G_n| = n!$, this means that

$$\mathrm{diam}(G_n) \asymp (\log |G_n|)^2,$$

hence the diameter is here rather small compared with the number of vertices.

Exercise 2.2.7. We consider here some specific features of trees, which we recall are connected forests.

(1) Show that the diameter of a finite tree $T_{d,k}$ with $d \geqslant 2$ and $k \geqslant 0$ is equal to $2k$, and is achieved by the distance between any two distinct vertices labeled with words $(s_1, \ldots, s_k)$ and $(s'_1, \ldots, s'_k)$ of (maximal) length k with $s_1 \neq s'_1$.

(2) Show that if T is a tree, then for any two vertices x and y, there exists a unique geodesic on T with extremities x and y (the image of all paths of length $d_T(x, y)$ between two vertices x and y of T is the same).

(3) If $T = T_{d,k}$ with "root" vertex $x_0 = \emptyset$ and $0 \leqslant j \leqslant k$, show that

$$V' = \{x \in V_T \mid d_T(x_0, x) \leqslant j\}$$

induces a full subgraph isomorphic to $T_{d,j}$.

(4) If $T = T_{d,k}$ with root x_0 and $x \in T$ is any vertex, show that

$$V'' = \{y \in V_T \mid d_\Gamma(y, x_0) \geqslant d_\Gamma(y, x)\}$$

induces a full subgraph T'' of T which is also a tree.

(5) Let Γ be any graph with girth $\ell \geqslant 1$, and let $x_0 \in V$. Show that the subgraph of Γ induced by

$$V' = \Big\{x \in V \mid d_\Gamma(x_0, x) < \frac{\ell}{2}\Big\}$$

is a tree.

The following very simple fact gives a hint of special features of graphs, when considered as geometric objects:

Proposition 2.2.8. *Let Γ_1 and Γ_2 be graphs, and let $f : \Gamma_1 \longrightarrow \Gamma_2$ be a graph map. Then f is always distance-decreasing, i.e., we have*

$$d_{\Gamma_2}(f(x), f(y)) \leqslant d_{\Gamma_1}(x, y) \tag{2.3}$$

for any x, $y \in \Gamma_1$. In particular, if f is surjective on vertices, the diameter of Γ_2 is at most that of Γ_1, and if f is an isomorphism, it is isometric.

Proof. This inequality follows from the observation that any path

$$\gamma : P_k \longrightarrow \Gamma_1$$

in Γ_1, between x, $y \in \Gamma_1$, gives a corresponding one $f \circ \gamma$ in Γ_2, of the same length, between $f(x)$ and $f(y)$. The distance in Γ_2 between $f(x)$ and $f(y)$ is computed using a minimum over a set which contains these particular paths, and that implies that $d_{\Gamma_2}(f(x), f(y)) \leqslant d_{\Gamma_1}(x, y)$.

The remainder is easy: if f is surjective, for any two vertices z, $w \in V_{\Gamma_2}$, we can write $z = f(x)$, $w = f(y)$ for some x, $y \in V_{\Gamma_1}$ and

$$d_{\Gamma_2}(z, w) = d_{\Gamma_2}(f(x), f(y)) \leqslant d_{\Gamma_1}(x, y) \leqslant \mathrm{diam}(\Gamma_1),$$

which gives $\mathrm{diam}(\Gamma_2) \leqslant \mathrm{diam}(\Gamma_1)$. □

Exercise 2.2.9. Here is an application of graphs and connected components to group theory, due to Bauer and Knutson (see [42, Lemma, p. 98]). Let $k \geqslant 2$ be an integer, $G = \mathfrak{S}_k$ the symmetric group on k letters. We suppose given a subgroup H of G such that: (i) H acts transitively on $\{1, \ldots, k\}$; (ii) H contains at least one transposition; (iii) H contains a cycle of length $p > k/2$ such that p is *prime*. The goal is to prove that, in fact, we have $H = G = \mathfrak{S}_k$.

Let $\Gamma = (V, E)$ be the simple graph with $V = \{1, \ldots, k\}$ and with an edge between any pair $(i, j) \in V \times V$ such that $i \neq j$ and the transposition $(i\ j)$ is in H. Assumption (ii) means that the edge set is not empty.

(1) Show that any connected component in Γ is a complete graph.

(2) Show that it is enough to show that Γ is connected in order to prove that $H = G$.

(3) Show that the action of G on $\{1, \ldots, k\}$ induces an action of G on Γ by automorphisms. Show then that G acts transitively on the set of all connected components of Γ. Deduce that all such components are isomorphic.

(4) Show that a p-cycle $\sigma \in H$ as in (iii) must fix (globally, not necessarily pointwise) each component of Γ, and conclude from this.

Exercise 2.2.10. Let Γ be a graph, and let Γ^s be the associated simple graph (Example 2.1.13). Show that $d_{\Gamma^s}(x, y) = d_\Gamma(x, y)$ for all x, $y \in V$.

Exercise 2.2.11 (Uniqueness of bipartite decompositions). (1) Let Γ be a connected *bipartite* graph with a bipartite decomposition $V = V_0 \cup V_1$. If $x_0 \in V_0$, show that

$$V_0 = \{y \in V \mid \text{there is a path of even length joining } x \text{ to } y\}. \tag{2.4}$$

(2) Deduce that the partition of edges $V = V_0 \cup V_1$ which exhibits the bipartiteness of a connected bipartite graph is unique (i.e., if $W_0 \cup W_1$ is another such partition, we have $(W_0, W_1) = (V_0, V_1)$ or $(W_0, W_1) = (V_1, V_0)$).

(3) Let Γ be an arbitrary connected graph, and define W by the right-hand side in (2.4). Compute W when Γ is *not* bipartite.

(4) Show that a forest is always bipartite.

(5) Show that if a graph Γ is finite and *not* bipartite, then its girth is finite. In fact, show that $\mathrm{girth}(\Gamma) \leqslant 2\,\mathrm{diam}(\Gamma) + 1$, and that this is best possible.

Another important construction related to paths in graphs is the "universal cover" of a graph. We only describe this for a non-empty connected graph Γ. Fix a vertex $v_0 \in V$. Then we define a simple graph $\widehat{\Gamma}_{v_0} = (\widehat{V}, \widehat{E})$ as follows:

- The vertex set $\widehat{V}$ is the set of all non-backtracking paths of length $k \geqslant 0$ in Γ starting at v_0, i.e., all graph maps $\gamma\colon P_k \to \Gamma$ where $k \geqslant 0$ is an integer such that $\gamma(0) = v_0$, and such that γ is non-backtracking (see Definition 2.2.2).
- The edge set $\widehat{E} \subset \widehat{V}^{(2)}$ is the set of pairs $\{\gamma_1, \gamma_2\}$ where γ_1 has length $k \geqslant 0$ and γ_2 has length $k + 1$, and the restriction of γ_2 to $\{0, \ldots, k\}$ is equal to γ_1. (Note in particular that $\gamma_1 \neq \gamma_2$).

There is a natural graph map $\pi\colon \widehat{\Gamma}_{v_0} \to \Gamma$, such that for $\gamma \in \widehat{V}$ of length k, we have $\pi(\gamma) = \gamma(k)$ (the end-point of the path) and for an edge $\alpha = \{\gamma_1, \gamma_2\}$ in $\widehat{E}$, with γ_1 of length k and γ_2 of length $k + 1$, then $\pi_*(\alpha) = \gamma_{2,*}(k + 1)$ (recall that the edge $k + 1$ of P_{k+1} joins k to $k + 1$).

To check that this is indeed a graph map, note that by definition of graph maps, the extremities of the edge $\gamma_{2,*}(k + 1)$ are $\{\gamma_2(k), \gamma_2(k + 1)\}$; by definition of $\widehat{E}$, we have $\gamma_2(k) = \gamma_1(k)$ since they are adjacent, so $\mathrm{ep}(\gamma_{2,*}(k + 1)) = \{\gamma_1(k), \gamma_2(k + 1)\} = \{\pi(\gamma_1), \pi(\gamma_2)\}$, which confirms (2.1).

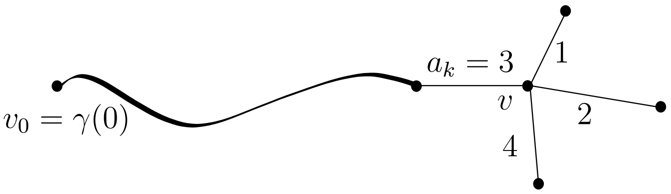


FIGURE 2.2. The universal cover of a regular graph

DEFINITION 2.2.12. The graph $\widehat{\Gamma}_{v_0}$ is called the *universal cover* of the graph Γ based at v_0.

EXERCISE 2.2.13. Let Γ be a non-empty connected graph.

(1) Let v_0 and v_1 be vertices of Γ. Show that the universal covers based at v_0 and v_1 are isomorphic.

(2) Show that $\widehat{\Gamma}_{v_0}$ is connected. Let $\widehat{v}_0$ be the vertex of $\widehat{\Gamma}_{v_0}$ corresponding to the path of length 0 with image v_0. Show then that $\widehat{\pi}\colon \widehat{\widehat{\Gamma}}_{\widehat{v}_0} \to \widehat{\Gamma}_{v_0}$ is an isomorphism.

Here is an important example of computing the universal cover.

PROPOSITION 2.2.14. *Let $d \geqslant 1$ be an integer. Let Γ be a finite non-empty connected d-regular graph. For any $v_0 \in \Gamma$, the universal cover $\widehat{\Gamma}_{v_0}$ of Γ based at v_0 is isomorphic to the infinite d-regular tree T_d* (Example 2.1.20).

REMARK 2.2.15. The isomorphism in this proposition is very far from unique, because trees have large automorphism groups, and this explains why the construction involves many choices (in particular, why it may look more complicated than it maybe should).

PROOF. We denote $\widehat{\Gamma} = \widehat{\Gamma}_{v_0}$. Since both T_d and $\widehat{\Gamma}$ are simple graphs, it is enough to define a graph map $f\colon T_d \to \widehat{\Gamma}$ that is bijective on vertices. We will define for any $k \geqslant 0$, a graph map $f_k\colon T_{d,k} \to \widehat{\Gamma}$ that is injective on vertices and has image the set of non-backtracking paths of length k, with the property that f_{k+1} extends f_k. The desired isomorphism f is then defined by $f(x) = f_k(x)$ for any $k \geqslant 0$ such that $x \in T_{d,k}$.

For $k = 0$, the map f_0 maps the one-vertex graph $T_{d,0}$ to the unique path of length 0 at v_0. Now assume that $k \geqslant 0$ and that f_k has been defined; we will construct f_{k+1} inductively by extending f_k (hence, the extension property will certainly be true). Let x be a vertex in $T_{d,k}$. By definition, $x = a_1 \cdots a_k$ is a word of length k in the alphabet $\{1, \ldots, d\}$, without repeated consecutive letters.

Let $\gamma = f_k(a_1 \cdots a_k)$. By the induction assumption, this is a non-backtracking path of length k in Γ. Let $v = \gamma(k)$ be the "end" of γ. Since Γ is d-regular, we can fix (arbitrarily) a bijection j between the alphabet $\{1, \ldots, d\}$ and the set of edges of Γ of which v is an extremity. We may do so in such a way that the "last" edge $\gamma_*(k)$ of γ corresponds by this bijection j to the last letter a_k of the word x. Then we define f_{k+1} simultaneously for all vertices $a_1 \cdots a_k a_{k+1}$ of $T_{d,k+1}$ with "prefix" equal to x by mapping

$$a_1 \cdots a_k a_{k+1}, \qquad a_{k+1} \neq a_k$$

to the (non-backtracking) path obtained by "concatenating" the path γ and the edge $j(a_{k+1}) \neq \gamma_*(k)$ (see Figure 2.2, illustrating a case with $d = 4$ and $a_k = 3$).

For any vertex $y \in T_{d,k+1}$, there exists a unique $x \in T_{d,k}$ such that x and y are joined by an edge, hence the argument above defines a map from the vertices of $T_{d,k+1}$ to those of $\widehat{\Gamma}$ that extends f_k. By construction, the image of f_{k+1} is contained in the

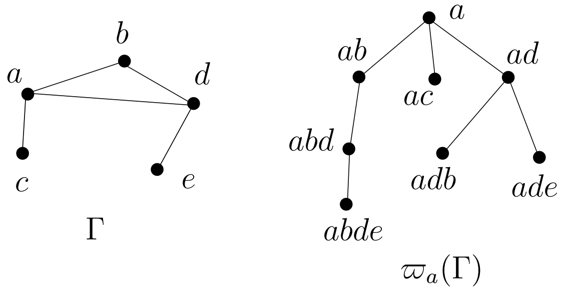


FIGURE 2.3. A path tree

set of non-backtracking paths of length $k+1$. In fact, the image is equal to this set of non-backtracking paths of length $k+1$, since any such path is the concatenation of a non-backtracking path of length k (which belongs by induction to the image of f_k) and of an edge which is different from the "last" edge of the latter. By induction, we have therefore defined a bijection f from the vertices of T_d to those of $\widehat{\Gamma}$. This bijection is a graph map, by definition of the edges in $\widehat{\Gamma}$, since we mapped a word $a_1 \cdots a_{k+1}$ to a non-backtracking path of length $k+1$ extending $f_k(a_1 \cdots a_k)$. Hence we have obtained an isomorphism f. □

EXERCISE 2.2.16. Let Γ be a non-empty finite graph and $v_0 \in V$ a vertex of Γ. The *path graph* (also called *path tree*, when it is connected) of Γ starting at v_0, denoted $\varpi_{v_0}(\Gamma)$, is the simple graph with

- Vertex set given by the set of all *injective* paths in Γ starting at v_0, i.e., all graph maps $\gamma\colon P_k \to \Gamma$ where $k \geqslant 0$ is an integer such that $\gamma(0) = v_0$, and such that γ is injective.
- Edge set $\widehat{E} \subset \widehat{V}^{(2)}$ given by pairs $\{\gamma_1, \gamma_2\}$ where γ_1 has length $k \geqslant 0$ and γ_2 has length $k+1$, and the restriction of γ_2 to $\{0, \ldots, k\}$ is equal to γ_1. (Note in particular that $\gamma_1 \neq \gamma_2$).

Note that injective paths are necessarily non-backtracking, which means that $\varpi_{v_0}(\Gamma)$ is a full subgraph of the universal cover $\widehat{\Gamma}_{v_0}$. Moreover, by the pigeon-hole principle, any injective path has length $\leqslant |\Gamma|$, hence $\varpi_{v_0}(\Gamma)$ is a finite tree. See Figure 2.3 for an illustration.

Assume that Γ is a simple graph.

(1) Let $\widehat{v}_0$ be the vertex of $\varpi_{v_0}(\Gamma)$ corresponding to the path of length 0 at v_0. Show that $\varpi_{v_0}(\Gamma) - \widehat{v}_0$ is the disjoint union, over vertices v in Γ adjacent to v_0, of the path trees $\varpi_v(\Gamma - v_0)$. (Intuitively, once the starting vertex of an injective path from v_0 has been removed, we obtain an injective path starting from a well-defined vertex adjacent to v_0, and conversely).

(2) Let v be a vertex adjacent to v_0, and let v_0v denote the vertex in $\varpi_{v_0}(\Gamma)$ corresponding to the edge from v_0 to v. Show that the graph $\varpi_{v_0}(\Gamma) - v_0 - v_0v$ is the disjoint union of graphs isomorphic to $\varpi_y(\Gamma - v_0 - v)$, where y runs over vertices $y \neq v_0$ adjacent to v, and of graphs isomorphic to $\varpi_y(\Gamma - v_0)$ for all $y \neq v$ adjacent to v_0. (Drawing a picture for a simple example will clarify this).

For the most part, this book will be concerned with graphs that are "sparse" in the sense of having "small" degree. Nevertheless, one particular result that plays an important auxiliary role in Chapter 6 turns out to depend on a property that is really about dense bipartite graph. We discuss this here, the application to the so-called Balog-Gowers-Szemerédi Theorem is found in Appendix A (see Theorem A.3.6).

Proposition 2.2.17. *Let Γ be a finite simple bipartite graph, with edege set $V = V_1 \cup V_2$, all edges joining V_1 and V_2. We assume that*

$$|E| \geqslant \frac{|V_1||V_2|}{\alpha}$$

for some $\alpha \geqslant 1$. Then there exist subsets $U_1 \subset V_1$ and $U_2 \subset V_2$, such that

$$|U_1| \geqslant \frac{|V_1|}{4\sqrt{2}\alpha}, \qquad |U_2| \geqslant \frac{|V_2|}{4\alpha}$$

and such that for any $(v_1, v_2) \in U_1 \times U_2$, there exist at least $2^{-12}\alpha^{-4}|V_1||V_2|$ paths of length three in Γ from v_1 to v_2.

To understand this statement, note that the case $\alpha = 1$ is obvious, since the graph Γ is then a complete bipartite graph, and the result holds with $U_1 = V_1$, $U_2 = V_2$ and $|V_1||V_2|$ paths of length three between two points (v_1, v_2) (namely the paths passing through arbitrary intermediate points, one in V_2 and one if V_1). Hence, intuitively, the statement means that in a fairly dense bipartite graph, one can find "large" subsets of both sides of the bipartite decomposition which are "almost" complete, as far as paths of length 3 are concerned. This can be interpreted as a form of stability of certain properties of complete graphs under perturbation; such ideas will come back often in Chapter 6.

Proof. For $v \in V = V_1 \cup V_2$, we denote by N_v the set of neighbors of v; for $(v, w) \in V$, we denote by $m_{v,w}$ the number of paths of length 2 from v to w, and by $t_{v,w}$ the number of paths of length 3 from v to w. Note that $m_{v,w} = |N_v \cap N_w|$. The proof follows three steps.

Step 1. Fix ε such that $0 < \varepsilon < 1$. We claim that we can find $W_1 \subset V_1$ with $|W_1| \geqslant (\sqrt{2}\alpha)^{-1}|V_1|$ such that the number of pairs $(v, w) \in W_1 \times W_1$ with $m_{v,w} \geqslant \frac{1}{2}\varepsilon\alpha^{-2}|V_2|$ is $\geqslant (1-\varepsilon)|W_1|^2$.

In other words, we want to construct a "big" subset $W_1 \times W_1$ of $V_1 \times V_1$ on which the function $(v, w) \mapsto m_{v,w}$ is mostly "not too small". It is natural to begin by estimating from below the average of $m_{v,w}$ on $V_1 \times V_1$, but the fact that we insist on a "square" subset where the function is large means that the approach is not as simple as averaging and using a generic argument.

Nevertheless, we begin by showing that the function is large on average. By the assumption on the number of edges and the Cauchy-Schwarz inequality, we get

$$\frac{1}{\alpha^2} \leqslant \frac{1}{|V_1|^2|V_2|^2}\Bigl(\sum_{\substack{(v_1,v_2)\in V_1\times V_2\\ v\sim w}} 1\Bigr)^2 = \frac{1}{|V_1|^2|V_2|^2}\Bigl(\sum_{v_2\in V_2}\sum_{\substack{v_1\in V_1\\ v_1\sim v_2}} 1\Bigr)^2$$
$$\leqslant \frac{1}{|V_1|^2|V_2|^2}|V_2|\sum_{v_2\in V_2}\Bigl(\sum_{\substack{v_1\in V_1\\ v_1\sim v_2}} 1\Bigr)^2 = \frac{1}{|V_1|^2|V_2|}\sum_{(v,w)\in V_1^2}\sum_{\substack{v_2\in V_2\\ v\sim v_2,\ w\sim v_2}} 1$$

or in other words

$$\frac{1}{|V_1|^2}\sum_{(v,w)\in V_1^2}\frac{m_{v,w}}{|V_2|} \geqslant \frac{1}{\alpha^2}.$$

Next, let $B \subset V_1 \times V_1$ be the set of pairs (v, w) such that $m_{v,w} < \frac{1}{2}\varepsilon\alpha^{-2}|V_2|^2$. We then have obviously

$$\frac{1}{|V_1|^2} \sum_{(v,w)\in B} \frac{m_{v,w}}{|V_2|} < \frac{\varepsilon}{2\alpha^2}.$$

Comparing with the average, we deduce that

$$\frac{1}{|V_1|^2} \sum_{(v,w)\in V_1^2} \Bigl(1 - \varepsilon^{-1}\mathbf{1}_B(v,w)\Bigr)\frac{m_{v,w}}{|V_2|} \geqslant \frac{1}{2\alpha^2}.$$

Writing $m_{v,w}$ as the sum of 1 over the elements $u \in V_2$ such that u is a neighbor of v and w, we deduce

$$\frac{1}{|V_2|} \sum_{u\in V_2} \frac{1}{|V_1|^2} \sum_{\substack{(v,w)\in V_1^2 \\ u\in N_v\cap N_w}} \Bigl(1 - \varepsilon^{-1}\mathbf{1}_B(v,w)\Bigr) \geqslant \frac{1}{2\alpha^2}.$$

As a consequence, there exists some $u \in V_2$ such that the inner average over (v, w) is $\geqslant \frac{1}{2}\alpha^{-2}$. We define $W_1 = N_u$. We have $W_1 \subset V_1$ since the graph Γ is bipartite, and we claim that W_1 has the desired properties. Indeed, from the defining property

$$\frac{1}{|V_1|^2} \sum_{\substack{(v,w)\in V_1^2 \\ u\in N_v\cap N_w}} \Bigl(1 - \varepsilon^{-1}\mathbf{1}_B(v,w)\Bigr) \geqslant \frac{1}{2\alpha^2}$$

we get first

$$\frac{1}{2\alpha^2} \leqslant \frac{1}{|V_1|^2} \sum_{\substack{(v,w)\in V_1^2 \\ u\in N_v\cap N_w}} 1 = \frac{|N_u|^2}{|V_1|^2},$$

so $|W_1| \geqslant (\sqrt{2}\alpha)^{-1}|V_1|$, and next

$$\frac{1}{\varepsilon}|W_1^2 \cap B| \leqslant \sum_{\substack{(v,w)\in V_1^2 \\ u\in N_v\cap N_w}} \varepsilon^{-1}\mathbf{1}_B(v,w) \leqslant \sum_{\substack{(v,w)\in V_1^2 \\ u\in N_v\cap N_w}} 1 = |N_u|^2 = |W_1|^2,$$

which by definition of B is the last required condition for W_1.

Step 2. We first assume that the minimal degree of a vertex in V_1 is at least $(2\alpha)^{-1}|V_2|$. By Step 1 with $\varepsilon = (16\alpha)^{-1}$, there exists $W_1 \subset V_1$, of size $\geqslant (\sqrt{2}\alpha)^{-1}|V_1|$, such that at least $(1 - (16\alpha)^{-1})|W_1|^2$ pairs $(v, w) \in W_1^2$ are connected by $\geqslant \varepsilon(2\alpha)^{-2} = (32\alpha^3)^{-1}|V_2|$ edges.

In particular, if B denotes the set of pairs $(v, w) \in W_1^2$ connected by at most $(8\alpha)^{-2}|V_2|$ edges, then B is contained in the "exceptional" set, and has order at most $(16\alpha)^{-1}|W_1|^2$.

We define $U_1 \subset W_1$ to be the set of $v \in W_1$ such that

$$|\{w \in W_1 \mid (v, w) \in B\}| \leqslant \frac{1}{8\alpha}|W_1|.$$

We have then

$$\sum_{w\in W_1 - U_1} \frac{|W_1|}{8\alpha} \leqslant |B| \leqslant \frac{1}{16\alpha}|W_1|^2,$$

which means that $|W_1 - U_1| \leqslant \frac{1}{2}|W_1|$, and consequently $|U_1| \geqslant (2\sqrt{2}\alpha)^{-1}|V_1|$.

Next, let $U_2 \subset V_2$ be the set of all $u \in V_2$ such that $|N_u \cap W_1| \geqslant (4\alpha)^{-1}|W_1|$. We claim that the sets U_1 and U_2 have the properties required to prove the lemma. The lower bound for $|U_1|$ has already been shown. To estimate $|U_2|$, we observe that since the minimal degree of a vertex in V_1 is at least $(2\alpha)^{-1}|V_2$, by assumption, we have

$$\sum_{w\in W_1}\sum_{\substack{u\in V_2\\ v\sim u}} 1 \geqslant \frac{|W_1||V_2|}{2\alpha},$$

and therefore

$$\begin{aligned}|W_1||U_2| &\geqslant \sum_{u\in U_2}\sum_{\substack{v\in W_1\\ v\sim u}} 1\\ &= \sum_{v\in W_1}\sum_{\substack{u\in U_2\\ v\sim u}} 1 - \sum_{u\notin V_2}\sum_{\substack{v\in W_1\\ v\sim u}} 1\\ &\geqslant \frac{|W_1||V_2|}{2\alpha} - \frac{|V_2||W_1|}{4\alpha} = \frac{|V_2||W_1|}{4\alpha},\end{aligned}$$

which gives $|U_2| \geqslant (4\alpha)^{-1}|V_2|$.

It remains to estimate the number $t_{v,u}$ of paths of length 3 between v and u, for $v \in U_1$ and $u \in U_2$. By definition of U_2 and U_1, the vertex u has at least $(4\alpha)^{-1}|W_1|$ neighbors w in W_1, and at most $(8\alpha)^{-1}|W_1|$ of these are such that $(v, w) \in B$. Hence u has at least $(8\alpha)^{-1}|W_1|$ neighbors w such that $(v, w) \notin B$. For each such w, there are at least $(32\alpha^3)^{-1}|V_2|$ paths of length 2 from v to w, say passing through $u_1 \in V_2$. Then the paths of length 3

$$v \sim u_1 \sim w \sim u$$

are all distinct, and they show that

$$t_{v,u} \geqslant \frac{1}{8\alpha}\frac{1}{32\alpha^3}|W_1||V_2| \geqslant \frac{1}{2^{3+5+1/2}\alpha^4}|V_1||V_2|.$$

This is, in this case, stronger than the claim of the proposition.

Step 3. Now consider the general case. Let $\tilde{V_1}$ be the set of vertices $v \in V_1$ with degree $\geqslant (2\alpha)^{-1}|V_2|$. We then apply Step 2 to the full (bipartite) subgraph $\tilde{\Gamma}$ with vertex set $\tilde{V_1} \cup V_2$. We write $\beta = |V_1|/|\tilde{V_1}| \geqslant 1$. Then

$$|\tilde{E}| \geqslant |E| - \frac{|V_1||V_2|}{2\alpha} \geqslant \frac{|V_1||V_2|}{2\alpha} = \beta\frac{|\tilde{V_1}||V_2|}{2\alpha} = \frac{|\tilde{V_1}||V_2|}{\gamma},$$

where $\gamma = 2\alpha\beta^{-1}$. Step 2 gives subsets $U_1 \subset \tilde{V_1} \subset V_1$ and $U_2 \subset V_2$ with

$$\begin{aligned}|U_1| &\geqslant \frac{|\tilde{V_1}|}{2\sqrt{2}\gamma} = \frac{|V_1|}{4\sqrt{2}\alpha},\\ |U_2| &\geqslant \frac{|V_2|}{4\gamma} \geqslant \frac{|V_2|}{8\alpha}\end{aligned}$$

and the property that $v \in U_1$ and $u \in U_2$ are joined by at least $(2^8\sqrt{2}\gamma^4)^{-1}|\tilde{V_1}||V_2|$ paths of length 3. Since

$$\frac{|\tilde{V_1}||V_2|}{2^{8+1/2}\gamma^4} = \beta^3\frac{|V_1||V_2|}{2^{12+1/2}\alpha^4} \geqslant \frac{|V_1||V_2|}{2^{13}\alpha^4},$$

the sets U_1 and U_2 satisfy the conditions required. □

2.3. Cayley graphs, action graphs, Schreier graphs

We will now define the *Cayley graphs*, which are used to get a geometric vision of groups and their properties. These will be among the most important examples of graphs in later chapters when considering expansion.

Definition 2.3.1 (Cayley graph). Let G be a group and let $S \subset G$ be any subset which is *symmetric*, in the sense that $s \in S$ if and only if $s^{-1} \in S$. The *Cayley graph* of G with respect to S is the graph (V, E, ep) where the set of vertices is $V = G$, the edges are given by

$$E = \{\{g, gs\} \mid g \in G,\ s \in S\} \subset V^{(2)}$$

and ep is the inclusion map $E \to V^{(2)}$. This graph is denoted $\mathcal{C}(G, S)$.

In other words, to draw $\mathcal{C}(G, S)$, we use the elements of the group as vertices, and draw an edge between x and y if and only if $x^{-1}y \in S$; since S is symmetric, this is equivalent with $y^{-1}x \in S$. This graph is not always a simple graph: although it has no multiple edges, it may have loops. In fact, this happens if and only if $1 \in S$, in which case there is a loop at every vertex.

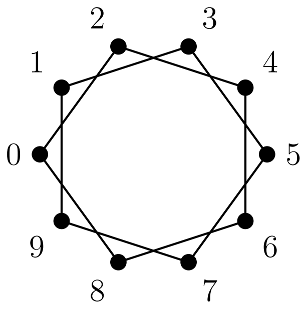
If S is finite, then $\mathcal{C}(G, S)$ is $|S|$-regular (there are $|S - \{1\}|$ edges from $g \in G$ with distinct extremities, because S is symmetric so $\{g, gs\} = \{gs, (gs)s^{-1}\}$, and one possible loop if $1 \in S$).

For a lively and insightful discussion of some of the many aspects of Cayley graphs that we will not discuss in this book, we refer to the book [50] of de la Harpe.

Example 2.3.2. (1) For $m \geqslant 3$, the cycle C_m can be seen as (i.e., it is isomorphic to) the Cayley graph $\mathcal{C}(\mathbf{Z}/m\mathbf{Z}, \{\pm 1\})$, as the reader is invited to check. For $m = 2$, this is not the case (because $1 = -1$ in $\mathbf{Z}/2\mathbf{Z}$); indeed, $\mathcal{C}(\mathbf{Z}/2\mathbf{Z}, \{1\})$ is isomorphic to P_2.

Similarly, for all $m \geqslant 2$, the complete graph K_m is also isomorphic to a Cayley graph of $\mathbf{Z}/m\mathbf{Z}$, but with respect to $S = \mathbf{Z}/m\mathbf{Z} - \{0\}$. This already shows that Cayley graphs can look quite different for the same group G when we change the set S.

(2) Here is a picture of the Cayley graph $\mathcal{C}(\mathbf{Z}/10\mathbf{Z}, \{\pm 2\})$:



Note that this graph is not connected.

(3) If $G = \mathbf{Z}$ and $S = \{\pm 1\}$, we obtain an infinite path (extending indefinitely in both directions).

(4) The graph G_n defined in Example 2.1.6, (4), is isomorphic to the Cayley graph of the symmetric group $\mathfrak{S}_n$ with respect to the (symmetric) subset

$$S_n = \{\tau, (1\ 2\ \cdots\ n)^{\pm 1}\}. \tag{2.5}$$

Indeed, if we use the deck of cards $D_n = \{1, \ldots, n\}$, the isomorphism (say f) maps $\sigma \in \mathfrak{S}_n$ to the arrangement $(\sigma(1), \ldots, \sigma(n))$ of the deck (read left-to-right as being top-to-bottom), which respects the edges: from

$$f(\sigma\tau) = (\sigma(2), \sigma(1), \sigma(3), \ldots, \sigma(n))$$

we see that the edge $\{\sigma, \sigma\tau\}$ corresponds to switching the first two cards, while

$$f(\sigma\sigma_n) = (\sigma(2), \sigma(3), \ldots, \sigma(n), \sigma(1))$$

and

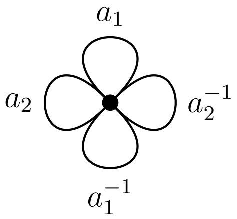
$$f(\sigma\sigma_n^{-1}) = (\sigma(n), \sigma(1), \ldots, \sigma(n-1))$$

do correspond to putting the top card at the bottom, and conversely. We will simply refer to the graphs G_n as Cayley graphs from now on.

The reader should check visually that the graph G_4 is connected and bipartite. As we will soon see, these facts reflect some basic group-theoretic properties of $\mathfrak{S}_n$ and of S_n.

(5) Let $n \geqslant 2$ be an integer and let G be a free group on n generators $(a_1, \ldots, a_n)$ (see for instance [50, Ch. II] for an introduction to free groups, and Appendix B.1 for some basic facts). The Cayley graph of G with respect to the symmetric set $S = \{a_1, a_1^{-1}, \ldots, a_n, a_n^{-1}\}$ is isomorphic to the infinite $(2n)$-regular tree T_{2n} (Example 2.1.20).

To see this, we will use Proposition 2.2.14, but we invite the reader to devise a different argument. This proposition shows that it suffices to prove that $\mathcal{C}(G, S)$ is isomorphic to the universal cover of *some* finite $(2n)$-regular graph Γ. We define $\Gamma = (\{1\}, S, \mathrm{ep})$, where 1 is the neutral element in G and ep is the unique (!) map from S to $\{1\}$. So Γ is a single vertex with $2n$ loops attached, and is $(2n)$-regular (as illustrated below with $n = 2$; we remark in passing that this is one example where the usefulness of allowing multiple edges and loops is particularly striking).



The idea is to represent each power s^m of a generator $s \in S$, with $m \geqslant 1$, by the non-backtracking path of length m obtained by going "around" the loop s, then around s^{-1}, then around s, etc. For a products $s_1^m s_2^n$ of two powers (or a longer product), we simply perform this "dance" for s_1 first, and then for s_2. Since there is only one vertex, there is no problem with starting again at the right place... The reader should convince herself that this leads to a bijection between reduced words in the generators $\{a_1, \ldots, a_n\}$ and non-backtracking paths in Γ. We now present the details.

Any path γ of length k in Γ is a concatenation of edges of Γ (viewed as paths of length 1), i.e., it can be viewed as a k-tuple of elements of S. Such a path is non-backtracking if, and only if, consecutive edges are different. If we gather together pairs of successive edges of the form (s, s^{-1}) for some $s \in S$, we obtain a representation of γ as a concatenation of subpaths represented by

$$(s, s^{-1}, s, \ldots, s^{(-1)^{j-1}})$$

for some integer $j \geqslant 1$ (the unique path of length 0 is represented by an "empty" concatenation). Moreover, we may ensure that two consecutive subpaths of this shape, say

$$(s_1, s_1^{-1}, s_1, \ldots, s_1^{(-1)^{j_1-1}}), \qquad (s_2, s_1^{-1}, s_2, \ldots, s_2^{(-1)^{j_2-1}})$$

satisfy $s_1 \neq s_2$ and $s_1 \neq s_2^{-1}$ (these are excluded, either by the non-backtracking condition, e.g., if j_1 is odd and $s_1 = s_2$, or because they allow us to merge the two subpaths into a longer one, e.g. if j_1 is odd and $s_1 = s_2^{-1}$); once this is done, the representation is in fact unique (as the reader should check). We then map γ to the product

$$f(\gamma) = s_1^{j_1} s_2^{j_2} \cdots$$

in the free group G, viewed as a vertex of $\mathcal{C}(G, S)$.

For example, for $n = 3$, the non-backtracking path corresponding to the sequence

$$(a_1^{-1}, a_2, a_3, a_1^{-1}, a_1, a_3^{-1}, a_2, a_2^{-1}, a_2, a_3^{-1})$$

is mapped to the element $a_1^{-1} a_2 a_3 a_1^{-2} a_3^{-1} a_2^3 a_3^{-1}$ of G.

The construction shows that $f \colon \widehat{\Gamma}_1 \to \mathcal{C}(G, S)$ is a graph map (both graphs are simple graphs). It is bijective, for instance because it is elementary to define an inverse map from G to $\widehat{\Gamma}_1$: express $g \in G$ as a reduced word in the generators $\{a_1, \ldots, a_n\}$, say

$$g = s_1^{j_1} \cdots s_m^{j_m}$$

where $s_{i+1} \notin \{s_i, s_i^{-1}\}$ for all i, and consider the non-backtracking path obtained by concatenating the edges

$$(s_1, s_1^{-1}, \ldots, s_1^{(-1)^{j_1 - 1}}), \cdots, (s_m, s_m^{-1}, \ldots, s_m^{(-1)^{j_m - 1}})$$

(for instance, for $n = 3$, the element $g = a_2^3 a_1^{-2} a_2^{-1} a_3 a_2$ is represented by the path $(a_2, a_2^{-1}, a_2, a_1^{-1}, a_1, a_2^{-1}, a_3, a_2)$). Hence f is the desired isomorphism.

Remark 2.3.3. Using Serre's definition of graphs, the following is the most natural definition of Cayley graphs. Let G be a group and let $\widetilde{S}$ be *any* subset of G (not necessarily symmetric). Recall (Remark 2.1.2 (5)) that a graph in Serre's sense is a tuple (V, E, o, e, i). Define the vertex set $V = G$, the edge set $E = G \times \widetilde{S} \times \{-1, 1\}$. The origin and end maps are given by

$$o(g, s, 1) = g, \quad o(g, s, -1) = gs, \quad e(g, s, 1) = gs, \quad e(g, s, -1) = g$$

and the involution is $i(g, s, \varepsilon) = (g, s, -\varepsilon)$ (so the edge $(g, s, 1)$ goes from g to gs, and the opposite oriented edge $(g, s, -1) = i(g, s, 1)$ goes from gs go g).

Recall also that the graph Γ (in our sense) associated to such a tuple (V, E, o, e, i) is $\Gamma = (V, E/i, \mathrm{ep})$ where $\mathrm{ep}(\alpha) = \{o(\alpha), e(\alpha)\}$ for any edge α. If $\widetilde{S}$ and $\widetilde{S}^{-1}$ do not intersect, then Γ is naturally isomorphic to the Cayley graph $\mathcal{C}(G, S)$ where S is the symmetric set $\widetilde{S} \cup \widetilde{S}^{-1}$. Indeed, the condition $\widetilde{S} \cap \widetilde{S}^{-1} = \emptyset$ shows that E/i can be identified with the edge set of $\mathcal{C}(G, S)$ by the map $(g, s, \pm 1) \mapsto \{g, gs\}$.

If $\widetilde{S} \cap \widetilde{S}^{-1}$ is not empty, we can think of removing from $\widetilde{S}$ those elements that come with their inverse, since these only amount apparently to duplicating some edges. However, this cannot be done if some element of $\widetilde{S}$ is its *own* inverse. Indeed, suppose that some element s of $\widetilde{S}$ is a non-trivial involution, so $s = s^{-1}$. Let $S = \widetilde{S} \cup \widetilde{S}^{-1}$. Then in $\mathcal{C}(G, S)$, we have a *single* edge $\{g, gs\}$ joining g and gs, whereas in the graph $(V, E/i, \mathrm{ep})$, on the other hand, we have *two* edges joining them, namely (the classes of) $(g, s, \pm 1)$ and $(gs, s, \pm 1)$, since $(gs) \cdot s = gs^2 = g$.

The simplest example of this behavior is $G = \mathbf{Z}/2\mathbf{Z}$ and $\widetilde{S} = \{1\}$. Then $S = \{1\}$, and $\mathcal{C}(G, S)$ is P_2 (as already observed), while the "Serre version" Γ of the Cayley graph is (isomorphic to) the cycle C_2 (a 2-regular graph). Similarly, for $G = \mathfrak{S}_n$ and the generating set S_n of Example (4) above, Serre's definition does not give the 3-regular graph of Example 2.1.6 (4), but 4-regular graphs where the single edges corresponding to the generator τ are doubled.

For our purposes in studying expander graphs, there is no practical effect of this small difference. Since Definition 2.3.1 has also some slight advantages, we will continue using it, and we will only make passing references to Serre's definition.

The geometric notions of the previous section are particularly interesting when applied to Cayley graphs. In particular, we have a group-theoretic interpretation of connectedness and of the distance in Cayley graphs:

Proposition 2.3.4 (Metric properties of Cayley graphs). *Let G be a group and S a symmetric subset of G. Let $\Gamma = \mathcal{C}(G,S)$ be the corresponding Cayley graph.*

(1) *The Cayley graph Γ is connected if and only if S is a generating set of G.*

(2) *Denote $\|x\|_S = d_\Gamma(1,x)$. Then the distance d_Γ satisfies*

$$d_\Gamma(x,y) = \|x^{-1}y\|_S, \tag{2.6}$$

for all x, $y \in G = V_\Gamma$, and in particular it is left-invariant, i.e.

$$d_\Gamma(xy,xz) = d_\Gamma(y,z)$$

for all x, y, z in G. Moreover

$$\|x\|_S = \min\{k \geqslant 0 \mid x = s_1\cdots s_k \textit{ for some } s_i \in S\}, \tag{2.7}$$

which is called the word length *of x with respect to S.*

Proof. Statement (1) should be intuitively clear, since paths in $\mathcal{C}(G,S)$ join two elements which differ by multiplication by an element in S, but let us give a proof. First, we assume that $\Gamma = \mathcal{C}(G,S)$ is connected. For any $x \in G$, let $\gamma : P_k \longrightarrow \Gamma$ be a path between 1 and x (of some arbitrary length). If x_i is the element $\gamma(i) \in G$ for $0 \leqslant i \leqslant k-1$, we have $x_0 = 1$ and $x_k = x$, and by definition of the edges in Γ, there exists $s_i \in S$ such that

$$\mathrm{ep}(f_*(i)) = \{x_{i-1}, x_{i-1}s_i\} = \{x_{i-1}, x_i\}$$

for $1 \leqslant i \leqslant k$, i.e., $x_i = x_{i-1}s_i$. By induction this gives

$$x = x_k = x_{k-1}s_k = \cdots = s_1s_1\cdots s_k$$

so that x is in the subgroup of G generated by S, and since it was arbitrary, this subgroup must indeed be equal to G.

The converse is basically already proved now: if S generates G, and x, $y \in G$ are arbitrary vertices of the Cayley graph, we can find $k \geqslant 0$ and elements $s_i \in S$, $1 \leqslant i \leqslant k$, such that

$$x^{-1}y = s_1\cdots s_k,$$

and then there is a path $\gamma : P_k \longrightarrow \Gamma$ defined by

$$\begin{aligned} f(0) = x, \quad f(i) = xs_1\cdots s_i, &\qquad 1 \leqslant i \leqslant k,\\ f_*(i) = \{xs_0\cdots s_{i-1}, xs_0\cdots s_i\}, &\qquad 1 \leqslant i \leqslant k, \end{aligned}$$

which links $f(0) = x$ to $f(k) = y$.

(2) The formulas (2.6) and (2.7) are implicit in what was done before: given x, $y \in G$, there is for any $k \geqslant 0$ a bijection, which we just constructed, between paths $\gamma : P_k \longrightarrow \Gamma$ between x and y, and k-tuples $(s_1,\ldots,s_k) \in S^k$ such that

$$y = xs_1s_2\cdots s_k.$$

The minimal possible k for given x and y is the distance between x and y, so that (2.7) follows, and since the equation above is equivalent with $x^{-1}y = s_1\cdots s_k$, this means also that

$$d_\Gamma(x,y) = \|x^{-1}y\|_S,$$

proving (2.6). □

Exercise 2.3.5. Prove that the set S_n given by (2.5) generates $\mathfrak{S}_n$, and hence that the graphs G_n of Example 2.3.2, (3), are all connected. In fact, show that there exist constants $c > 0$ and $C > 0$ such that the diameter of G_n satisfies

$$cn^2 \leqslant \operatorname{diam}(G_n) \leqslant Cn^2 \tag{2.8}$$

for all $n \geqslant 3$. [Hint: This is a fairly classic exercise. As described by Diaconis and Saloff-Coste [32, §3, Ex. 1], it can be convenient to think of this in terms of card shuffling.]

Cayley graphs do not only give a geometric "representation" of groups, the construction is compatible with homomorphisms, i.e., with possible "relations" between groups: whenever we have a homomorphism

$$G \xrightarrow{f} H$$

of groups, and a subset $S \subset G$, we get an induced graph map

$$(f, f_*) : \mathcal{C}(G, S) \longrightarrow \mathcal{C}(H, f(S))$$

which is defined by the map f itself on the vertices, and by the definition

$$f_*(\{g, gs\}) = f(\{g, gs\}) = \{f(g), f(g)f(s)\}$$

(*"qui s'impose"*) for any edge $\{g, gs\} \in E_{\mathcal{C}(G,S)}$. Obviously, this association maps the identity to the identity of the Cayley graph, and is compatible with composition (in the language of categories, it is a *functor*) on the category of groups with a subset. We also see that (f, f_*) is an embedding whenever f is injective.

As another example of relation between groups and their Cayley graphs, here is a bipartiteness criterion:

Proposition 2.3.6 (Bipartiteness criterion for Cayley graphs). *Let G be a group, and let S be a symmetric generating set of G. Then $\mathcal{C}(G, S)$ is bipartite if and only if there exists a surjective group homomorphism*

$$\varepsilon : G \longrightarrow \{\pm 1\}$$

such that $\varepsilon(s) = -1$ for all $s \in S$. In particular, if $1 \in S$, the Cayley graph $\mathcal{C}(G, S)$ is not bipartite.

Although this can be related to Proposition 2.1.25, the proof is simple enough to spell out in full.

Proof. First of all, suppose ε exists with the properties indicated. Then we can partition the vertex set $V = G$ as

$$V = \varepsilon^{-1}(1) \cup \varepsilon^{-1}(-1)$$

and both subsets are non-empty since ε is supposed to be surjective. Consider an edge in the Cayley graph: it is of the form $\{g, gs\}$ for some $g \in G$ and $s \in S$, and since

$$\varepsilon(gs) = \varepsilon(g)\varepsilon(s) = -\varepsilon(g),$$

it follows that the extremities are distinct, one being in $\varepsilon^{-1}(1)$ while the other is in $\varepsilon^{-1}(-1)$. Hence we can make $\mathcal{C}(G, S)$ into a bipartite graph using this partition.

Conversely, suppose $\mathcal{C}(G, S)$ is bipartite, with $V = G = V_0 \cup V_1$ a partition in inputs and outputs. We then claim that

$$\varepsilon(g) = \begin{cases} 1 & \text{if } g \in V_0 \\ -1 & \text{if } g \in V_1. \end{cases}$$

is a surjective group homomorphism $G \longrightarrow \{\pm 1\}$ such that $\varepsilon(s) = -1$ for $s \in S$.

Certainly, ε is well-defined and is surjective. Moreover, since any $s \in S$ is an extremity of the edge $\{1, s\}$, where $1 \in V_0$, the bipartiteness implies that $s \in V_1$, and hence $\varepsilon(s) = -1$.

There only remains to check that ε is a homomorphism in order to conclude. For this, we claim that $\varepsilon(g)$ can be computed as $\varepsilon(g) = (-1)^k$ for any k such that there exists elements $s_1, \ldots, s_k \in S$ with

$$g = s_1 \cdots s_k$$

(in other words, $\varepsilon(g) = (-1)^{\ell(\gamma)}$ for any path γ between 1 and g). It is easy enough to see this: we start with $s_1 \in V_1$ (by the above), then the edge $\{s_1, s_1 s_2\}$, with one extremity in V_1, implies that $s_1 s_2 \in V_0$, and then similarly $s_1 s_2 s_3 \in V_1$, and so on (by induction, the element $s_1 \cdots s_k$ is in V_0 if k is even, and in V_1 when k is odd, which amounts to this formula $\varepsilon(g) = (-1)^k$).

We now write $g = s_1 \cdots s_k$, $h = t_1 \cdots t_m$ with s_i, $t_j \in S$ (note that since the Cayley graph is connected, the set S is a set of generators!), and obtain

$$\varepsilon(gh) = \varepsilon(s_1 \cdots s_k t_1 \cdots t_m) = (-1)^{k+m} = \varepsilon(g)\varepsilon(h),$$

as desired. □

Example 2.3.7. (1) Consider $G = \mathfrak{S}_n$, the symmetric group on n letters, and the generating set $S = \{$ transpositions in G $\}$. Then $\mathcal{C}(G, S)$ is bipartite, the corresponding homomorphism being the signature $\varepsilon : \mathfrak{S}_n \longrightarrow \{\pm 1\}$.

(2) For the Cayley graphs $G_n = \mathcal{C}(\mathfrak{S}_n, S_n)$ discussed in Example 2.3.2, (3), note that we have $\varepsilon(\tau) = -1$, $\varepsilon((1\ 2 \cdots\ n)) = (-1)^{n-1}$, so that G_n is bipartite if and only if n is even. (For instance, this occurs for G_4, which we drew earlier.)

(3) The first two examples show that bipartiteness is not purely a condition on the group involved, but also depends on the choice of generators. In particular, in situations where having a bipartite graph is a problem (as happens with the behavior of random walks, as we will see in Section 3.2), one can often efficiently bypass the issue for a Cayley graph $\mathcal{C}(G, S)$ by considering instead $\mathcal{C}(G, S \cup \{1\})$, which is not bipartite. Graphically, adding 1 to S amounts to replacing the graph $\mathcal{C}(G, S)$ with the graph with the same vertices, but with an extra loop added at each vertex.

As we will see, for instance when discussing the relation between expansion and diameter in Section 3.5, Cayley graphs are in some ways better behaved than general graphs (even general regular graphs). Very often, this comes from the fact that Cayley graphs are highly symmetric, as the following simple proposition observes:

Proposition 2.3.8 (Automorphisms of Cayley graphs). *Let G be a group and $S \subset G$ an arbitrary symmetric set. The group G acts faithfully and transitively by graph automorphisms on $\mathcal{C}(G, S)$, the automorphism f_g associated to $g \in G$ being given by*

$$f_g(x) = gx, \qquad f_{g,*}(\{x, xs\}) = \{gx, gxs\}$$

for all vertices $x \in G$ and edges $\{x, xs\}$ of the Cayley graph.

The very simple proof is left to the reader. The following corollary expresses the girth of a Cayley graph in group-theoretic terms. It is one of the very few places in this book where our definition of Cayley graphs imposes a restriction.

Corollary 2.3.9. *Let G be a group and $S \subset G$ a symmetric subset that contains no non-trivial involution. Let $\Gamma = \mathcal{C}(G, S)$ be the corresponding Cayley graph. The girth of Γ is then equal to the length of the shortest non-trivial relation among the elements* *of S, namely* $\operatorname{girth}(\Gamma)$ *is the smallest $m \geqslant 1$ for which there exist $(s_1, \ldots, s_m)$ in S, with $s_i s_{i+1} \neq 1$ for all i, such that $s_1 s_2 \cdots s_m = 1$.*

In particular, if G is finite, the girth of the Cayley graph Γ is finite.

Proof. First of all, by composing with a suitable automorphism of Γ, we can replace an arbitrary embedding of a cycle $C_m \hookrightarrow \Gamma$ with one starting from the identity element. Denote by m the integer defined in the statement of the corollary, and let $k = \operatorname{girth}(\Gamma)$.

(1) If k is finite, and $\gamma : C_k \hookrightarrow \Gamma$ is cycle of length k starting at 1, the edges $\gamma_*(\{i-1, i\})$ are of the form $\{g_i, g_i s_i\}$ for some $g_i \in G$ and $s_i \in S$. Following the cycle, we see that the relation $s_1 \cdots s_k = 1$ holds in G. Since the map $C_k \to \Gamma$ is an embedding, we also obtain $s_i s_{i+1} \neq 1$ for all i, hence $m \leqslant k$.

(2) Conversely, let $s_1 \cdots s_m = 1$ be a relation of minimal length in G, with $s_i \in G$ and $s_i \neq s_{i+1}$. Identifying the vertex set $\mathbf{Z}/m\mathbf{Z}$ of C_m with $\{0, \ldots, m-1\}$, we define

$$\gamma(0) = 1, \quad \gamma(i) = s_1 \cdots s_i \text{ for } 1 \leqslant i < m,$$

and

$$\gamma_*(\{i-1, i\}) = \{s_1 \cdots s_{i-1}, s_1 \cdots s_i\} \in E_\Gamma,$$

for $1 \leqslant i \leqslant m$. This defines a graph map $C_m \to \Gamma$, because the relation $s_1 \cdots s_m = 1$ ensures that the last edge "comes back", as it should, to the origin. We claim that, because we selected a relation of minimal length, γ is an embedding. Indeed, assume that i and $j \neq i$ are such that $0 \leqslant i < j < m$ and $\gamma(i) = \gamma(j)$. Then we find that

$$s_1 \cdots s_i = s_1 \cdots s_i s_{i+1} \cdots s_j$$

and hence

$$s_{i+1} \cdots s_j = 1,$$

which is a relation of length $j - i$. By definition of m, this means that $j - i \geqslant m$, which is a contradiction. □

Remark 2.3.10. The restriction on S is needed because, for instance, the Cayley graph $\mathcal{C}(\mathbf{Z}/2\mathbf{Z}, \{1\})$ is a path of length 1, hence a tree, and therefore has infinite girth, whereas the minimal length of a relation in that case is 2. The reader should check that the restriction on S disappears if one uses the Cayley graphs following Serre's definition, as sketched in Remark 2.3.3.

After defining Cayley graphs as a way to "geometrize" the algebraic structure of a group, it is natural to try to do something similar to a set on which the group acts. To keep a uniform presentation in comparison with Cayley graphs, we consider right actions (left actions can of course be handled similarly). Thus we consider a group G and a set X with a right action

$$(g, x) \mapsto x \cdot g$$

of G on X. Given a subset S of G, we can visualize the action geometrically by using X as a set of vertices, and putting edges between any two points x, $x \cdot s$ for $s \in S$. If we take $X = G$ with G acting by right multiplication, we will recover the Cayley graphs.

A suitable definition turns out to be a bit tricky, due partly to our "coding" of graphs. We wish that the action graph should be regular, and that the edges with one extremity at a given vertex $x \in X$ correspond naturally to the elements of S. The presence of possible fixed points turns out to create complications to achieve exactly this (see Example 2.3.17 below for an enlightening illustration).

We will use the method of Example 2.1.8 (3) to define the graph, i.e., we specify the edges between any pair of vertices. The resulting graph will be called the *action graph*.

For $(x, y) \in X \times X$, let $E^+(x, y) = \{s \in S \mid x \cdot s = y\}$. We define $E(x, x) = E^+(x, x)$ for $x \in X$ and for $x \neq y$, we put

$$E(x, y) = (E^+(x, y) \cup E^+(y, x))/(s \sim s^{-1}),$$

(the equivalence classes for the equivalence relation that identifies s and s^{-1}). We then have $E(x, y) = E(y, x)$ for all x and y.

Lemma 2.3.11. *The projection map $E^+(x, y) \to E(x, y)$ is a bijection for all x and y in X.*

Proof. This is true by definition if $x = y$. Otherwise, note that $s \in E^+(x, y)$ if and only if $s^{-1} \in E^+(y, x)$. □

Note that if $s \in S$ is an involution, we have $s \in E^+(x, y)$ and $s \in E^+(y, x)$ whenever $x \cdot s = y$.

Definition 2.3.12 (Action graph). Let G be a group, $S \subset G$ a symmetric subset and X a set on which G acts on the right. The *action graph* of X with respect to S, denoted $\mathcal{A}(X, S)$, is the graph $\mathcal{A}(X, S) = (X, E, \mathrm{ep})$ where E is the disjoint union of the sets $E(x, y)$ for $(x, y) \in X \times X$, and $\mathrm{ep}(\alpha) = \{x, y\}$ if $\alpha \in E$ belongs to $E(x, y)$.

We do not use $E^+(x, y)$ in this definition, because $E^+(x, y) \neq E^+(y, x)$ in general. However, by definition, the adjacency matrix of $\mathcal{A}(X, S)$ is given by $a(x, y) = |E(x, y)| = |E^+(x, y)|$. The neighbors of $x \in X$ are the elements $y = x \cdot s$ for some $s \in S$, and if an edge α is represented by $s \in E^+(x, y)$, then we have $\mathrm{ep}(\alpha) = \{x, x \cdot s\}$.

Lemma 2.3.13. *Let $x \in X$. The map sending $s \in S$ to the edge corresponding to the element of $E(x, x \cdot s)$ corresponding to $s \in E^+(x, x \cdot s)$ is a bijection from S to the set E_x of edges in $\mathcal{A}(X, S)$ with one extremity equal to x.*

Proof. We have a disjoint union

$$E_x = \bigcup_{y \in X} E(x, y) = E(x, x) \cup \bigcup_{y \neq x} E(x, y) = E^+(x, x) \cup \bigcup_{y \neq x} E(x, y),$$

and Lemma 2.3.11 shows that for any $y \neq x$, the map sending $s \in E^+(x, y)$ to the corresponding edge in $E(x, x \cdot s)$ is a bijection, hence the result follows. □

Remark 2.3.14. We define an action graph for a left action of G on X by an immediate adaptation of the definition: we use $E^+(x, y) = \{s \in S \mid s \cdot x = y\}$ to define edges.

Definition 2.3.15 (Relative Cayley graphs, Schreier graphs). Let G be a group, $S \subset G$ a symmetric subset and $H \subset G$ a subgroup.

(1) The *Schreier* graph of $H\backslash G$ with respect to S is the action graph $\mathcal{A}(H\backslash G, S)$, where G acts on $H\backslash G$ by $Hx \cdot g = H(xg)$ for $g \in G$ and $Hx \in H\backslash G$.

(2) If $H \triangleleft G$ is normal in G with quotient $K = G/H = H\backslash G$ and canonical projection $\pi \colon G \longrightarrow K$, and if G acts on K by multiplication on the right

$$k \cdot g = k\pi(g),$$

then we call the associated Schreier graph the *relative Cayley graph* of K with respect to S. It is also denoted $\mathcal{C}(K, S)$.

Exercise 2.3.16. Let $H \triangleleft G$ be a normal subgroup and $S \subset G$ a symmetric subset. Denote by $\pi \colon G \to G/H$ the canonical surjection. Show that the relative Cayley graph $\mathcal{A}(G/H, S)$ is naturally isomorphic to the Cayley graph $\mathcal{C}(G/H, \pi(S))$ (which justifies our notation $\mathcal{C}(G/H, S)$).

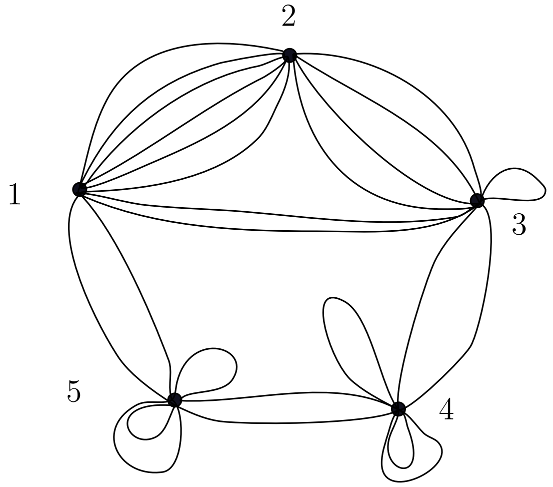
Example 2.3.17. The following example will clarify the complicated-looking definition of action graphs. We take $G = \mathfrak{S}_5$, the symmetric group on 5 letters, with symmetric generating set $S = \{\tau, \mu, \mu^{-1}, \sigma, \sigma^{-1}\}$ where

$$\tau = (1\ 2), \quad \mu = (1\ 2\ 3\ 4\ 5), \quad \sigma = (1\ 2\ 3)$$

(in cycle notation; the first element $\tau = (1\ 2)$ is of order 2, so equal to its inverse).

The group G acts on $X = \{1, \ldots, 5\}$ on the left by evaluation, namely by $\sigma \cdot i = \sigma(i)$, and this action is transitive, hence isomorphic to the multiplication action of G on right cosets of the stabilizer subgroup $H = \{\sigma \in G \mid \sigma(1) = 1\}$, so that the action graph $\mathcal{A}(X, S)$ is isomorphic to the Schreier graph $\mathcal{A}(G/H, S)$.

First, here is a geometric picture of the graph $(X, X \times S, \mathrm{ep})$, where $\mathrm{ep}(x, s) = \{x, s \cdot x\}$, which shows all connections between elements of X arising by the action of S (in other words, the set of edges between x and y is the disjoint union of $E^+(x, y)$ and $E^+(y, x)$):

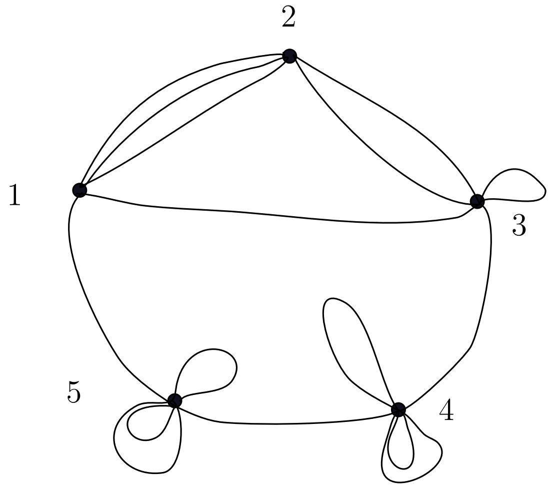


The reader should check that this is correct. For instance, there are indeed $25 = |S| \times |X|$ edges, and the six edges with extremities $\{1, 2\}$ are given by

$$(\tau, 1), \quad (\tau, 2), \quad (\mu, 1), \quad (\mu^{-1}, 2), \quad (\sigma, 1), \quad (\sigma^{-1}, 1),$$

while the three loops around 4 are $(\tau, 5)$, $(\sigma, 5)$ and $(\sigma^{-1}, 5)$. Note in particular that this graph is not regular.

The corresponding action graph $\Gamma = \mathcal{A}(X, S) \simeq \mathcal{A}(G/H, S)$, on the other hand, is the following graph:



There are now only three edges between 1 and 2, corresponding to the quotient in the equivalence relation defining $E(1,2)$, but the loops are preserved. The graph is, indeed, 5-regular.

Note also that the associated simple graph is different from both graphs (it is simply the cycle C_5).

Exercise 2.3.18. Continuing Exercise 2.1.5, give a definition of action graphs using the coding you defined, which corresponds to the definition above.

Proposition 2.3.19. *Let G be a group and $S \subset G$ a symmetric generating set.*

(1) *Let X be a set on which G acts on the right, and $x_0 \in X$. The orbit map*

$$f \begin{cases} G & \longrightarrow & X \\ g & \mapsto & x_0 \cdot g \end{cases}$$

induces a natural graph map $\mathcal{C}(G,S) \longrightarrow \mathcal{A}(X,S)$, i.e., there exists a map f_* between the edge sets of both graphs so that (f, f_*) is a graph map.

(2) *More generally, let $f : X \to Y$ be a morphism of sets with right G-action, i.e., we have $f(x \cdot g) = f(x) \cdot g$ for all $x \in X$. Then f induces in a similar way a graph map*

$$\mathcal{A}(X,S) \longrightarrow \mathcal{A}(Y,S).$$

Proof. (1) We need to define the map f_* between edges so that (f, f_*) is a graph map. Let $\{g, gs\}$ be an edge in the Cayley graph. We map it to the edge represented by $s \in E^+(x_0 \cdot g, x_0 \cdot gs)$ in the action graph; this is well-defined, because if we interpret the original edge as $\{gs, (gs)s^{-1}\}$, then $s^{-1} \in E^+(gs, g)$ corresponds to the same edge in the action graph.

(2) The construction is entirely similar and left to the reader. □

Exercise 2.3.20. Let G act on the left on a set X.

(1) If $S \subset G$ is a symmetric generating set, show that $\mathcal{A}(X,S)$ has no loop if and only if the elements of S act without fixed points on X.

(2) If $X = G/H$ with the action of G by left-multiplication, show that $\mathcal{A}(G/H,S)$ has no loops if and only if $S \cap xHx^{-1} = \emptyset$ for all $x \in G$.

(3) Let $k \geqslant 2$ be an integer and let $G = \mathfrak{S}_k$ acting on $X = \{1, \ldots, k\}$ by evaluation. Find (or show that there exists) a symmetric generating set S of G such that $\mathcal{A}(X,S)$ has no loops.

(4) Find a criterion for the action graph to be connected.

CHAPTER 3

Expansion in graphs

3.1. Expansion in graphs

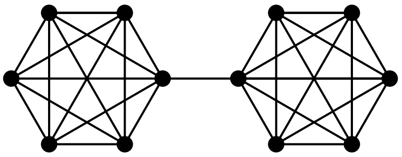
In this section, we begin the study of *expansion properties* of graphs. This will lead to the definition of an expander family of graphs, and the main results of this chapter will be the equivalence of different notions of expanders.

The goal is to find a quantitative invariant that can be used to measure a very high level of connectedness of a graph. Of course, assuming a graph is known to be connected, the diameter is the first natural invariant that comes to mind: for a fixed number of vertices, a graph with smaller diameter is "better connected".

However, as discussed in Chapter 1, we also wish to be able to detect (using our invariant) that the graph is "robust", by which we mean that it can not be disconnected too easily.

For instance, consider a graph Γ_m given by taking the disjoint union of two copies Γ and Γ' of a complete graph K_m, for some $m \geqslant 2$, and adding a single edge between chosen vertices $x_1 \in \Gamma$ and $x_2 \in \Gamma'$:

(3.1)



We clearly have $\operatorname{diam}(\Gamma_m) = 3$, for any m, which shows that Γ_m has very small diameter. But if we remove the single additional edge between x_1 and x_2, we obtain a disconnected graph. This behavior is not desirable in many applications, and leads to the definition of the "expansion constant", or Cheeger constant, of a graph (the name is motivated by the geometric analogue defined by Cheeger in [28], which will also make an appearance in Example 5.4.7 in Section 5.4).

Definition 3.1.1 (Expansion constant). Let $\Gamma = (V, E, \mathrm{ep})$ be a finite graph.

(1) For any disjoint subsets of vertices $V_1, V_2 \subset V$, we denote by $\mathcal{E}(V_1, V_2)$ or $\mathcal{E}_\Gamma(V_1, V_2)$ the set of edges of Γ with one extremity in V_1 and one extremity in V_2,

$$\mathcal{E}(V_1, V_2) = \{\alpha \in E \mid \mathrm{ep}(\alpha) \cap V_1 \neq \emptyset, \quad \mathrm{ep}(\alpha) \cap V_2 \neq \emptyset\}.$$

and we denote by $\mathcal{E}(V_1)$ or $\mathcal{E}_\Gamma(V_1)$ the set $\mathcal{E}(V_1, V - V_1)$ of edges with one extremity in V_1, and one outside V_1.

(2) The *expansion constant* $h(\Gamma)$ is defined by

$$h(\Gamma) = \min\Bigl\{\frac{|\mathcal{E}(W)|}{|W|} \in [0, +\infty[\Bigm| \emptyset \neq W \subset V \text{ and } |W| \leqslant \tfrac{1}{2}|\Gamma|\Bigr\},$$

with the convention that $h(\Gamma) = +\infty$ if Γ has at most one vertex.

In other words, $h(\Gamma)$ is the smallest possible ratio between the number of edges exiting from W and the size of W, when W is a set of vertices that is non-empty, but not too big. This will provide a measure of robustness, in the following sense: the larger $h(\Gamma)$ is, the more difficult it is to disconnect a largish subset of V from the rest of the graph. This is expressed in the following result:

Proposition 3.1.2. *Let $\Gamma = (V, E, \mathrm{ep})$ be a finite graph with at least two vertices, so that $h(\Gamma) < +\infty$.*

(1) We have $h(\Gamma) > 0$ if and only if Γ is connected.

(2) If $W \subset V$ is a subset of vertices with $|W| = \delta|V|$ where $0 < \delta \leqslant \frac{1}{2}$, one must remove at least $\delta h(\Gamma)|V|$ edges from Γ to disconnect W from the rest of the graph.

Proof. (1) The condition $h(\Gamma) = 0$ means that there exists some $W \subset V$, non-empty, of size $\leqslant |\Gamma|/2$, such that $\mathcal{E}(W)$ is empty. In particular $V - W$ is also of size $\geqslant 1$. Let $x \in W$ and $y \notin W$ be two vertices. Then there is no path in Γ between x and y, since such a path would have to cross from W to $V - W$ at some point (we leave as an exercise to make this rigorous). Therefore Γ is not connected.

Conversely, if Γ is not connected, there are at least two connected components in Γ, and at least one of them, say W, must have size $|W| \leqslant |\Gamma|/2$. Since W is not empty and $\mathcal{E}(W) = \emptyset$, we get $h(\Gamma) \leqslant |\mathcal{E}(W)|/|W| = 0$.

(2) Once we explain the meaning of the sentence, it will become clear: we say that removing a set C of edges disconnects W from $V - W$ if $\mathcal{E}(W) \subset C$, i.e., all edges that go from W to "somewhere else" are contained in C. Then since

$$|\mathcal{E}(W)| \geqslant h(\Gamma)|W| = \delta h(\Gamma)|V|,$$

by definition of $h(\Gamma)$, our statement is just a reformulation of the definition. □

Example 3.1.3. (1) Consider the complete graph K_m with $m \geqslant 2$ vertices. Any two subsets of the vertices with the same cardinality are equivalent (i.e., there is an automorphism of the graph mapping one to the other), and hence

$$h(K_m) = \min_{1 \leqslant j \leqslant m/2} \frac{1}{j}|\mathcal{E}(\{1, \ldots, j\})| = \min_{1 \leqslant j \leqslant m/2} (m - j) = m - \left\lfloor \frac{m}{2} \right\rfloor$$

(since there are $j(m-j)$ edges in K_m from $\{1, \ldots, j\}$ to its complement $\{j+1, \ldots, m\}$).

(2) Consider now $\Gamma = C_m$, the cycle with $m \geqslant 2$ vertices. The subsets of size $\leqslant m/2$ that expand least are given by the images W of paths in C_m of length $\mathrm{diam}(C_m) = \lfloor \frac{m}{2} \rfloor \leqslant m/2$ (this is intuitively clear, and the proof is left as an exercise). In this case $\mathcal{E}(W)$ has two elements (one edge from each end of the path), and therefore

$$h(C_m) = \frac{2}{\lfloor \frac{m}{2} \rfloor} \leqslant \frac{4}{m-1}. \tag{3.2}$$

Note that the inequality $h(C_m) \leqslant 4/(m-1)$ follows, even if one does not know that paths are the least expanding subsets, since

$$h(C_m) \leqslant \frac{|\mathcal{E}(W)|}{|W|}$$

by definition for any subset W.

(3) Let Γ be a graph like the one in (3.1): two copies of K_m joined by a single edge α. Then if we take W to be the first copy of K_m, we see that $\mathcal{E}(W) = \{\alpha\}$, hence

$$h(\Gamma) \leqslant \frac{1}{m}.$$

(4) Let $T = T_{d,k}$ be a finite tree with degree $d \geqslant 3$ and depth $k \geqslant 1$. The expansion constant can be bounded from above by taking as subset W one of the subtrees "below a

neighbor of the root", i.e., if x_0 is the root and x_1 is a vertex indexed with a single letter of the alphabet (e.g., $x_1 = 1$), we let

$$W = \bigcup_{2\leqslant j\leqslant k} \{(1, s_2, \ldots, s_j) \in V_T\}$$

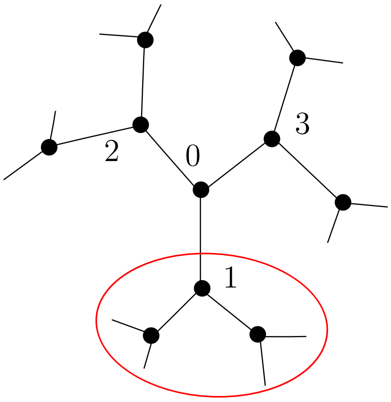
which (see Exercise 2.2.7, (4)), can be written equivalently as

$$W = \{y \in V_T \mid d_T(y, x_0) \geqslant d_T(y, 1)\}.$$

We then have $|W| = \frac{|T|-1}{d} \leqslant \frac{|T|}{2}$, and therefore

$$h(T) \leqslant \frac{|\mathcal{E}(W)|}{|W|}.$$

It is clear from the picture



that $\mathcal{E}(W)$ contains a *single* edge, the one joining 0 to 1 (in other words, to "escape" from the subtree induced by W, one *must* pass through the root), and therefore

$$h(T) \leqslant \frac{1}{|W|} = \frac{d}{|T|-1}.$$

These examples are already instructive. In particular, they show that $h(\Gamma)$ behaves in a way consistent with our goal: the "super"-connected complete graphs have $h(\Gamma)$ very large, while large, easily-disconnected graphs, like C_m or those of Example (3) have quite small expansion constants.

Although the arguments were highly elementary, they also show that it is much easier to give an upper-bound for $h(\Gamma)$ than a lower-bound: since the expansion constant is defined as a minimum, a single well-chosen subset W may lead to a good upper-bound, while we need to know which sets are the worst behaved in order to give a non-trivial lower-bound. This is confirmed by the wide gap in the following trivial bounds:

LEMMA 3.1.4 (Trivial bounds). *For any finite connected graph Γ with at least two vertices,*

$$\frac{2}{|\Gamma|} \leqslant h(\Gamma) \leqslant \min_{x\in V} \operatorname{val}(x),$$

where we recall that $\operatorname{val}(x)$ *is the valency of a vertex x.*

PROOF. For the lower-bound, we just note that since Γ is connected, we must have $|\mathcal{E}(W)| \geqslant 1$ for any non-empty proper subset W in V, hence

$$\frac{|\mathcal{E}(W)|}{|W|} \geqslant \frac{1}{|W|} \geqslant \frac{2}{|\Gamma|}$$

if $1 \leqslant |W| \leqslant |\Gamma|/2$. On the other hand, for the upper-bound, take $W = \{x\}$ where $x \in V$ (which satisfies $|W| \leqslant |\Gamma|/2$), and note that $|\mathcal{E}(W)| = \mathrm{val}(x)$, hence $h(\Gamma) \leqslant \mathrm{val}(x)$ for all $x \in V$. □

We now come to a proper result: we show that a large $h(\Gamma)$ implies that the diameter of a graph is relatively small. This means that the expansion constant does control this more natural-looking invariant.

Proposition 3.1.5 (Expansion and diameter). *Let Γ be a finite non-empty connected graph. We have*

$$\mathrm{diam}(\Gamma) \leqslant 2\frac{\log \frac{|\Gamma|}{2}}{\log\Bigl(1+\frac{h(\Gamma)}{v}\Bigr)} + 3 \tag{3.3}$$

where $v = \max_{x \in V} \mathrm{val}(x)$ is the maximal valency.

The intuitive idea is the following: to "join" x to y with a short path, we look at how many elements there are at increasing distance from x and y; the definition of the expansion constant gives a geometrically increasing *lower-bound* on the number of new elements when we increase the distance by one, and at some point the sets which can be reached in n steps from both sides are so big that they have to intersect, giving a distance at most $2n$ by the triangle inequality.

The following lemma is the crucial step:

Lemma 3.1.6. *Let Γ be a finite non-empty connected graph and $x \in V$. For any $n \geqslant 0$, let $\mathcal{B}_x(n)$ be the ball of radius n around x, i.e.*

$$\mathcal{B}_x(n) = \{y \in V \mid d_\Gamma(x,y) \leqslant n\}.$$

Then, with v denoting the maximal valency of Γ, we have

$$|\mathcal{B}_x(n)| \geqslant \min\Bigl(\frac{|\Gamma|}{2}, \Bigl(1+\frac{h(\Gamma)}{v}\Bigr)^n\Bigr).$$

Proof. It is enough to show that if $n \geqslant 0$ is such that $|\mathcal{B}_x(n)| \leqslant |\Gamma|/2$, then we have

$$|\mathcal{B}_x(n+1)| \geqslant \Bigl(1+\frac{h(\Gamma)}{v}\Bigr)|\mathcal{B}_x(n)|,$$

since $\mathcal{B}_x(0) = \{x\}$. To prove this inequality, we observe simply that if $\alpha \in \mathcal{E}(\mathcal{B}_x(n))$ is an edge exiting from $\mathcal{B}_x(n)$, its extremity which is not in $\mathcal{B}_x(n)$ is in $\mathcal{B}_x(n+1) - \mathcal{B}_x(n)$, i.e., is at distance $n+1$ from x: this is a "new" point.

It is possible that multiple edges α starting from $\mathcal{B}_x(n)$ lead to the same y, but since all these edges share the extremity y, the maximal number of edges leading to y is $\mathrm{val}(y) \leqslant v$, so that

$$|\mathcal{B}_x(n+1) - \mathcal{B}_x(n)| \geqslant \frac{|\mathcal{E}(\mathcal{B}_x(n))|}{v} \geqslant \frac{h(\Gamma)}{v}|\mathcal{B}_x(n))|,$$

by definition of $h(\Gamma)$, using the assumption that $|\mathcal{B}_x(n)| \leqslant |\Gamma|/2$. Then we get

$$|\mathcal{B}_x(n+1)| = |\mathcal{B}_x(n)| + |\mathcal{B}_x(n+1) - \mathcal{B}_x(n)| \geqslant \Bigl(1+\frac{h(\Gamma)}{v}\Bigr)|\mathcal{B}_x(n)|,$$

as desired. □

Proof of Proposition 3.1.5. Let x, $y \in V$ be two arbitrary vertices; we are going to estimate $d_\Gamma(x,y)$ from above. For this, we denote

$$\beta = 1 + \frac{h(\Gamma)}{v},$$

and we denote by $n \geqslant 1$ the smallest integer such that

$$\beta^n \geqslant \frac{|\Gamma|}{2},$$

(which is possible since $\beta > 1$, in view of the connectedness of Γ). Then by Lemma 3.1.6, applied to x and y, we find that

$$|\mathcal{B}_x(n)| \geqslant \frac{|\Gamma|}{2}, \qquad |\mathcal{B}_y(n)| \geqslant \frac{|\Gamma|}{2}.$$

In fact, we must have $|\mathcal{B}_x(n+1)| > |\Gamma|/2$ (because either this is true for $\mathcal{B}_x(n)$, or else $|\mathcal{B}_x(n)| = |\Gamma|/2$ and then there are some vertices at distance $n+1$), and therefore

$$\mathcal{B}_x(n+1) \cap \mathcal{B}_y(n) \neq \emptyset,$$

which means that $d_\Gamma(x,y) \leqslant 2n+1$ by passing through an intermediate point z lying in this intersection...

Since x and y were arbitrary, we have $\operatorname{diam}(\Gamma) \leqslant 2n+1$, and since

$$n = \left\lceil \frac{\log \frac{|\Gamma|}{2}}{\log \beta} \right\rceil \leqslant \frac{\log \frac{|\Gamma|}{2}}{\log \beta} + 1,$$

we obtain the diameter bound that we stated. □

Example 3.1.7. One checks easily that, for the complete graphs and the cycles, this translates to the following asymptotic upper bounds on the diameter:

$$\operatorname{diam}(K_m) \ll \log m, \qquad \operatorname{diam}(C_m) \ll m \log m$$

for $m \geqslant 2$. Both are off by a factor of size $\log m = \log |K_m| = \log |C_m|$ from the actual values.

We can now define expander graphs, which encapsulate the idea of graphs which are both relatively sparse and highly, and robustly, connected.

Definition 3.1.8 (Expander graphs). A *family* $(\Gamma_i)_{i \in I}$ of finite non-empty connected graphs $\Gamma_i = (V_i, E_i, \mathrm{ep})$ is an *expander family*, or a *family of expanders*, if there exist constants $v \geqslant 1$ and $h > 0$, independent of i, such that:

(1) The number of vertices $|V_i|$ "tends to infinity", in the sense that for any $N \geqslant 1$, there are only finitely many $i \in I$ such that Γ_i has at most N vertices.

(2) For each $i \in I$, we have

$$\max_{x \in V_i} \operatorname{val}(x) \leqslant v,$$

i.e., the maximal valency of the graphs is bounded independently of i.

(3) For each $i \in I$, the expansion constant satisfies

$$h(\Gamma_i) \geqslant h > 0,$$

i.e., it is bounded away from 0 by a constant independent of i.

We will say that a pair (h, v) for which the two properties above hold are *expansion parameters* of the family.

Remark 3.1.9. Most often, the index set I is just the set of positive integers, so that we have a *sequence* of expander graphs. But it is sometimes convenient to allow more general index sets.

Let us review these conditions. The first is, to some extent, a matter of convention: if Γ is a fixed non-empty connected graph, it has bounded valency, of course, as well as positive expansion constant, and hence a "constant" family with $\Gamma_i = \Gamma$ for all i would qualify as expanders if the number of vertices was allowed to remain bounded. But since our intuition is that a family of expanders should allow us to construct arbitrarily large graphs (measured with the number of vertices) which are "sparse" and "super-connected", it is not of interest to just repeat a single graph infinitely many times.

The second condition is our interpretation of sparsity. The point is that if the valency of vertices of a graph Γ is $\leqslant k$, the number of edges is controlled by the number of vertices, namely

$$|E_\Gamma| \leqslant k|V_\Gamma|.$$

The number of edges is seen here (as discussed in Chapter 1) as a "cost" involved in constructing the graph. Bounding the valency means that we ensure that the cost scales linearly with the number of vertices.

Finally, the last condition is a connectedness and robustness assertion. It is natural in view of our examples and of Proposition 3.1.5. It is the best to hope for here, since the trivial bound of Lemma 3.1.4 shows that one can not do better than having $h(\Gamma)$ bounded from below for a family of graphs with bounded valency.

In fact, combining the conditions of sparseness and expansion, we can now derive the following result, which shows that expanders have quite a small diameter, relative to the number of vertices:

Corollary 3.1.10 (Diameter of expanders). *Let (Γ_i) be an expander family of graphs. Then we have*

$$\mathrm{diam}(\Gamma_i) \ll \log(3|\Gamma_i|)$$

for all i, where the implied constant depends only on the expansion parameters (h, v) of the family. [1]

Note that the examples of finite trees $T_{d,k}$, with $d \geqslant 3$ fixed, show that the converse to this statement is not true: the sequence $(T_{d,k})_{k\geqslant 1}$ is a sequence of graphs which have valency bounded by d, and diameter $2k \ll \log |T_{d,k}|$, but they are not expanders.

Proof. Let J be the set of the (finitely many) indices $i \in I$ such that $|\Gamma_i| \leqslant \frac{1}{3}e^3$. We apply Proposition 3.1.5: denoting

$$v = \max_{i\in I} \max_{x\in\Gamma_i} \mathrm{val}(x) < +\infty, \qquad h = \inf_{i\in I} h(\Gamma_i) > 0,$$

and

$$\xi = \frac{1}{\log(1 + h/v)} > 0,$$

we get first

$$\begin{aligned}\mathrm{diam}(\Gamma_i) \leqslant 2\xi \log(\tfrac{1}{2}|\Gamma_i|) + 3 &\leqslant 2\xi \log(\tfrac{1}{2}|\Gamma_i|) + \log(3|\Gamma_i|)\\ &\leqslant (2\xi + 1)\log(3|\Gamma_i|),\end{aligned}$$

[1] We use $3|\Gamma_i|$ to avoid any problem with the possible exceptional i's where $|\Gamma_i| = 1$, and because $\log 3 \geqslant 1$; this is old analytic number theory lore...

for $i \notin J$. We can then get an estimate valid for all i, e.g., by writing $\operatorname{diam}(\Gamma_i) \leqslant C \log(3|\Gamma_i|)$ with

$$C = \max(2\xi + 1, \max_{j \in J} \operatorname{diam}(\Gamma_j)) \tag{3.4}$$

for all $i \in I$. □

Although there are infinite sequences of graphs with diameter growing asymptotically slower than the logarithm of the number of vertices, as we have seen with complete graphs, this result is in fact *best possible* under the sparseness condition, as far as the order of magnitude is concerned:

Lemma 3.1.11. *Let Γ be a non-empty finite graph with maximal valency $\leqslant k$, where $k \geqslant 1$ is an integer. Then*

$$\operatorname{diam}(\Gamma) \geqslant \frac{\log(|\Gamma|)}{\log k}.$$

Proof. Let $x \in \Gamma$ be a fixed vertex. By induction, we get immediately $|\mathcal{B}_x(n)| \leqslant k^n$ for $n \geqslant 0$, since for each $y \in \mathcal{B}_x(n-1)$, there are at most k vertices joined to y at distance n from x. If $d = \operatorname{diam}(\Gamma)$, we have $\mathcal{B}_x(d) = \Gamma$ by definition and hence $|\Gamma| \leqslant k^d$, which is the desired estimate. □

Thus we see that, *if they exist*, expander families are essentially optimal graphs when it comes to combining sparsity and strong connectedness (or expansion) properties.

Another useful property of expanders arises immediately from Lemma 3.1.6:

Proposition 3.1.12 (Growth of metric balls). *Let (Γ_i) be an expander family. Then the metric balls in Γ_i are uniformly exponentially expanding, in the sense that there exists $\gamma > 1$, independent of i, such that for any graph Γ_i in the family, we have*

$$|\mathcal{B}_x(n)| \geqslant \min\Big(\frac{|\Gamma|}{2}, \gamma^n\Big),$$

for all $x \in \Gamma_i$ and $n \geqslant 0$. In fact, one can take $\gamma = 1 + h/v$, where (h, v) are expansion parameters of the family.

Proof. This is a direct consequence of Lemma 3.1.6. □

Exercise 3.1.13. Let (Γ_i) be an expander family such that all Γ_i are d-regular for some *fixed* integer $d \geqslant 1$. Show that $d \geqslant 3$. (We will see later that d-regular expander families do exist for all $d \geqslant 3$.)

Exercise 3.1.14 (Some Cayley graphs of $\mathfrak{S}_n$). We consider again the Cayley graphs $G_n = \mathcal{C}(\mathfrak{S}_n, S_n)$ of Example 2.3.2. Could (G_n) be an expander family? For the moment, we only know an upper bound (2.8) for the diameter that is a bit too weak, but is not very far off from the estimate

$$\operatorname{diam}(G_n) \ll \log|G_n| \ll n \log n$$

that would be necessary for an expander. However, we will see here concretely that (G_n) is *not* an expander. (We will also present later, in Proposition 3.5.8, another proof of this using the results of Section 3.4, which in fact produces a better upper bound for $h(G_n)$.)

It is convenient here to see $\mathfrak{S}_n$ as acting by permutations of $\mathbf{Z}/n\mathbf{Z}$. With this interpretation, the generators σ_n and σ_n^{-1} act on $\mathbf{Z}/n\mathbf{Z}$ by

$$\sigma_n(i) = i + 1, \qquad \sigma_n^{-1}(i) = i - 1$$

for $i \in \mathbf{Z}/n\mathbf{Z}$.

Define then

$$W_n = \{\sigma \in \mathfrak{S}_n \mid \text{ there is no } i \in \mathbf{Z}/n\mathbf{Z} \text{ such that } \sigma(i+1) = \sigma(i)+1\} \subset \mathfrak{S}_n.$$

(1) Show that

$$\frac{|\mathcal{E}(W_n)|}{|\mathfrak{S}_n|} \ll \frac{1}{n}.$$

(2) Show that

$$\frac{1}{3} \leqslant \liminf_{n\to+\infty} \frac{|W_n|}{|\mathfrak{S}_n|} \leqslant \limsup_{n\to+\infty} \frac{|W_n|}{|\mathfrak{S}_n|} \leqslant \frac{1}{2},$$

and conclude that $h(G_n)) \ll n^{-1}$. [Hint: You can use inclusion-exclusion.]

The reader may have wondered why the expansion constant $h(\Gamma)$ was defined using the quantities $|\mathcal{E}(W)|$, measuring a set of edges, instead of their extremities outside W, which are the vertices that one can reach in one step from W. In other words, why not study what might be called the *vertex-expansion constant* defined by

$$\widetilde{h}(\Gamma) = \min_{1\leqslant|W|\leqslant|\Gamma|/2} \frac{|\partial W|}{|W|}, \tag{3.5}$$

where

$$\partial W = \{x \in V_\Gamma \mid x \notin W, \quad d_\Gamma(x,y) = 1 \text{ for some } y \in W\}$$

is the *boundary* of W?

The answer is that the definition is to some extent a convention, but that the one we used fits better with the idea of measuring “robustness”: two vertices x, y, with $x \in W$, $y \notin W$, which are linked with more than one edge (the case where the counting of edges diverges from that of vertices) are “better connected” than if there is only one edge with $\mathrm{ep}(\alpha) = \{x,y\}$, since cutting one of them would not disrupt the connection.

However, in the setting of expander families, it turns out that there is no difference in the class of graphs distinguished by the expansion constant and the variant (3.5). This follows from an easy lemma:

Lemma 3.1.15. *Let $\Gamma = (V, E, \mathrm{ep})$ be a non-empty finite graph with maximal valency v and let $W \subset V$ be any subset. We have*

$$\frac{1}{v}|\mathcal{E}(W)| \leqslant |\partial W| \leqslant |\mathcal{E}(W)|.$$

In particular, a family (Γ_i) of graphs of increasing size is an expander family if and only if it has bounded valency, and there exists $\widetilde{h} > 0$ such that

$$\widetilde{h}(\Gamma_i) \geqslant \widetilde{h}$$

for all $i \in I$.

Proof. Consider the map

$$\begin{cases} \mathcal{E}(W) & \longrightarrow & \partial W \\ \alpha & \mapsto & \mathrm{ep}(\alpha) \cap (V - W) \end{cases}$$

which sends an edge in $\mathcal{E}(W)$ to the one among its extremities which is not in W. By definition, this map is surjective, which gives the second inequality of the lemma, and there are at most v edges which map to any given $x \in \partial W$, which means that $|\mathcal{E}(W)| \leqslant v|\partial W|$. □

In the case of a bipartite graph, yet another variant of the expansion constant is obtained by looking just at subsets of the input vertices. Precisely, for a finite bipartite graph $\Gamma = (V, E, ep)$ with a decomposition $V = V_0 \cup V_1$, we let

$$\widehat{h}(\Gamma) = \min(h_0, h_1), \qquad h_i = \min_{\substack{W \subset V_i \\ 1 \leqslant |W| \leqslant |V_i|/2}} \frac{|\partial W|}{|W|} \tag{3.6}$$

(note that all the vertices in the boundary of a subset of V_i are in the other part). We then have:

Lemma 3.1.16. *Let $\Gamma = (V, E, \mathrm{ep})$ be a finite bipartite graph with a bipartiteness decomposition $V = V_0 \cup V_1$ and with maximal valency $v \geqslant 1$. Assume further that $|V_0| = |V_1|$. We have*

$$\frac{\widehat{h}(\Gamma) - 1}{2} \leqslant h(\Gamma) \leqslant v\widehat{h}(\Gamma).$$

Proof. The upper-bound is easy and left as an exercise. For the lower-bound, we may assume that $\widehat{h}(\Gamma) \geqslant 1$, say $\widehat{h}(\Gamma) = 1 + \delta$ with $\delta \geqslant 0$; we must then check that $h(\Gamma) \geqslant \frac{\delta}{2}$. Let then $W \subset V$ be any subset with $1 \leqslant |W| \leqslant \frac{1}{2}|V|$. We write $W = W_0 \cup W_1$ with $W_i = W \cap V_i$. Up to exchanging V_0 and V_1, which does not affect $\widehat{h}(\Gamma)$, we can assume that $|W_1| \leqslant |W_0|$, and in particular that $|W_0| \geqslant \frac{1}{2}|W|$. We now distinguish two cases:

(1) If $|W_0| \leqslant \frac{1}{2}|V_0|$, we deduce by definition that

$$|\partial W_0| \geqslant (1 + \delta)|W_0|.$$

Among the neighbors of W_0, at most $|W_1| \leqslant |W_0|$ belong to W. So W_0 has at least $\delta|W_0| \geqslant \frac{\delta}{2}|W|$ neighbors (in V_1) which are not in W. Hence we get

$$\frac{|\mathcal{E}(W)|}{|W|} \geqslant \frac{\delta}{2}.$$

(2) If $|W_0| > \frac{1}{2}|V_0|$, we deduce

$$|\partial W_0| \geqslant (1 + \delta)\frac{|V_0|}{2},$$

by applying the definition to a subset of W_0 of size $\lceil |V_0|/2 \rceil$. But $|W_1| \leqslant |V_0|/2$ since

$$|W| = |W_0| + |W_1| \leqslant |V| = \frac{|V_0|}{2},$$

(recall that we assume that V_0 and V_1 have the same number of elements). Thus ∂W_0 contains at least $\frac{\delta}{2}|V_0| = \frac{\delta}{4}|V| \geqslant \frac{\delta}{2}|W|$ neighbors not in W, and hence we obtain again

$$\frac{|\mathcal{E}(W)|}{|W|} \geqslant \frac{\delta}{2}.$$

□

The expander property can be thought of as relatively qualitative, and in particular it is fairly robust to certain changes of the structure (especially of the edges) of the graphs. Here is a fairly convenient lemma in this direction:

Lemma 3.1.17 (Comparison of expansion constants). *Let $\Gamma_1 = (V_1, E_1, \mathrm{ep})$ and $\Gamma_2 = (V_2, E_2, \mathrm{ep})$ be non-empty finite graphs with distances d_1 and d_2 respectively, and maximal valencies bounded by v_1 and v_2 respectively, and let $f : V_1 \longrightarrow V_2$ be a surjective map such that:*

(1) *For all $y \in V_2$, the set $f^{-1}(y)$ has the same cardinality $d \geqslant 1$, in particular $|V_1| = d|V_2|$;*

(2) *There exists $C > 0$ for which $d_2(f(x), f(y)) \leqslant Cd_1(x, y)$ for all x, $y \in V_1$.*

Then we have

$$h(\Gamma_2) \geqslant \frac{h(\Gamma_1)}{w},$$

where $w > 0$ depends only on (C, v_1, v_2), namely

$$w = v_1 \sum_{j=1}^{\lfloor C \rfloor} v_2^{j-1}.$$

We emphasize that f is *not* assumed to be a graph map, so that the condition $d_2(f(x), f(y)) \leqslant d_1(x, y)$ is not automatic (as it is for graph maps, as stated in Proposition 2.2.8).

Proof. Let $W \subset V_2$ be a non-empty set of vertices with $|W| \leqslant |V_2|/2$. Assumption (1) implies that $W' = f^{-1}(W) \subset V_1$ is (non-empty and) of size $\leqslant |V_1|/2$. Thus we get

$$|\mathcal{E}(W')| \geqslant h(\Gamma_1)|W'| = dh(\Gamma_1)|W|.$$

Since we are not assuming that f is a graph map, we do not know what it will do to edges, and hence we "convert" this inequality to the boundary vertices of W', getting

$$dh(\Gamma_1)|W| \leqslant |\mathcal{E}(W')| \leqslant v_1|\partial W'|$$

(using Lemma 3.1.15).

Using (1) once more, we have $|\partial W'| = d|f(\partial W')|$, hence

$$h(\Gamma_1)|W| \leqslant v_1|f(\partial W')|.$$

Since any $x \in \partial W'$ satisfies $d_1(x, x_0) = 1$ for some $x_0 \in W'$, assumption (2) gives

$$f(\partial W') \subset W'' = \{y \in V_2 - W \mid d_2(y, W) \leqslant C\}.$$

By induction on $j \geqslant 1$, we have

$$|\{y \in V_2 \mid y \notin W, \quad d_2(y, W) = j\}| \leqslant v_2^{j-1}|\partial W| \leqslant v_2^{j-1}|\mathcal{E}(W)|,$$

(using again Lemma 3.1.15) and hence

$$|f(\partial W')| \leqslant |W''| \leqslant \Bigl(\sum_{j=1}^{\lfloor C \rfloor} v_2^{j-1}\Bigr)|\mathcal{E}(W)|,$$

which leads to the inequality

$$|\mathcal{E}(W)| \geqslant \frac{h(\Gamma_1)}{v}|W|, \qquad v = v_1 \sum_{j=1}^{\lfloor C \rfloor} v_2^{j-1},$$

and hence to the conclusion. □

Here are applications to expanders, showing that certain "perturbations" of a family of expander graphs do not affect its expansion features.

Corollary 3.1.18. *Let (Γ_i) be a family of expander graphs, $\Gamma_i = (V_i, E_i, \mathrm{ep})$ with maximal valency bounded by v.*

(1) *For $i \in I$, let Γ'_i be any graph with the same vertex set as Γ_i and with "more edges", i.e., $E_{\Gamma'_i} \supset E_i$. Then (Γ'_i) is also an expander graph provided the maximal valency of Γ'_i remains bounded.*

(2) *The family of simple graphs* (Γ_i^s) *is a family of expanders.*

(3) *More generally assume that, for any* $i \in I$, *we are given graphs* $\Gamma_i' = (V_i', E_i', \mathrm{ep})$ *with maximal valency bounded by* w, *and bijections* $V_i \xrightarrow{f_i} V_i'$, *such that*

$$d_{\Gamma_i'}(f_i(x), f_i(y)) \leqslant C d_{\Gamma_i}(x, y) \tag{3.7}$$

for some fixed constant $C \geqslant 0$.

Then the family $(\Gamma_i')_{i\in I}$ *is also an expanding family. Precisely, it satisfies*

$$\inf_{i\in I} h(\Gamma_i') \geqslant \delta \inf_{i\in I} h(\Gamma_i),$$

where $\delta > 0$ *satisfies*

$$\delta^{-1} = v \sum_{j=1}^{\lfloor C \rfloor} w^{j-1},$$

i.e., if (h, v) *are expansion parameters for* (Γ_i), *then* $(\delta h, w)$ *are expansion parameters for* (Γ_i').

The example to keep in mind for (3) is when $V_i' = V_i$ and f_i is simply the identity. This means that Γ_i' is a graph with the same vertices, but with edges "rearranged" in some way, and the condition (3.7) says that the distance between vertices in the new graphs (using the modified edges) is distorted at most by a constant factor from the distance in the original ones. This will be particularly useful for Cayley graphs.

Proof. In each case, we only need apply Lemma 3.1.17 to compare each Γ_i with a graph Γ_i', which has the same set of vertices (i.e., the graphs (Γ_1, Γ_2) in the Lemma are (Γ_i, Γ_i')), and with f being the identity (so that Condition (1) of the Lemma is automatic).

In (1), because we *added* edges, we have $d_{\Gamma_i'}(x, y) \leqslant d_{\Gamma_i}(x, y)$ for each x and y, so we can take $C = 1$ in Condition (2) of the Lemma. In (2), with $\Gamma_i' = \Gamma_i^s$, although we may have removed some edges, we have not changed the distance (Exercise 2.2.10), so Condition (2) holds again with $C = 1$. Finally, in (3), Condition (2) is precisely given by (3.7). □

Remark 3.1.19. Some of the previous results (Corollary 3.1.10, Proposition 3.1.12 and (3) in this Corollary) can be interpreted in two ways: first, as a rough qualitative expression of properties of expanders (logarithmic growth of the diameter, exponential growth of balls, stability under "localized" changes of edge sets), but also as *quantitative* expressions of these properties, since in each case one can write down precise inequalities in terms of the expansion parameters of the family. As is often the case, the actual value of the constants appearing in such inequalities should not be considered as particularly significant for a first understanding of the intuitive meaning. Nevertheless, it is very important for certain applications that it is indeed possible to control these constants explicitly.

At this point, the most pressing question is: *do expanders really exist*? In all the easy examples of graphs (with bounded valency) for which we computed the expansion constant, it tends to 0 as the number of vertices goes to infinity, even in the case of finite trees where the diameter, at least, has the right order of magnitude. A pessimist's attitude might be that this is a bad sign.

An optimist might observe that, in the case of the "best" candidates so far (the finite trees $T_{d,k}$ with $d \geqslant 3$ fixed and $k \to +\infty$), there are many subsets of vertices which *do* have large expansion ratio $|\mathcal{E}(W)|/|W|$. Roughly speaking, as long as W is a set of vertices that only contains a few elements at the maximal distance k from the root of the tree, there will be many edges "escaping" further away from the root, in fact typically as many as the size of W. In other words, one might imagine that adding an edges to each of the far vertices, reconnecting them to the middle of the tree, *might* have a chance of producing graphs with good expansion constant.

We will not actually proceed this way; but, indeed, the optimists are in the right here: expanders do exist, and in fact exist in cheerful abundance. We will prove this in Chapter 4 using four different methods, in particular in Section 4.1 using probabilistic methods, as originally done by Barzdin and Kolmogorov [5], and independently by Pinsker [95]. The reader may skip to that section right now, since it is independent of what follows.

However, what we will do in the next two sections is provide definitions of other families of graphs, which turn out to be equivalent with the class of expanders, and which are often more flexible – and indeed, more important, in some applications.

3.2. Random walks

The definition of expansion constant (and consequently of expander graphs) does not provide an easy or direct way of computing $h(\Gamma)$. In terms of the number of vertices, which is a natural parameter coding the size of a graph of bounded valency, the exact determination of $h(\Gamma)$ requires looking at all subsets containing at most half of the vertices, a number of sets which is exponentially large in terms of $|\Gamma|$. In this section and the next, we will describe another invariant that can be estimated, in practice, much more easily, and which controls to some extent the expansion constant. In particular, we will be able to give an alternative definition of expander graphs.

The idea can be motivated by looking at the proof of Proposition 3.1.5: to show that the distance between x and y is "small", we looked at bigger and bigger balls around the two vertices, until they intersect. We are going to study what happens when we move in this way among the vertices of the graph. And because we do not know how to actually choose the best path at each step, we will consider *random* walks, and study their asymptotic behavior.

Definition 3.2.1 (Random walk on a graph). Let $\Gamma = (V, E, \mathrm{ep})$ be a countable graph with bounded valency at each vertex. A *random walk on* Γ is a sequence $(X_n)_{n\geqslant 0}$ of V-valued random variables, defined on a common probability space $(\Omega, \Sigma, \mathbf{P})$, with joint distribution satisfying the following rule: for any $n \geqslant 0$, and any vertices x_0, ..., x_n and $y \in V$, with

$$d_\Gamma(x_i, x_{i+1}) \leqslant 1, \qquad 0 \leqslant i \leqslant n-1,$$

we have

$$\begin{aligned}(3.8)\quad \mathbf{P}\{X_{n+1} = y \mid (X_n, \ldots, X_0) = (x_n, \ldots, x_0)\} &= \mathbf{P}\{X_{n+1} = y \mid X_n = x_n\} \\ &= \begin{cases} 0 & \text{if } d_\Gamma(x_n, y) > 1, \\ \dfrac{|\{\alpha \in E \mid \mathrm{ep}(\alpha) = \{x_n, y\}|\}}{\mathrm{val}(x_n)} & \text{if } d_\Gamma(x_n, y) = 0 \text{ or } 1. \end{cases}\end{aligned}$$

In other words, if X_n is at the vertex x, then X_{n+1} is determined by moving at step $n+1$ to an adjacent vertex y, using a randomly, uniformly, chosen edge connecting x to y, the choice being independent of the past history of the walk. This includes the possibility that $X_{n+1} = x$, which may only happen if there is a loop at x.

The distribution of the step X_0 of the walk is called the *initial distribution*. It is characterized by the probabilities $\mathbf{P}(X_0 = x)$ for $x \in V$. If we have $X_0 = x_0$ almost surely for a certain vertex $x_0 \in V$, i.e., at time 0, the walk always starts at x_0, then the random walk is called the random walk *starting from* x_0.

Remark 3.2.2. (1) We recall that if A, $B \in \Sigma$ are events in a probability space $(\Omega, \Sigma, \mathbf{P})$, the conditional probability of A knowing B, denoted $\mathbf{P}(A \mid B)$, is defined by

$$\mathbf{P}(A \mid B) = \begin{cases} \dfrac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}, & \text{if } \mathbf{P}(B) \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

(2) In probabilistic terms, the definition says that (X_n) is a *Markov chain* with *state space* V and with *transition matrix* $P = (P(x,y))_{x,y \in V}$ given by

$$P(x,y) = \begin{cases} 0 & \text{if } d_\Gamma(x,y) > 1, \\ \dfrac{a(x,y)}{\mathrm{val}(x)} & \text{if } d_\Gamma(x,y) = 0 \text{ or } 1, \end{cases}$$

where $a(x,y)$ is the (x,y)-coefficient of the adjacency matrix (Definition 2.1.4). In fact, this can be shortened to

$$P(x,y) = \frac{a(x,y)}{\mathrm{val}(x)} \tag{3.9}$$

for all x and y, since $a(x,y) = 0$ when x and y are not joined by at least one edge.

For an introduction to Markov chains, in greater generality but with an emphasis which is similar to the topics of this book, we refer to the book [76] of Levin, Peres and Widmer. A book dedicated to random walks on graphs in particular (especially in the infinite case) is [119] by Woess. We note that the subject of random walks is quite fascinating, both as an intrinsic subject (with its own problems) and because of its interactions with other fields; our presentation will be far from doing it justice!

For any given initial probability distribution μ_0 on V, determined by the probabilities $\mu_0(x)$ for $x \in V$, there is a random walk on Γ (on some probability space) for which X_0 has distribution μ_0, i.e.

$$\mathbf{P}(X_0 = x) = \mu_0(x).$$

This existence statement is a standard fact in probability theory, which we will not prove (except for reducing it to another standard probabilistic statement in the important case when Γ is a Cayley graph, see Example 3.2.7 below); details can be found in [119, 1.B]. We observe, however, that this random walk is unique (given μ_0) in the sense that the joint distribution of the process (X_n), i.e., all values of probabilities of the type

$$\mathbf{P}(X_{n_1} = x_1, \ldots, X_{n_k} = x_k)$$

for $k \geqslant 1$, $n_1 < n_2 < \cdots < n_k$, and $(x_1, \ldots, x_k) \in V^k$, depend *only* on μ_0, and not on specific features of the construction. To give an example, let

$$q(w,x,y,z) = \mathbf{P}((X_0, X_1, X_2, X_3) = (w,x,y,z))$$

be the joint law of the first four steps. We have

$$\begin{aligned} q(w,x,y,z) &= \mathbf{P}(X_3 = z \mid (X_0, X_1, X_2) = (w,x,y))\, \mathbf{P}((X_0, X_1, X_2) = (w,x,y)) \\ &= P(y,z)\, \mathbf{P}(X_2 = y \mid (X_0, X_1) = (w,x))\, \mathbf{P}((X_0, X_1) = (w,x)) \\ &= P(y,z)P(x,y)\, \mathbf{P}(X_1 = x \mid X_0 = w)\, \mathbf{P}(X_0 = w) \\ &= P(y,z)P(x,y)P(w,x)\mu_0(\{w\}), \end{aligned}$$

which is determined by P and μ_0, as claimed.

Another elementary consequence of the Markov property is the following useful fact: "starting" a walk after some steps of a random walk (X_n) on Γ also leads to a similar random walk on the graph.

Proposition 3.2.3. *Let* $\Gamma = (V, E, \mathrm{ep})$ *be a countable graph with finite valency at each vertex and let* (X_n) *be a random walk on* Γ*. Fix an integer* $m \geqslant 0$*, and define* $Y_n = X_{m+n}$ *for* $n \geqslant 0$*. Then* (Y_n) *is a random walk on* Γ *with initial distribution* μ_1 *given by the law of* X_m*, i.e., by*

$$\mu_1(A) = \mathbf{P}(X_m \in A)$$

for any $A \subset V$*.*

The proof is also left as an exercise. Indeed, the reader may want to try to prove the following stronger version:

Exercise 3.2.4 (Markov property and stopping time). Let $\Gamma = (V, E, \mathrm{ep})$ be a countable graph with finite valency at each vertex, and (X_n) a random walk on Γ.

(1) Let τ be a random variable taking non-negative integer values such that, for any $n \geqslant 0$, the event $\{\tau = n\}$ can be described only using X_0, ..., X_n. For $n \geqslant 0$, define $Y_n = X_{\tau+n}$, or in other words

$$Y_n(\omega) = X_{\tau(\omega)+n}(\omega)$$

in terms of elementary events $\omega \in \Omega$. Show that (Y_n) is a random walk on Γ with initial distribution given by the law of X_τ (note that we can take τ to be constant, equal to some integer $m \geqslant 0$, and the result is then Proposition 3.2.3).

(2) Show that if $A \subset V$ is a fixed subset, the "hitting time"

$$\tau_A = \min\{n \geqslant 0 \mid X_n \in A\} \in \{0, 1, \ldots, \} \cup \{+\infty\},$$

has the desired property (here we allow τ to take the value infinity, in case A is never reached from certain starting points).

Random variables τ with the property above that $\{\tau \leqslant n\}$ can be described using the process "up to time n" are called "stopping times" and are very important in the development of random walks in general; see [76, §6.2] for an introduction and examples.

Exercise 3.2.5. We present here another somewhat similar result that shows how random walks are "preserved". Let $\Gamma = (V, E, \mathrm{ep})$ be a countable graph with finite valency at each vertex, and (X_n) a random walk on Γ. Fix some integer $m \geqslant 0$ and $x \in V$. Let S be the event $\{X_m = x\}$. We assume that $\mathbf{P}(S) \neq 0$. We then consider S as probability space with the conditional measure $\mathbf{P}_S(B) = \mathbf{P}(S \cap B)/\mathbf{P}(S)$ for $B \subset S$.

(1) Show that the sequence $Y_n = X_{n+m}$, restricted to S, defines a random walk on Γ with initial distribution $Y_0 = x$. (One says that (Y_m) is the random walk (X_n) *conditioned to start from* x *at time* m).

(2) More generally, let τ be a stopping time, and S the event $\{\tau = m\}$. If $\mathbf{P}(S) > 0$, consider S as probability space with the conditional measure as described in (1). Show that the sequence $Y_n = X_{n+m}$ restricted to S defines a random walk on Γ with initial distribution $Y_0 = X_m$.

Remark 3.2.6. A somewhat subtle point in defining random walks (or Markov chains in general) is that the distributions of the individual variables (X_n) are determined uniquely by the weaker conditions

$$\mathbf{P}(X_{n+1} = y \mid X_n = x) = P(x, y) \tag{3.10}$$

(without requiring information on conditioning further back in the history). Indeed, for instance, we have

$$\mathbf{P}(X_1 = y) = \sum_{x \in V} \mathbf{P}(X_1 = y \mid X_0 = x)\, \mathbf{P}(X_0 = x) = \sum_{x \in V} \mu_0(\{x\}) P(x, y),$$

by (3.10) and then inductively

$$\mathbf{P}(X_{n+1} = y) = \sum_{x \in V} \mathbf{P}(X_{n+1} = y \mid X_n = x)\, \mathbf{P}(X_n = x) = \sum_{x \in V} \mathbf{P}(X_n = x) P(x, y),$$

for $n \geqslant 0$, which leads to this result. In the remainder of this section, we will concentrate our attention on the distribution of a single (X_n), so the reader should not be surprised to see us use only (3.10) in our treatment. However, in Section 5.2, we will consider some applications of random walks where the full property (3.8) is used (see Proposition 5.2.7).

Example 3.2.7 (Random walk on a Cayley graph). Let $\Gamma = \mathcal{C}(G, S)$ be the Cayley graph of a countable group G with respect to a finite symmetric set $S \subset G$. Then Γ is countable with finite degree $|S|$ at each vertex. Let (X_n) be a random walk on Γ starting at $x_0 = 1 \in S$; then the distribution of the n-th step is given by

$$\mathbf{P}(X_n = g) = \frac{1}{|S|^n} \sum_{\substack{(s_1, \ldots, s_n) \in S^n \\ s_1 \cdots s_n = g}} 1.$$

Conversely, using this formula, we can construct the random walk on Γ. To do this, let $(\xi_n)_{n \geqslant 1}$ be a sequence of *independent* S-valued random variables, each identically uniformly distributed, so that

$$\mathbf{P}(\xi_n = s) = \frac{1}{|S|},$$

for each $s \in S$ and $n \geqslant 1$, and (expressing independence)

$$\mathbf{P}(\xi_{n_1} = s_1, \ldots, \xi_{n_k} = s_k) = \frac{1}{|S|^k}$$

for any $k \geqslant 0$, any choices of distinct indices $n_1, \ldots,\ n_k$ and any $(s_1, \ldots, s_k) \in S^k$. Define the sequence (X_n) by

$$X_0 = 1, \qquad X_n = \xi_1 \cdots \xi_n.$$

This sequence of G-valued random variables is a random walk on Γ with $X_0 = 1$. To obtain a random walk with another initial distribution μ_0, pick any G-valued random variable X_0 with distribution given by μ_0, and let

$$X_n = X_0 \xi_1 \cdots \xi_n.$$

Exercise 3.2.8. Check that (X_n) is indeed a random walk on $\mathcal{C}(G, S)$, i.e., that (3.8) holds.

We get an analogue of Proposition 3.2.3:

Proposition 3.2.9. *Let G be a countable group and S a finite symmetric subset of G. Let $\Gamma = \mathcal{C}(G, S)$ and let (X_n) be a random walk on Γ starting at $X_0 = 1$ with independent increments (ξ_n), as above. For any m and $n \geqslant 0$, $Y_n = X_m^{-1} X_{m+n}$ is independent of X_m and distributed like X_n. Similarly, X_n^{-1} is distributed like X_n, i.e., X_n is* symmetric.

Proof. Indeed, we have

$$Y_n = \xi_{m+1} \cdots \xi_{m+n},$$

which is visibly independent of ξ_1, ..., ξ_m, since all the increments are independent, and is distributed like

$$\xi_1 \cdots \xi_n = X_n$$

since all increments are also identically distributed (and, again, independent).

Similarly, we have $X_n^{-1} = \xi_n^{-1} \cdots \xi_1^{-1}$, and since S is symmetric, we know that ξ_i is symmetrically-distributed, so that X_n^{-1} has the same distribution as X_n. □

Take for instance the case of the group $\mathbf{Z}^r$ for some $r \geqslant 1$, and denote by S the vectors in the canonical basis of $\mathbf{Z}^r$ together with their opposites. The resulting random walk on $\mathcal{C}(\mathbf{Z}^r, S)$ is called the *simple random walk* on $\mathbf{Z}^r$. This is historically the first example of a random walk to have been studied, and Polyá proved the first important theorem in the subject: if $r = 1$ or $r = 2$, then the simple random walk on $\mathbf{Z}^r$ is recurrent, in the sense that

$$\mathbf{P}(X_n = 0 \text{ for some } n \geqslant 1) = 1$$

(i.e., almost surely, a random walk on $\mathbf{Z}$ or $\mathbf{Z}^2$ will return to the origin; it is then fairly simple to see that, almost surely, the walker will come back infinitely often), while the simple random walk is *not* recurrent when $r \geqslant 3$ (see, e.g., [76, Ch. 21]).

Exercise 3.2.10. Let $\Gamma = (V, E, \mathrm{ep})$ be a finite connected graph. For any vertex x of Γ, we denote by τ_x the random variable on Γ such that

$$\tau_x = \min\{k \mid V = \{X_0^{(x)}, \ldots, X_k^{(x)}\}\}$$

where $(X_n^{(x)})$ is the random walk on Γ starting at x (in other words, $\tau_x = k$ means that the random walk has visited all vertices from time 0 to k, but not before). We define $T_x = \mathbf{E}(\tau_x)$. The maximum of T_x over $x \in V$ is called the *cover time* $T(\Gamma)$ of Γ.

(1) Show that $T(\Gamma)$ is finite.

(2) Show that $T(K_n) \sim n \log(n)$ for the complete graph K_n. [Hint: Compare with the *coupon collector problem*, see [76, 2.2].]

(3) Show that $T(C_n) = n(n-1)/2$ for the n-cycle with $n \geqslant 2$. [Hint: Show that $T(C_n)$ is the expected value of the (random) time t_n taken by the random walk on $\mathcal{C}(\mathbf{Z}, \pm 1)$ to have n distinct values; show that $\mathbf{E}(t_n) = \mathbf{E}(t_{n-1}) + n - 1$ for $n \geqslant 1$.]

In the case of a random walk on a finite *connected* graph, it is relatively simple to understand qualitatively the asymptotic distribution of X_n. Indeed, with an exception in the case of bipartite graph, this distribution converges to the uniform distribution on the set of vertices. This means that if n is very large, then the walker is, roughly speaking, as likely to be located at time n at any of the vertices of the graph, independently of the starting point or of the steps taken to get there.

The basic philosophy of this section (and the next) is to investigate the rate of convergence to this limiting distribution. As we will see, it is always exponentially fast (in a precise sense), but the rate of exponential convergence is a crucial invariant of the graph. It is not too hard to understand, intuitively, that the more the graph is highly connected – in the sense of having a large expansion constant – the faster the random walk should mix and become uniform. Indeed, there exists a precise relation of this type.

We start by establishing the asymptotic distribution of the random walks on a finite connected graph. This is a special case of the basic results of the theory of finite Markov chains, but it very easy to prove from scratch. We start with definitions that do not require the graph to be finite.

Definition 3.2.11 (Measure and functions on a graph). Let $\Gamma = (V, E, \mathrm{ep})$ be a countable graph with V and E non-empty, without isolated vertex and with finite valencies, i.e., such that $1 \leqslant \mathrm{val}(x) < +\infty$ for all $x \in V$.

(1) The graph measure ν_Γ on Γ is the measure on V defined by

$$\nu_\Gamma(\{x\}) = \mathrm{val}(x)$$

for $x \in V$. If Γ is finite, then the normalized graph measure on Γ is the probability measure on V defined by

$$\mu_\Gamma(\{x\}) = \frac{\mathrm{val}(x)}{N}$$

for all $x \in V$, where

$$N = \sum_{x \in V} \mathrm{val}(x) > 0.$$

(2) The space of functions on Γ is the space $L^2(\Gamma, \nu_\Gamma)$, i.e., it is the vector space of all functions

$$\varphi : \Gamma \to \mathbf{C}$$

such that the series

$$\sum_{x \in V} \mathrm{val}(x) |\varphi(x)|^2 \tag{3.11}$$

converges. It is equipped with the corresponding Hilbert space structure, i.e., with inner product

$$\langle \varphi_1, \varphi_2 \rangle_\Gamma = \sum_{x \in V} \mathrm{val}(x) \varphi_1(x) \overline{\varphi_2(x)}.$$

If Γ is finite, we have $L^2(\Gamma, \nu_\Gamma) = L^2(\Gamma, \mu_\Gamma)$ as vector spaces; the inner product on $L^2(\Gamma, \mu_\Gamma)$ is

$$\langle \varphi_1, \varphi_2 \rangle = \frac{1}{N} \sum_{x \in V} \mathrm{val}(x) \varphi_1(x) \overline{\varphi_2(x)}.$$

There is also a useful formula for the norm $\|\varphi\|$ of $\varphi \in L^2(\Gamma, \nu_\Gamma)$ in terms of the adjacency matrix (see Definition 2.1.4), namely

$$\begin{aligned} \|\varphi\|^2 = \sum_{x \in V} \mathrm{val}(x) |\varphi(x)|^2 &= \sum_{x \in V} |\varphi(x)|^2 \sum_{y \in V} a(x, y) \\ &= \sum_{x, y \in V} a(x, y) |\varphi(x)|^2 \end{aligned} \tag{3.12}$$

where $(a(x, y))$ is the adjacency matrix, i.e.

$$a(x, y) = |\{\alpha \in E \mid \mathrm{ep}(\alpha) = \{x, y\}\}|.$$

(There is no issue in interchanging the sums over x and y in this argument, even if V is infinite, since the functions involved are non-negative.)

We remark immediately an important fact: the constant function 1 belongs to the space $L^2(\Gamma, \nu_\Gamma)$ if and only if Γ is finite; its norm is then $N^{1/2}$, and its norm in $L^2(\Gamma, \mu_\Gamma)$ is 1.

Exercise 3.2.12 (An alternative formula). Let $\Gamma = (V, E, \mathrm{ep})$ be a finite graph, and $\varphi \in L^2(\Gamma, \mu_\Gamma)$ of mean zero, i.e., $\langle \varphi, 1 \rangle = 0$. Show that

$$\|\varphi\|^2 = \frac{1}{2N^2} \sum_{x, y \in V} \mathrm{val}(x) \, \mathrm{val}(y) |\varphi(x) - \varphi(y)|^2. \tag{3.13}$$

Remark 3.2.13. (1) As usual, we will often drop the subscript Γ when the context is clear. We will most often simply write $\|\varphi\|$ for the norm of a function in $L^2(\Gamma,\nu_\Gamma)$ or in $L^2(\Gamma,\mu_\Gamma)$. There will be no danger of confusion between the two inner products in the case of a finite graph, because we will use *exclusively* the normalized measure μ_Γ in that case, unless specified otherwise.

(2) If Γ is d-regular for some $d \geqslant 1$, then the measure ν_Γ is d times the counting measure and (if Γ is finite) the measure μ_Γ is simply the normalized probability counting measure on V, namely

$$\mu_\Gamma(W) = \frac{|W|}{|V|}$$

for all $W \subset V$. This case will in fact occur very often, so the reader may read the remainder of this section first with this case in mind. For finite graphs, we have the comparison relation

$$\frac{v_-}{v_+}\frac{|W|}{|V|} \leqslant \mu_\Gamma(W) \leqslant \frac{v_+}{v_-}\frac{|W|}{|V|} \tag{3.14}$$

for all $W \subset V$, where

$$v_- = \min_{x\in V}\operatorname{val}(x), \qquad v_+ = \max_{x\in V}\operatorname{val}(x).$$

(3) Finally, we will also have the occasion to use the supremum norm

$$\|\varphi\|_\infty = \max_{x\in V}|\varphi(x)|$$

for $\varphi \in L^2(\Gamma,\nu_\Gamma)$. It is always well-defined for the graphs we consider, since for the series (3.11) to converge, the values $\varphi(x)$ must tend to 0 in the sense that only finitely many are larger than any given $\varepsilon > 0$.

In the finite case, since any two norms on a finite-dimensional vector space are equivalent, the supremum norm is comparable to the Hilbert space norm. Precisely, we have

$$\|\varphi\| \leqslant \|\varphi\|_\infty \leqslant \Big(\frac{N}{v_-}\Big)^{1/2}\|\varphi\|, \tag{3.15}$$

where the left-hand inequality is a classical fact which holds for any probability measure, while the right-hand inequality follows from

$$\max|\varphi(x)|^2 \leqslant \frac{1}{v_-}\sum_{x\in V}\operatorname{val}(x)|\varphi(x)|^2 = \frac{N}{v_-}\|\varphi\|.$$

(4) Here is a technical remark: we assumed that there is no isolated vertex (i.e., $\operatorname{val}(x) \geqslant 1$ for all $x \in V$) because otherwise the inner product is not positive-definite on the space of all functions satisfying (3.11). As dictated by measure theory, the "correct" definition of $L^2(\Gamma,\nu_\Gamma)$ is as the quotient of the space of functions on V satisfying (3.11) by the subspace of functions φ which are zero "almost everywhere" with respect to ν_Γ. Such a function φ is one which is zero on all vertices x with $\operatorname{val}(x) \geqslant 1$, which explain why the isolated vertices would be "invisible" from this point of view.

Exercise 3.2.14. Let $\Gamma = (V, E, \mathrm{ep})$ be a finite graph, and let (X_n) be the random walk on Γ with initial distribution the uniform probability measure μ_Γ. Show that, for all n, the random variable X_n is also distributed like μ_Γ. (Of course, X_n is *not* the same random variable as X_0, it simply has the same distribution.)

The basic lemma is a simple identity that connects the steps of the random walk with the *Markov operator* of the graph.

Definition 3.2.15. Let $\Gamma = (V, E, \mathrm{ep})$ be a countable graph with finite valencies and no isolated vertex. The *Markov averaging operator* on $L^2(\Gamma, \nu_\Gamma)$ is the linear map

$$M_\Gamma : \begin{cases} L^2(\Gamma, \nu_\Gamma) & \longrightarrow & L^2(\Gamma, \nu_\Gamma) \\ \varphi & \mapsto & M\varphi \end{cases}$$

such that

$$(M_\Gamma\varphi)(x) = \frac{1}{\mathrm{val}(x)} \sum_{\substack{\alpha\in E \\ \mathrm{ep}(\alpha)=\{x,y\}}} \varphi(y) = \frac{1}{\mathrm{val}(x)} \sum_{\substack{y\in V \\ d_\Gamma(x,y)\leqslant 1}} a(x,y)\varphi(y).$$

We will often simply write M instead of M_Γ when only one graph is involved.

The notation in the first formula for $M\varphi$ should be clear: the value at x of the function $M\varphi$ is the average, over all edges of which x is an extremity, of the values of φ at the "other" extremity, *including* the values $\varphi(x)$ corresponding to the possible loops at x. For instance, if Γ is the action graph $\mathcal{A}(\mathfrak{S}_5, S)$ of Example 2.3.17, we have

$$(M\varphi)(3) = \frac{1}{5}(\varphi(1) + 2\varphi(2) + \varphi(3) + \varphi(4)),$$

and

$$(M\varphi)(1) = \frac{1}{5}(3\varphi(2) + \varphi(3) + \varphi(5)).$$

We note that the condition $d_\Gamma(x,y) \leqslant 1$ can be omitted in the second expression for $M\varphi$, since the quantity

$$a(x,y) = |\{\alpha \in E \mid \mathrm{ep}(\alpha) = \{x,y\}\}|$$

is zero unless $d_\Gamma(x,y) \leqslant 1$. In terms of transition probabilities (3.9), we can write

$$(M\varphi)(x) = \sum_{y\in V} P(x,y)\varphi(y). \tag{3.16}$$

We also note that M is well-defined, i.e., the function $M\varphi$ also belongs to $L^2(\Gamma, \nu_\Gamma)$ if φ does, and in fact the operator M has norm $\leqslant 1$ as an endomorphism of the Hilbert space $L^2(\Gamma, \nu_\Gamma)$ (additional spectral properties of M will be discussed later). Indeed, applying the Cauchy-Schwarz inequality, we get

$$|(M\varphi)(x)|^2 \leqslant \frac{1}{\mathrm{val}(x)^2}\Bigl(\sum_y a(x,y)\Bigr)\Bigl(\sum_y a(x,y)|\varphi(y)|^2\Bigr) = \frac{1}{\mathrm{val}(x)}\sum_y a(x,y)|\varphi(y)|^2,$$

hence

$$\sum_{x\in V} \mathrm{val}(x)|(M\varphi)(x)|^2 \leqslant \sum_{x\in V}\sum_{y\in V} a(x,y)|\varphi(y)|^2 = \sum_{y\in V} \mathrm{val}(y)|\varphi(y)|^2 = \|\varphi\|^2. \tag{3.17}$$

Lemma 3.2.16 (Markov operator and random walk). *Let $\Gamma = (V, E, \mathrm{ep})$ be a countable graph with finite valencies and no isolated vertex. Let (X_n) be a random walk on Γ. For any function $\varphi \in L^2(\Gamma, \nu_\Gamma)$ and $n \geqslant 0$, we have*

$$\mathbf{E}(\varphi(X_{n+1})) = \mathbf{E}((M\varphi)(X_n)).$$

Proof. If Γ is finite, then one proof is to say that this statement is linear in terms of φ, and that the defining condition (3.8) of a random walk implies that the identity holds for φ the characteristic function of a singleton y. Since these functions form a basis of $L^2(\Gamma, \nu_\Gamma)$ when Γ is finite, the result follows.

In general, observe first that since φ and $M\varphi$ are in $L^2(\Gamma, \nu_\Gamma)$, they are also bounded, hence the random variables $\varphi(X_{n+1})$ and $(M\varphi)(X_n)$ are both bounded, and consequently integrable. Then we compute

$$\begin{aligned}
\mathbf{E}(\varphi(X_{n+1})) &= \sum_{y \in V} \varphi(y)\, \mathbf{P}(X_{n+1} = y) \\
&= \sum_{y \in V} \varphi(y) \sum_{x \in V} \mathbf{P}(X_n = x)\, \mathbf{P}(X_{n+1} = y \mid X_n = x) \\
&= \sum_{y \in V} \varphi(y) \sum_{x \in V} \frac{1}{\mathrm{val}(x)}\, \mathbf{P}(X_n = x) |\{\alpha \in E \mid \mathrm{ep}(\alpha) = \{x, y\}\}| \\
&= \sum_{x \in V} \psi(x)\, \mathbf{P}(X_n = x) = \mathbf{E}(\psi(X_n))
\end{aligned}$$

where

$$\psi(x) = \frac{1}{\mathrm{val}(x)} \sum_{y \in V} \varphi(y) a(x, y) = (M\varphi)(x).$$

The interchange of the sums over x and y is justified since

$$\Bigl| \frac{1}{\mathrm{val}(x)} \varphi(y)\, \mathbf{P}(X_n = x) a(x, y) \Bigr| = \frac{1}{\mathrm{val}(x)} |\varphi(y)|\, \mathbf{P}(X_n = x) a(x, y),$$

whose sum over x and y, by the same argument (for non-negative functions, where all manipulations are allowed) is equal to $\mathbf{E}(|\varphi|(X_{n+1}))$, which is finite. □

This basic induction relation gives immediately a "formula" for the distribution of the n-th step of a random walk in terms of the linear operator M and of the initial distribution.

Corollary 3.2.17. *Let $\Gamma = (V, E, \mathrm{ep})$ be a countable graph with finite valencies and no isolated vertex. Let (X_n) be a random walk on Γ. For any function $\varphi \in L^2(\Gamma, \nu_\Gamma)$ and any integer $n \geqslant 0$, we have*

$$\mathbf{E}(\varphi(X_n)) = \mathbf{E}((M^n \varphi)(X_0)),$$

In particular, if the random walk starts at x_0, we have

$$\mathbf{E}(\varphi(X_n)) = (M^n \varphi)(x_0).$$

We are now led to the investigation of the averaging operator M. Here, the choice of the measure μ_Γ is important, because the crucial self-adjointness property of M depends on it.

Proposition 3.2.18 (Spectral properties of the Markov operator). *Let $\Gamma = (V, E, \mathrm{ep})$ be a countable graph with finite valencies and no isolated vertex. Let M be the Markov averaging operator for Γ.*

(1) *For any function $\varphi \in L^2(\Gamma, \nu_\Gamma)$, we have*

$$\langle (\mathrm{Id} - M)\varphi, \varphi \rangle_\Gamma = \frac{1}{2} \sum_{x, y \in V} a(x, y) |\varphi(x) - \varphi(y)|^2, \tag{3.18}$$

$$\langle (\mathrm{Id} + M)\varphi, \varphi \rangle_\Gamma = \frac{1}{2} \sum_{x, y \in V} a(x, y) |\varphi(x) + \varphi(y)|^2. \tag{3.19}$$

(2) *The operator M is self-adjoint of norm $\leqslant 1$. It is bounded from above by the identity and from below by minus the identity.*

Part (2) combines three assertions: first, we have

$$\langle M\varphi_1, \varphi_2\rangle_\Gamma = \langle \varphi_1, M\varphi_2\rangle_\Gamma$$

for any functions φ_1, $\varphi_2 \in L^2(\Gamma, \nu_\Gamma)$ (self-adjointness), next, we have $\|M\varphi\|_\Gamma \leqslant \|\varphi\|$ for $\varphi \in L^2(\Gamma, \nu_\Gamma)$ (norm at most 1), and finally we have

$$-\langle \varphi, \varphi\rangle_\Gamma \leqslant \langle M\varphi, \varphi\rangle_\Gamma \leqslant \langle \varphi, \varphi\rangle_\Gamma \tag{3.20}$$

for all $\varphi \in L^2(\Gamma, \nu_\Gamma)$.

Proof. We start by proving the self-adjointness, which is a key property (ultimately, it relates to the fact that we are working with *unoriented* graphs). We have by definition

$$\begin{aligned}\langle M\varphi_1, \varphi_2\rangle_\Gamma &= \sum_{x\in V} \mathrm{val}(x)(M\varphi_1)(x)\overline{\varphi_2(x)}\\ &= \sum_{x\in V} \overline{\varphi_2(x)} \sum_{y\in V} \varphi_1(y)a(x,y)\\ &= \sum_{x,y\in V} a(x,y)\varphi_1(y)\overline{\varphi_2(x)} \qquad (3.21)\end{aligned}$$

and since $a(x,y) = a(y,x)$, this is also $\langle \varphi_1, M\varphi_2\rangle_\Gamma$. The interchange of the sums over x and y (in the infinite case) is possible because

$$\sum_{x,y} |a(x,y)\varphi_1(y)\overline{\varphi_2(x)}| = \langle M|\varphi_1|, |\varphi_2|\rangle_\Gamma < +\infty$$

by the same computation for non-negative elements of $L^2(\Gamma, \nu_\Gamma)$.

We now prove the formulas (3.18) and (3.19). Both are very similar and we deal only with the first one. Using (3.12), the symmetry of the adjacency matrix and (3.21), we have

$$\begin{aligned}2(\langle \varphi, \varphi\rangle_\Gamma - \langle M\varphi, \varphi\rangle_\Gamma) &= 2\sum_{x,y\in V} a(x,y)|\varphi(x)|^2 - 2\sum_{x,y\in V} a(x,y)\varphi(x)\overline{\varphi(y)}\\ &= \sum_{x,y\in V} a(x,y)|\varphi(x)|^2 + \sum_{x,y\in V} a(x,y)|\varphi(y)|^2 - 2\sum_{x,y\in V} a(x,y)\varphi(x)\overline{\varphi(y)}\end{aligned}$$

which is equal to

$$\sum_{x,y\in V} a(x,y)|\varphi(x) - \varphi(y)|^2$$

These formulas (3.22) and (3.23) immediately imply (3.20). From this, we get

$$|\langle M\varphi, \varphi\rangle_\Gamma| \leqslant \|\varphi\|^2$$

for all $\varphi \in L^2(\Gamma, \nu_\Gamma)$, and since it is standard that

$$\|M\| = \sup_{\varphi\neq 0} \frac{|\langle M\varphi, \varphi\rangle_\Gamma|}{\|\varphi\|^2}$$

for a self-adjoint operator, this gives $\|M\| \leqslant 1$; we could also notice that we already proved that $\|M\varphi\| \leqslant \|\varphi\|$ (see (3.17)) when checking that M is well-defined. □

Remark 3.2.19. Let $\Gamma = (V, E, \mathrm{ep})$ be a countable graph with bounded valencies. Define the *adjacency operator* A_Γ on $L^2(\Gamma, \nu_\Gamma)$ by

$$(A_\Gamma\varphi)(x) = \sum_{y\sim x} a(x,y)\varphi(y)$$

for any function $\varphi \in L^2(\Gamma, \nu_\Gamma)$. (In other words, this is the operator whose matrix is the adjacency matrix in the basis of characteristic functions of vertices of Γ, which justifies the abuse of notation involved in using the same letter as for the adjacency matrix). In general, we have $A_\Gamma = \mathrm{val} \cdot M_\Gamma$, where val denotes the linear map of multiplication by the function $x \mapsto \mathrm{val}(x)$. Since the valency is assumed to be bounded, we have $A_\Gamma\varphi \in L^2(\Gamma, \nu_\Gamma)$, which shows that A_Γ is an endomorphism of $L^2(\Gamma, \nu_\Gamma)$.

If Γ is d-regular for some integer $d \geqslant 1$, then we have $A_\Gamma = dM_\Gamma$, and A_Γ is also a self-adjoint endomorphism of $L^2(\Gamma, \nu_\Gamma)$. If Γ is *not* regular, then A_Γ is still self-adjoint, but with respect to a different inner-product, namely

$$(\varphi_1, \varphi_2) = \sum_{x\in V} \varphi_1(x)\overline{\varphi_2(x)}$$

(this is simply because the adjacency matrix is real and symmetric). See also Remark 3.4.2 below.

Corollary 3.2.20. *Let* $\Gamma = (V, E, \mathrm{ep})$ *be a non-empty* finite *graph with no isolated vertex.*

(1) *The Markov operator M is diagonalizable in an orthonormal basis of $L^2(\Gamma, \mu_\Gamma)$, its eigenvalues are real numbers, and all eigenvalues have absolute value at most* 1. *For $\varphi \in L^2(\Gamma, \mu_\Gamma)$, we have*

$$\langle(\mathrm{Id} - M)\varphi, \varphi\rangle = \frac{1}{2N}\sum_{x,y\in V} a(x,y)|\varphi(x) - \varphi(y)|^2, \tag{3.22}$$

$$\langle(\mathrm{Id} + M)\varphi, \varphi\rangle = \frac{1}{2N}\sum_{x,y\in V} a(x,y)|\varphi(x) + \varphi(y)|^2, \tag{3.23}$$

and

$$\langle M\varphi_1, \varphi_2\rangle = \frac{1}{N}\sum_{x,y\in V} a(x,y)\varphi_1(y)\overline{\varphi_2(x)} \tag{3.24}$$

where the inner product are with respect to the probability measure μ_Γ.

(2) *The* 1*-eigenspace* $\ker(M - 1)$ *of M has dimension equal to the number of connected components of Γ, and is spanned by the characteristic functions of these connected components. In particular, if Γ is connected, then we have* $\ker(M - 1) = \mathbf{C}$, *spanned by constant functions.*

(3) *If Γ is connected, the (-1)-eigenspace* $\ker(M + 1)$ *is zero* unless *Γ is bipartite. In that case, it is one-dimensional and spanned by a function $\varepsilon_\pm$ equal to* 1*, resp.* -1*, on the set of inputs, resp. outputs, of a bipartite decomposition of V.*

(4) *If Γ is bipartite, then the spectrum of M is symmetric: if λ is an eigenvalue of M, then so is $-\lambda$.*

Proof. (1) Since $\mu_\Gamma = \nu_\Gamma/N$ is a multiple of ν_Γ, the self-adjointness of M acting on $L^2(\Gamma, \nu_\Gamma)$ that we proved is equivalent to the self-adjointness on $L^2(\Gamma, \mu_\Gamma)$. Since Γ is finite, the space $L^2(\Gamma, \mu_\Gamma)$ is finite-dimensional, so by linear algebra, the endomorphism M is diagonalizable in an orthonormal basis of $L^2(\Gamma, \mu_\Gamma)$, and its eigenvalues are real. The formulas (3.22) and (3.23) are simply restatements, for the inner-product of $L^2(\Gamma, \mu_\Gamma)$, of the formulas (3.18) and (3.19), and similarly (3.24) restates (3.21).

(2) We next investigate the structure of $\ker(M-1)$ using (3.22), though there is a nice "geometric" computation also (see the exercise below). If $M\varphi = \varphi$, we get immediately the identity

$$\sum_{x,y \in V} a(x,y)|\varphi(x) - \varphi(y)|^2 = 0,$$

from (3.22). By positivity, this is equivalent with

$$\varphi(x) = \varphi(y)$$

whenever $a(x,y) \neq 0$, i.e., φ has the same value at all extremities of any edge. If we fix any $x_0 \in V$, and use induction on $d_\Gamma(x_0, x)$, we get $\varphi(x) = \varphi(x_0)$ for all x reachable by a path from x_0. This means that φ is constant on each connected component of Γ. The converse is easy: if φ is constant on each connected component, the definition shows that it does satisfy $M\varphi = \varphi$. Hence $\ker(M - 1)$ is the space spanned by characteristic functions of connected components in the graph. (Note that this computation seems to also apply to the case of infinite graphs; however, a non-zero constant function is *not* an element of $L^2(\Gamma, \nu_\Gamma)$ if Γ is infinite).

(3) We deal similarly with the possible -1 eigenvalue, for which we restrict our attention to connected graphs for simplicity. The reader should first check that, if Γ is bipartite, then the function $\varepsilon_\pm$ defined in the statement of the theorem is indeed in $\ker(M + 1)$. We now proceed to show that it generates the (-1)-eigenspace.

Let φ be such that $M\varphi = -\varphi$. We get from (3.23) that

$$\varphi(x) = -\varphi(y)$$

for all x and y connected by an edge. If $\gamma : P_2 \longrightarrow \Gamma$ is any path of length 2 with $\gamma(0) = x$, $\gamma(2) = y$, it follows that

$$\varphi(x) = -\varphi(\gamma(1)) = \varphi(y).$$

Iterating, we obtain $\varphi(x) = \varphi(\gamma(2k))$ for any path γ of even length $2k$. Now we fix some $x_0 \in V$, and let W be the set of vertices in Γ which are the other extremity of a path $\gamma : P_{2k} \longrightarrow \Gamma$ of even length with $\gamma(0) = x_0$ (in particular, $x_0 \in W$ using a path of length 0). We see that φ is constant, equal to $\varphi(x_0)$, on all of W. If $W = V$, it follows that φ is constant, hence $M\varphi = \varphi = -\varphi$, so $\varphi = 0$.

On the other hand, if $W \neq V$, we claim that $V_0 = W$, $V_1 = V - W$ is a bipartite partition of V. Indeed, let $\alpha \in E$ be an edge with extremities $\{x_1, x_2\}$. It is not possible that x_1 and x_2 are both in V_0: if that were to happen, then given any $y \in V_1$, we would get a path of even length joining x_0 to y by (1) going from x_0 to x_1 with a path of even length $2\ell_1$ (possible since $x_1 \in V_0$); (2) going to x_1 to x_2 by the path of length 1 given by α; (3) going from x_2 to x_0 with a path of even length $2\ell_2$ (again, because $x_2 \in V_0$); (4) going from x_0 to y, which is possible since Γ is connected, and possible with odd length $2\ell_3 + 1$ since $y \notin V_0$: the total length is

$$2\ell_1 + 1 + 2\ell_2 + 2\ell_3 + 1 \equiv 0 \pmod 2,$$

(see Figure 3.1 for the graphical illustration of this construction).

This contradicts the fact that $V_0 = W \neq V$. Similarly, we see that x_1, x_2 can not both be in V_1, and this concludes the proof that Γ is bipartite. It is now easy to finish determining φ: it is constant, equal to $\varphi(x_0)$, on V_0, and for any $x \in V_1$, finding $y \in V_0$ connected by an edge, we get $\varphi(y) = -\varphi(x) = -\varphi(x_0)$. Thus it is equal to $\varphi(x_0)\varepsilon_\pm$.

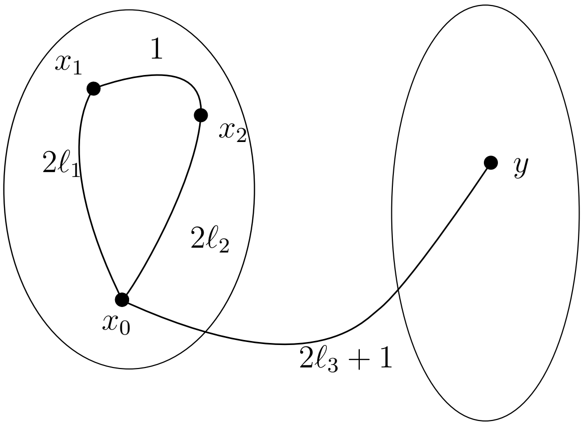


FIGURE 3.1. Bipartite graph

(4) Assume that Γ is bipartite with bipartite partition $V = V_1 \cup V_2$. Whenever $\varphi\colon V \to \mathbf{C}$ is a λ-eigenfunction of M, it follows that $\tilde{\varphi}$ defined by $\tilde{\varphi}(x) = \varphi(x)$ for $x \in V_1$ and $\tilde{\varphi}(x) = -\varphi(x)$ for $x \in V_2$ is a $-\lambda$-eigenfunction of M. □

EXERCISE 3.2.21. The following proof of $\|M\| \leqslant 1$, for Γ finite, does not use (3.20) and is also useful to keep in mind.

(1) Explain why the norm of M is the maximum of the absolute values of its eigenvalues.

(2) If λ is an eigenvalue, show directly that $|\lambda| \leqslant 1$. [Hint: Use the maximum norm instead of the L^2-norm.]

EXERCISE 3.2.22 (Maximum modulus principle). This exercise discusses the "geometric" computation of $\ker(M-1)$. We assume that Γ is a non-empty finite graph without isolated vertices.

(1) Show that if φ is the characteristic function of a connected component of Γ, we have $M\varphi = \varphi$.

(2) Show that, in order to prove that these characteristic functions span $\ker(M-1)$, it is enough to prove that a real-valued element of $\ker(M-1)$ is constant on each connected component of Γ.

(3) Let $W \subset V$ be a connected component. Let φ be a real-valued element of $\ker(M-1)$, let m be the maximum value of $\varphi(x)$ on W, and $x_0 \in W$ a vertex where $\varphi(x_0) = m$. Show that $\varphi(x) = m$ for all x connected to x_0 by at least one edge.

(4) Deduce that φ is equal to m on all of W and conclude.

(5) Using similar methods, determine $\ker(M+1)$.

EXERCISE 3.2.23 (Both sides have equal weight). Let Γ be a connected non-empty finite bipartite graph without isolated vertices, partitioned as $V = V_0 \cup V_1$ with all edges between V_0 and V_1. Show that

$$\mu_\Gamma(V_0) = \mu_\Gamma(V_1) = \frac{1}{2}.$$

The simple spectral properties of M are enough to understand the asymptotic behavior of a random walk on a *fixed* finite connected graph Γ. We define first the relevant invariant:

DEFINITION 3.2.24 (Equidistribution radius). Let $\Gamma = (V, E, \mathrm{ep})$ be a connected non-empty finite graph without isolated vertices. The *equidistribution radius* of Γ, denoted

ϱ_Γ, is the maximum of the absolute values $|\lambda|$ for λ an eigenvalue of M which is different from ± 1. Equivalently, ϱ_Γ is the spectral radius of the restriction of M to the subspace

$$L_0^2(\Gamma, \mu_\Gamma) = (\ker(M-1) \oplus \ker(M+1))^\perp,$$

i.e., (1) if Γ is not bipartite, the restriction to the space of $\varphi \in L^2(\Gamma, \mu_\Gamma)$ such that

$$\langle \varphi, 1 \rangle = \frac{1}{N} \sum_{x \in V} \mathrm{val}(x)\varphi(x) = 0,$$

and (2) if Γ is bipartite with bipartite partition $V_0 \cup V_1 = V$, the restriction to the space of $\varphi \in L^2(\Gamma, \mu_\Gamma)$ such that

$$\frac{1}{N} \sum_{x \in V} \mathrm{val}(x)\varphi(x) = 0, \qquad \frac{1}{N} \sum_{x \in V_0} \mathrm{val}(x)\varphi(x) = \frac{1}{N} \sum_{x \in V_1} \mathrm{val}(x)\varphi(x).$$

The equivalence of the stated definitions of ϱ_Γ, and of the subspace $L_0^2(\Gamma, \mu_\Gamma)$, are direct consequences of Proposition 3.2.18 (taking into account the assumption that Γ is connected). The following are also almost part of the definition:

Lemma 3.2.25. *Let $\Gamma = (V, E, \mathrm{ep})$ be a connected, non-empty, finite graph without isolated vertices. We have $0 \leqslant \varrho_\Gamma < 1$ and ϱ_Γ is given by*

$$\varrho_\Gamma = \max_{0 \neq \varphi \in L_0^2(\Gamma, \mu_\Gamma)} \frac{|\langle M\varphi, \varphi \rangle|}{\|\varphi\|^2}. \tag{3.25}$$

Proof. The inequality $\varrho_\Gamma < 1$ simply expresses the fact that M is self-adjoint with real eigenvalues of absolute value at most 1, so that, on the orthogonal complement $L_0^2(\Gamma, \mu_\Gamma)$ of the space spanned by the eigenspaces for ± 1, all eigenvalues have modulus < 1.

Similarly, the restriction of the self-adjoint operator M to $L_0^2(\Gamma, \mu_\Gamma)$ is self-adjoint, and its norm is ϱ_Γ. The formula (3.25) is then a standard property of endomorphisms of Hilbert spaces. □

Example 3.2.26. (1) Let $m \geqslant 3$ and consider the complete graph K_m on m vertices. The Markov operator is then given by

$$(M\varphi)(x) = \frac{1}{m-1} \sum_{y \neq x} \varphi(y).$$

for $\varphi \in L^2(K_m)$. Restricted to $L_0^2(K_m)$, which is the space of functions orthogonal to 1, since K_m is not bipartite for $m \geqslant 3$, this becomes

$$(M\varphi)(x) = \frac{1}{m-1} \Bigl(\sum_y \varphi(y) - \varphi(x) \Bigr) = -\frac{1}{m-1} \varphi(x).$$

Hence the unique eigenvalue of the Markov operator M is $-1/(m-1)$, with multiplicity $m-1$. In particular, we have $\varrho(K_m) = \frac{1}{m-1}$.

(2) The second simplest example of computation of ϱ_Γ is (probably) for the m-cycle C_m where $m \geqslant 2$. We will compute all eigenvalues of M and describe the eigenfunctions in that case. For a function

$$\varphi : \mathbf{Z}/m\mathbf{Z} \longrightarrow \mathbf{C}$$

(recall that we use $\mathbf{Z}/m\mathbf{Z}$ as vertex set for the cycle) and a vertex $x \in \mathbf{Z}/m\mathbf{Z}$, we have

$$M\varphi(x) = \frac{1}{2}(\varphi(x-1) + \varphi(x+1)).$$

In order to analyze this operator, we use the Fourier transform on $\mathbf{Z}/m\mathbf{Z}$, which is the linear map

$$\begin{cases} L^2(C_m) & \longrightarrow & L^2(C_m) \\ \varphi & \mapsto & \widehat{\varphi} \end{cases}$$

defined by

$$\widehat{\varphi}(a) = \frac{1}{m} \sum_{x \in \mathbf{Z}/m\mathbf{Z}} \varphi(x) e\Bigl(-\frac{ax}{m}\Bigr)$$

for $a \in \mathbf{Z}/m\mathbf{Z}$, where $e(z) = e^{2i\pi z}$ for $z \in \mathbf{C}$. We can also write

$$\widehat{\varphi}(a) = \langle \varphi, \chi_a \rangle$$

where $\chi_a(x)$ is the function $\chi_a(x) = e(ax/m)$, which is distinguished by being a *character* of $\mathbf{Z}/m\mathbf{Z}$, i.e., by $\chi_a(x+y) = \chi_a(x)\chi_a(y)$.

Because of this property, we see by a change of variable that if $\psi(x) = \varphi(x+b)$ for some fixed $b \in \mathbf{Z}/m\mathbf{Z}$, we have

$$\widehat{\psi}(a) = \frac{1}{m} \sum_{x \in \mathbf{Z}/m\mathbf{Z}} \varphi(x+b) e\Bigl(-\frac{ax}{m}\Bigr) = \chi_a(b) \frac{1}{m} \sum_{y \in \mathbf{Z}/m\mathbf{Z}} \varphi(y) e\Bigl(-\frac{ay}{m}\Bigr),$$

i.e., $\widehat{\psi} = \chi_a(b)\widehat{\varphi}$. In particular, it follows that

$$\widehat{M\varphi}(a) = \frac{\chi_a(1) + \chi_a(-1)}{2} \widehat{\varphi}(a) = \cos\Bigl(\frac{2\pi a}{m}\Bigr) \widehat{\varphi}(a).$$

In other words, M acts *diagonally* on Fourier transforms. But the Fourier transform is an isomorphism, as revealed by the inversion formula

$$\varphi(x) = \sum_{a \in \mathbf{Z}/m\mathbf{Z}} \widehat{\varphi}(a) \chi_a(x),$$

which means that φ is the Fourier transform of $a \mapsto m\widehat{\varphi}(-a)$. So we have found an explicit diagonalization of M for the cycle C_m. In fact, we find also directly that

$$M\chi_b = \cos\Bigl(\frac{2\pi b}{m}\Bigr)\chi_b,$$

and since these characters form an orthonormal basis of $L^2(\Gamma)$, they are a basis of eigenfunctions of M, with eigenvalues $\cos(2\pi b/m)$.

If m is odd, each eigenvalue except 1, for which $\ker(M-1)$ is one-dimensional, has a 2-dimensional eigenspace (spanned by χ_b and χ_{-b}), while if m is even, all eigenvalues except for 1 and -1 (which have 1-dimensional eigenspaces, in the second case because C_m is then bipartite) have a 2-dimensional eigenspace. In any case, we get

$$\varrho_{C_m} = \cos\Bigl(\frac{2\pi}{m}\Bigr) = 1 - \frac{2\pi^2}{m^2} + O(m^{-4})$$

for $m \geqslant 2$.

Exercise 3.2.27. Let G_3 be the Cayley graph $\mathcal{C}(\mathfrak{S}_3, S_3)$ which we drew in Example 2.1.6.

(1) Compute the matrix of the Markov operator of G_3 in the basis of characteristic functions of single points, and compute its spectrum and the equidistribution radius.

(2) Compute an orthonormal basis of $L^2(G_3)$ of eigenfunctions of M.

Corollary 3.2.28 (Convergence to equilibrium in random walks). *Let $\Gamma = (V, E, \mathrm{ep})$ be a connected, non-empty, finite graph without isolated vertices, and let (X_n) be a random walk on Γ.*

(1) *If Γ is* not *bipartite, then for any function $\varphi \in L^2(\Gamma)$, we have*

$$\Big|\mathbf{E}(\varphi(X_n)) - \langle \varphi, 1\rangle\Big| \leqslant \varrho_\Gamma^n \Big(\frac{N}{v_-}\Big)^{1/2} \|\varphi\|, \tag{3.26}$$

where $v_- = \min \mathrm{val}(x)$, and in particular

$$\lim_{n\to+\infty} \mathbf{P}(X_n = x) = \mu_\Gamma(x) = \frac{\mathrm{val}(x)}{N} \tag{3.27}$$

for all $x \in V$.

(2) *If Γ* is *bipartite with bipartite partition $V = V_0 \cup V_1$, then for any function $\varphi \in L^2(\Gamma)$, we have*

$$\Big|\mathbf{E}(\varphi(X_n)) - \Big\{m_0 + m_1 + (-1)^n (m_0 - m_1)(p_0 - p_1)\Big\}\Big| \leqslant \varrho_\Gamma^n \Big(\frac{N}{v_-}\Big)^{1/2} \|\varphi\|, \tag{3.28}$$

where

$$p_0 = \mathbf{P}(X_0 \in V_0), \qquad p_1 = \mathbf{P}(X_0 \in V_1),$$
$$m_0 = \frac{1}{N}\sum_{x\in V_0} \mathrm{val}(x)\varphi(x), \qquad m_1 = \frac{1}{N}\sum_{x\in V_1} \mathrm{val}(x)\varphi(x).$$

Proof. (1) The idea is to write

$$\varphi = \langle \varphi, 1\rangle + \varphi_0 = \alpha + \varphi_0$$

where $\alpha = \langle \varphi, 1\rangle$ is the average of φ and, by definition, we have $\varphi_0 \in L_0^2(\Gamma)$. Applying Corollary 3.2.17 and the fact that α is an eigenfunction of M with eigenvalue 1, we get

$$\mathbf{E}(\varphi(X_n)) = \mathbf{E}((M^n\varphi)(X_0)) = \alpha + \mathbf{E}((M^n\varphi_0)(X_0)).$$

By writing

$$\mathbf{E}((M^n\varphi_0)(X_0)) = \sum_{x\in V} \mathbf{P}(X_0 = x)(M^n\varphi_0)(x),$$

we get

$$|\,\mathbf{E}((M^n\varphi_0)(X_0))| \leqslant \|(M^n\varphi_0)\|_\infty,$$

and we are almost done. The last step is to compare this maximum norm with the L^2 norm, which we do with (3.15), from which we get

$$|\,\mathbf{E}((M^n\varphi_0)(X_0))| \leqslant \Big(\frac{N}{v_-}\Big)^{1/2} \|(M^n\varphi_0)\|.$$

Now the definition of ϱ_Γ (and the fact that M sends $L_0^2(\Gamma)$ to itself) leads immediately, by induction, to

$$\|(M^n\varphi_0)\| \leqslant \varrho_\Gamma^n \|\varphi_0\|,$$

from which the inequality (3.26) follows. The limit (3.27) is simply the special case when φ is the characteristic function of $x \in V$, in which case $\langle \varphi, 1\rangle = \mathrm{val}(x)/N$.

(2) The bipartite case is very similar, but we must now take into account the eigenvalue -1. We write

$$\varphi = \langle \varphi, 1\rangle + \langle \varphi, \varepsilon_\pm\rangle + \varphi_0,$$

with again $\varphi_0 \in L_0^2(\Gamma)$, and obtain

$$\mathbf{E}(\varphi(X_n)) = \mathbf{E}((M^n\varphi)X_0) = \langle \varphi, 1\rangle + (-1)^n \langle \varphi, \varepsilon_\pm\rangle\, \mathbf{E}(\varepsilon_\pm(X_0)) + \mathbf{E}((M^n\varphi_0)X_0).$$

The last term is estimated exactly as before:

$$|\mathbf{E}((M^n\varphi_0)X_0)| \leqslant \Big(\frac{N}{v_-}\Big)^{1/2}\varrho_\Gamma^n\|\varphi\|,$$

using $\varphi_0 \in L^2_0(\Gamma)$. We now note that

$$\langle\varphi, 1\rangle = m_0 + m_1, \qquad \langle\varphi, \varepsilon_\pm\rangle = m_0 - m_1$$

and that

$$\mathbf{E}(\varepsilon_\pm(X_0)) = p_0 - p_1,$$

to deduce that

$$\langle\varphi, 1\rangle + (-1)^n\langle\varphi, \varepsilon_\pm\rangle\,\mathbf{E}(\varepsilon_\pm(X_0)) = m_0 + m_1 + (-1)^n(m_0 - m_1)(p_0 - p_1).$$

□

Remark 3.2.29. (1) This result should also be remembered as a general template: in a number of applications, it is very useful to study more deeply the behavior of a sequence $\mathbf{E}(\varphi(X_n))$ by expanding φ in a full orthonormal basis of $L^2(\Gamma)$ of eigenfunctions of M (instead of just isolating the projection onto the 1 and (-1)-eigenspaces).

(2) The bipartite case is clearer when $X_0 = x_0$ is a fixed vertex, say $x_0 \in V_0$. Then $p_0 = 1$, $p_1 = 0$ and the "main term" for $\mathbf{E}(\varphi(X_n))$ becomes

$$\begin{aligned} m_0 + m_1 + (-1)^n(m_0 - m_1) &= (1 + (-1)^n)m_0 + (1 - (-1)^n)m_1 \\ &= \begin{cases} 2m_0 & \text{if } n \text{ is even,} \\ 2m_1 & \text{if } n \text{ is odd.} \end{cases} \end{aligned}$$

In this case, the sequence $\mathbf{E}(\varphi(X_n))$ does not converge in general (unless $m_0 = m_1$): it oscillates between the even terms which converge to twice the average of φ on V_0, and the odd ones which converge to twice the average of φ on V_1.

In particular, if φ is the characteristic function of a single vertex x_1, and (say) $x_1 \in V_1$, the probability that $X_n = x_1$ is zero, unless n is odd, and in that case it converges (exponentially fast) to $2\mu_\Gamma(x_1)$. The factor 2 can be interpreted as follows: since we know *a priori* that X_n is in V_1 for n odd, and $\mu_\Gamma(V_1) = \frac{1}{2}$ (Exercise 3.2.23), the probability that X_n be any fixed element of V_1 is twice as large than for a completely random vertex in V.

Example 3.2.30 (Time to reach equilibrium). For the characteristic function δ_x of a fixed vertex, we have

$$\|\delta_x\|^2 = \mu_\Gamma(x) = \frac{\mathrm{val}(x)}{N},$$

so that, if Γ is not bipartite, the precise statement is

$$\Big|\mathbf{P}(X_n = x) - \frac{\mathrm{val}(x)}{N}\Big| \leqslant \Big(\frac{\mathrm{val}(x)}{v_-}\Big)^{1/2}\varrho_\Gamma^n$$

for $n \geqslant 1$. When n is "small", this inequality is typically trivial, because the right-hand side still dominates the limiting value. A natural measure of the time when equidistribution becomes effective is the first index n when the "error" becomes comparable in size to $\frac{1}{2}\mu_\Gamma(x)$: for such n, the probability $\mathbf{P}(X_n = x)$ is at most off by a factor 2 from its limiting value. (Note that the statement also shows that the convergence is typically monotonic: the quality of approximation increases with n). To determine n, we simply write that we wish that

$$\varrho_\Gamma^n \leqslant \frac{(\mathrm{val}(x)v_-)^{1/2}}{2N},$$

and in the case of a k-regular graph, this means

$$n \geqslant \frac{\log(2|V|)}{\log(\varrho_\Gamma^{-1})}.$$

Hence, for n larger than some (possibly large, depending on how close ϱ_Γ is to 1) multiple of $\log |V|$, a random walk (X_n) becomes essentially equidistributed. This type of considerations turns out to play an important role in understanding the arguments of Chapter 6.

We see from Corollary 3.2.28 that ϱ_Γ controls the rate of convergence of a random walk on Γ to the normalized graph measure μ_Γ. On the other hand, it is intuitively natural to expect that the random walk should converge faster if the expansion constant $h(\Gamma)$ is larger, in which case there are intuitively "more" opportunities to explore (or get lost in) the graph. Hence one may expect that bounding ϱ_Γ away from 1 should be related to bounding $h(\Gamma)$ away from 0. This is indeed the case (up to a minor technical point), and we will discuss this in Section 3.3. Before this, we conclude this section with an important computation concerning random walks on infinite regular trees, which will play some role in Chapter 6. The reader may well decide to skip to the next section right away.

Proposition 3.2.31 (Kesten). *Let $d \geqslant 2$ be an integer. Let T_d be the infinite d-regular tree. The spectral radius of the Markov operator on $L^2(T_d, \nu_{T_d})$ is $\varrho = 2\sqrt{d-1}/d$.*

In particular, let $(X_n)_{n\geqslant 0}$ be the random walk on T_d with $X_0 = 1$. Then for all $n \geqslant 0$ and all $x \in T_d$, we have $\mathbf{P}(X_n = x) \leqslant r^{-n}$ where $r = d/(2\sqrt{d-1})$.

This result is due to Kesten [65, Th. 3]. Note that $r > 1$ as soon as $d \geqslant 3$, but $r = 1$ for $d = 2$ (which means that the proposition is trivial in that case). There is a very interesting probabilistic proof, for which we refer to the book of Woess [119, I.1.D]. We will give a more direct analytic argument.

Proof. Let $T = T_d$, and denote by x_0 its root vertex. Since the Markov operator is self-adjoint, its spectral radius coincides with its norm. We first show that the norm of M is $\leqslant \varrho$. It is enough to prove that

$$\langle M\varphi, \varphi\rangle_T \leqslant \varrho \|\varphi\|^2$$

for any $\varphi \in L^2(T, \nu_T)$. Since T is a simple graph, by (3.21), we have

$$\langle M\varphi, \varphi\rangle_T = \sum_{x\sim y} \varphi(x)\overline{\varphi(y)},$$

hence

$$\langle M\varphi, \varphi\rangle_T \leqslant \sum_{x\sim y} |\varphi(x)\varphi(y)|.$$

We use the notation $x > y$ to say that the distance of x to the root is larger than the distance of y to the root. We can then re-order the sum to write

$$\langle M\varphi, \varphi\rangle_T \leqslant 2 \sum_{\substack{x\sim y\\ x<y}} |\varphi(x)\varphi(y)|.$$

Now, for a suitable real parameter $\alpha > 0$ to be fixed later, and $x \sim y$ with $x < y$, we use the bound

$$2|\varphi(x)\varphi(y)| \leqslant \alpha^{-1}|\varphi(x)|^2 + \alpha|\varphi(y)|^2$$

(an easy form of the Cauchy-Schwarz inequality; the idea is to select α so that this inequality is as close as possible to an equality, which means that $\alpha|\varphi(x)|^2$ is as close as possible to $\alpha^{-1}|\varphi(y)|^2$, something that cannot be true with $\alpha = 1$, as the function must belong to $L^2(T)$). Hence

$$\langle M\varphi, \varphi\rangle_T \leqslant \alpha^{-1} \sum_{\substack{x\sim y\\ x<y}} |\varphi(x)|^2 + \alpha \sum_{\substack{x\sim y\\ x<y}} |\varphi(y)|^2.$$

Fix x in the first sum. If $x = x_0$ is the root, then there are d vertices y joined to x_0 at distance 1, but if $x \neq x_0$, there are only $d-1$ vertices joined to x at distance $d_T(x, x_0)+1$ of x_0. So the first sum is

$$(d-1)\alpha^{-1} \sum_x |\varphi(x)|^2 + \alpha^{-1}|\varphi(x_0)|^2.$$

In the second sum, on the other hand, if $y = x_0$, there is no vertex x with $x < x_0$, and if $y \neq x_0$, there is a unique x joined to y and closer to the root. So this sum is

$$\alpha \sum_{y\neq x_0} |\varphi(y)|^2.$$

We therefore get the bound

$$\langle M\varphi, \varphi\rangle_T \leqslant \frac{1}{d}\Big(\frac{d-1}{\alpha} + \alpha\Big)\|\varphi\|^2 + (\alpha^{-1} - \alpha)|\varphi(x_0)|^2$$

(recall that

$$\|\varphi\|^2 = d \sum_x |\varphi(x)|^2$$

by (3.11)). Taking $\alpha = \sqrt{d-1}$, so that $\alpha^{-1} - \alpha \leqslant 0$, it follows that $\langle M\varphi, \varphi\rangle \leqslant \varrho\|\varphi\|^2$, as claimed.

Now we derive the lower-bound for the norm of M. Let $n \geqslant 1$ be an integer. Define φ_n such that $\varphi_n(x) = 0$ if $d_T(x, x_0) > n$, and such that φ_n is constant and satisfies

$$\sum_{d_T(x,x_0)=k} |\varphi_n(x)|^2 = 1$$

on each sphere $d_T(x, x_0) = k$ with $0 \leqslant k \leqslant n$. It is then elementary that

$$\frac{\langle M\varphi_n, \varphi_n\rangle_T}{\|\varphi_n\|^2} \to \varrho$$

as $n \to +\infty$. This shows that the norm of M is $\geqslant \varrho$.

Finally, by Corollary 3.2.17, we have

$$\mathbf{P}(X_n = x) = \mathbf{E}(\delta_x(X_n)) = \mathbf{E}((M^n\delta_x)(X_0)) = (M^n\delta_x)(1)$$

where δ_x is the characteristic function of x. Since the inner product is not normalized, we have $\varphi(1) = \langle \varphi, \delta_1\rangle$ for any $\varphi \in L^2(T, \nu_T)$. Then

$$(M^n\delta_x)(1) = \langle M^n\delta_x, \delta_1\rangle \leqslant \|M\|^2\|\delta_x\|\|\delta_1\| = \varrho^n,$$

since $\|\delta_x\| = \|\delta_1\| = 1$. □

Exercise 3.2.32. This exercise explains the probabilistic argument to estimate the probability $\mathbf{P}(X_n = x_0)$ (see also [119, I.1.D]). For any vertex x of $T = T_d$, we denote by $(X_n^{(x)})$ the random walk on T starting at x, i.e, with the initial condition $X_0^{(x)} = x$. Define the stopping times

$$\tau_{x,y} = \min\{n \geqslant 0 \mid X_n^{(x)} = y\}$$
$$\tau_x^+ = \min\{n \geqslant 1 \mid X_n^{(x)} = x\}$$

for $y \in T$. So $\tau_{x,y}$ is the "time" when the walk from x reaches y for the first time (or $+\infty$ if it doesn't) and τ_x^+ is the time when the walk from x *comes back* to x for the first time.

Define the generating functions

$$G_{x,y}(z) = \sum_{n\geqslant 0} \mathbf{P}(X_n^{(x)} = y)z^n$$
$$F_{x,y}(z) = \mathbf{E}(z^{\tau_{x,y}}) = \sum_{n\geqslant 0} \mathbf{P}(\tau_{x,y} = n)z^n$$
$$U_x(z) = \mathbf{E}(z^{\tau_x^+}) = \sum_{n\geqslant 0} \mathbf{P}(\tau_x^+ = n)z^n.$$

(1) Show that

$$G_{x,x} = 1 + G_{x,x}U_x, \qquad G_{x,y} = F_{x,y}G_{y,y}$$
$$U_x(z) = \sum_{y\in T_k} \mathbf{P}(X_1^{(x)} = y)zF_{y,x}(z), \quad F_{x,y}(z) = \sum_{v\in T_k} \mathbf{P}(X_1^{(x)} = v)zF_{v,y}(z),$$

where the last relation holds when $x \neq y$.

(2) Show that there exists a power series F such that $F_{x,y} = F^{d_T(x,y)}$ for all x and y.

(3) Show that

$$F(z) = \frac{z}{d} + \Big(1 - \frac{1}{d}\Big)zF(z)^2$$

(4) Prove that

$$F(z) = \frac{d - \sqrt{d^2 - 4(d-1)z^2}}{2(d-1)z},$$

and that

$$U_x(z) = zF(z), \quad G_{x,x}(z) = G_{x_0,x_0}(z) = \frac{1}{1 - U_{x_0}(z)} = \frac{1}{1 - zF(z)}$$
$$G_{x,y} = G_{x_0,x_0}F^{d(x,y)}.$$

(5) Conclude using G_{x_0,x_0} that $\mathbf{P}(X_n = x_0) \leqslant \varrho^n$ for all n where $\varrho = 2\sqrt{d-1}/d$.

Exercise 3.2.33. This exercise continues the previous one, and describes Kesten's direct combinatorial computation of $\mathbf{P}(\tau_1^+ = n)$ for the random walk on $T = T_d$. This gives the formula for the function $U_{x_0}(z)$, hence also the formula for $G_{x_0,x_0}(z) = 1/(1 - U_{x_0}(z))$. Let 1 denote the root vertex of T_d and let $p_n = \mathbf{P}(\tau_1^+ = n)$.

(1) Show that $p_n = 0$ if n is odd and that

$$p_n = \frac{1}{d^n}|W_n^+|$$

where W_n^+ is the set of paths of length n in T_d, starting and ending at 1, and not passing through 1 otherwise.

(2) Let $n = 2m$ be even and let W_n be the set of all paths of length n starting and ending at 1 in T_d. To any $\gamma = (1, v_1, \ldots, v_{n-1}, v_n)$ in W_n with $v_n = 1$, we associate a continuous path $\kappa(\gamma)$ in $\mathbf{R}^2$ as follows. Starting from $(0, 0)$, we add (concatenate) an horizontal or vertical line segment of length 1 inductively for $0 \leqslant k \leqslant n-1$, joining the

point (x_k, y_k) at step k to (x_k+1, y_k) (horizontal step) if $d(1, v_{k+1}) = d(1, v_k) + 1$ and joining (x_k, y_k) to (x_k, y_k+1) (vertical step) if $d(1, v_{k+1}) = d(1, v_k) - 1$.

Show that $\kappa(\gamma)$ joins $(0, 0)$ to (m, m), and that $\gamma \in W_n^+$ if and only if $\kappa(\gamma)$ intersects the diagonal $x = y$ in $\mathbf{R}^2$ only at the points $(0, 0)$ and (m, m).

(3) Let C_n be the number of paths $\kappa(\gamma)$ as γ varies over W_n^+. Show that $|C_n| = n^{-1}\binom{2n-2}{n-1}$ (a Catalan number).

(4) Let $\gamma \in W_n^+$. Prove that the number of paths $\gamma' \in W_n^+$ with the same image $\kappa(\gamma)$ is equal to $d(d-1)^{n-1}$.

(5) With notation as in the previous exercise, deduce that

$$U_{x_0}(z) = \sum_{n \geqslant 0} p_n z^n = \frac{d - \sqrt{d^2 - 4(d-1)z^2}}{2(d-1)}.$$

Exercise 3.2.34. Let $r \geqslant 1$ be an integer and $\Gamma = \mathcal{C}(\mathbf{Z}^r, S)$, where $S = \{\pm e_i\}$ with (e_i) the canonical basis of $\mathbf{Z}^r$. Show that the spectral radius of the corresponding Markov operator is equal to 1.

The fact that the norm of the Markov operator is 1 for the infinite graph of the last exercise contrasts strongly with the result for a d-regular tree (with $d \geqslant 3$) of Proposition 3.2.31. This difference reflects a fundamental dichotomy in the theory and geometry of discrete group: finitely generated groups whose Cayley graphs, with respect to finite symmetric generating sets, have Markov operator with norm 1, like $\mathbf{Z}^r$, are called *amenable groups* (although they have many other equivalent definitions, which might be taken as the starting point instead of that characterization). Their properties are in many respect very different from those of groups where the norm is < 1, such as non-abelian free groups. The interested reader can find the basic steps in the study of amenable groups in Chapters 4 and 6 of the book of Ceccherini-Silberstein and Coornaert [25] (especially [25, Th. 6.12.9] contains the proof that the characterization in terms of spectral radius of amenable groups corresponds to the other definitions in [25, Ch. 4]). Further references can also be found there or in [78, §2.2].

We will use the following result in Section 3.6.

Lemma 3.2.35. *Let $d \geqslant 2$ and let T be a full subgraph of the infinite d-regular tree T_d. The norm of the adjacency operator A_T is $\leqslant 2\sqrt{d-1}$.*

Proof. Consider the subspace $E \subset L^2(T_d, \nu_{T_d})$ of functions with support in the set of vertices of T. Let p be the orthogonal projection from $L^2(T_d, \nu_{T_d})$ to E, which is the restriction of functions to T. The adjoint p^* of p is the inclusion map of E in $L^2(T_d, \nu_{T_d})$.

As a vector space, we can identify $L^2(T, \nu_T)$ with E by the isomorphism that extends a function $\varphi \in L^2(T, \nu_T)$ by zero outside T in T_d. Under this identification, we have $A_T = p \circ A_{T_d} \circ p^*$ (indeed, for $\varphi \in L^2(T, \nu_T)$, and $x \in T$, we have

$$A_T\varphi(x) = \sum_{y \sim_T x} a_T(x, y)\varphi(y) = \sum_{y \sim_{T_d}} a_{T_d}(x, y)\varphi(y) = \sum_{y \sim_{T_d} x} a_{T_d}(x, y)\varphi(y),$$

since T is a full subgraph of T_d and φ is zero outside T, which implies the desired statement).

Since the norm of p and p^* is $\leqslant 1$, the norm of A_T, as an endomorphism of E, is at most that of A_{T_d}. Since T_d is d-regular, the norm of A_{T_d} is $d\|M_{T_d}\| \leqslant 2\sqrt{d-1}$ by Kesten's result. □

3.3. Random walks and expansion

As promised, we will now describe the precise link between the expansion constant $h(\Gamma)$ and the equidistribution radius ϱ_Γ of a finite graph. As we mentioned already, there is a technical point to address. Indeed, being or not bipartite (or "very close", in the sense that there is an eigenvalue of the Markov operator M that is very close to -1) is a property essentially unrelated to being an expander, but it affects the rate of equidistribution. To make this clear, we make the following definition:

Definition 3.3.1 ("Absolute Expanders"). Let (Γ_i) be a family of finite, non-empty, connected graphs $\Gamma_i = (V_i, E_i, \mathrm{ep})$ with maximal valency $\leqslant v$ for all i, such that the number of vertices of Γ_i tends to infinity, in the same sense as in Definition 3.1.8. We say that (Γ_i) is a family of *absolute expanders* if and only if there exists $\varrho < 1$ such that

$$\varrho_{\Gamma_i} \leqslant \varrho < 1 \tag{3.29}$$

for all $i \in I$. When this is true, we say that (ϱ, v) are *equidistribution parameters* for the absolute expander family.

The precise link between expanders and absolute expanders is the content of the following result:

Theorem 3.3.2 (Random walk definition of expanders). (1) *A family of absolute expanders is an expander family.*

(2) *Conversely, let (Γ_i) be an expander family with $\Gamma_i = (V_i, E_i, \mathrm{ep})$. Let $\widetilde{\Gamma}_i$ be the "relaxed" graphs obtained from Γ_i by adding a loop at each vertex, i.e.,*

$$\widetilde{\Gamma}_i = (V_i, E_i \cup V_i, \mathrm{ep}')$$

with $\mathrm{ep}'(\alpha) = \mathrm{ep}(\alpha)$ for $\alpha \in E_i$ and $\mathrm{ep}'(x) = \{x\}$ for $x \in V_i$. Then $(\widetilde{\Gamma}_i)$ is a family of absolute expanders.

Remark 3.3.3. Since the vertices do not change, and only loops are added to the edges of the relaxed graphs, which has no effect on the value of $\mathcal{E}(W_1, W_2)$ for any subsets W_1, $W_2 \subset V$, we have $h(\widetilde{\Gamma}_i) = h(\Gamma_i)$.

Moreover, we only add one loop for each vertex, so that the maximal valency of the relaxed graphs has only been increased by 1. In particular, we see that (Γ_i) is an expander family *if and only if* $(\widetilde{\Gamma}_i)$ is an expander family. On the other hand, because we added loops, $\widetilde{\Gamma}_i$ is not bipartite, and hence -1 is not an eigenvalue of M. In fact, having added loops to all vertices allows us quite easily to show that there is no eigenvalue of M too close to -1, and this explains why the relaxed family has better equidistribution properties.

In fact, more is true: there are quantitative two-sided inequalities relating $h(\Gamma)$ and ϱ_Γ, from which the statement will immediately follow with relations between the expansion and equidistribution parameters. It will also be possible to see that, in general, the full converse of (1) is not true. However, there are many families of expanders which are absolute expanders without any addition of loops being needed.

By definition, ϱ_Γ is either the largest eigenvalue < 1 of M, or the negative of the smallest eigenvalue which is > -1. A convenient way to express this is to give names to the distance of the largest and smallest eigenvalues to 1 and -1.

Definition 3.3.4 (Normalized spectral gaps). Let Γ be a finite non-empty connected graph. The *normalized spectral gap* $\boldsymbol{\lambda}_1(\Gamma)$ is the smallest non-zero eigenvalue of $\mathrm{Id} - M$.

The *complementary normalized spectral gap* $\boldsymbol{\mu}_1(\Gamma)$ is the smallest non-zero eigenvalue of $\mathrm{Id} + M$.

The largest eigenvalue < 1 of M is therefore $1 - \boldsymbol{\lambda}_1$, and the smallest > -1 is $-1 + \boldsymbol{\mu}_1$. Thus we have

$$\varrho_\Gamma = \max(1 - \boldsymbol{\lambda}_1(\Gamma), \boldsymbol{\mu}_1(\Gamma) - 1).$$

Moreover we have

$$\boldsymbol{\lambda}_1(\Gamma) = \min_{0 \neq \varphi \perp 1} \frac{\langle (\mathrm{Id} - M)\varphi, \varphi\rangle}{\langle \varphi, \varphi\rangle} \tag{3.30}$$

$$= \min_{\varphi \text{ not constant}} \frac{\langle (\mathrm{Id} - M)\varphi, \varphi\rangle}{\|\varphi - \langle \varphi, \varphi\rangle\|^2}, \tag{3.31}$$

where the equality between these two characterizations follows from the fact that

$$\langle (\mathrm{Id} - M)\varphi, \varphi\rangle = \langle (\mathrm{Id} - M)\varphi_0, \varphi_0\rangle$$

for $\varphi_0 = \varphi - \langle \varphi, 1\rangle$, which is orthogonal to 1, so that the range of values in the minimum in the second definition is in fact identical to the one in the first.

The link between $h(\Gamma)$ and equidistribution becomes visible here. First by comparing with the definition of the expansion constant, also as a minimum, and then by using (3.22) which shows that the numerator is determined by the difference in values of φ on adjacent vertices, so that suitable choices of φ lead to the quantity $\mathcal{E}(W)$, as the following lemma shows:

Lemma 3.3.5. *Let Γ be a finite non-empty graph without isolated vertices. Let $W \subset V$ be a subset of vertices, $W' = V - W$, and let*

$$\varphi = \mathbf{1}_W - \mu_\Gamma(W),$$

the "centered" characteristic function of W. Then

$$\langle (\mathrm{Id} - M)\varphi, \varphi\rangle = \langle (\mathrm{Id} - M)\mathbf{1}_W, \mathbf{1}_W\rangle = \frac{|\mathcal{E}(W)|}{N}$$

and $\|\varphi\|^2 = \mu_\Gamma(W)\mu_\Gamma(W')$.

Proof. The formula (3.22) gives

$$\langle (\mathrm{Id} - M)\varphi, \varphi\rangle = \frac{1}{2N} \sum_{x,y \in V} a(x,y)(\varphi(x) - \varphi(y))^2$$

hence

$$\langle (\mathrm{Id} - M)\varphi, \varphi\rangle = \frac{1}{2N} \sum_{x,y \in V} a(x,y)(\mathbf{1}_W(x) - \mathbf{1}_W(y))^2.$$

The only non-zero terms in this sum are those where, on the one hand, x and y are adjacent, and on the other hand, one of them is in W and the other is not. The two cases $x \in W$, $y \notin W$ and $x \notin W$, $y \in W$ have equal contribution, and hence

$$\langle (\mathrm{Id} - M)\varphi, \varphi\rangle = \frac{1}{N} \sum_{\substack{x \in W \\ y \notin W}} a(x,y) = \frac{|\mathcal{E}(W)|}{N}.$$

The formula for $\|\varphi\|^2$ is a simple computation: since φ is orthogonal to constants, we have

$$\|\varphi\|^2 = \|\mathbf{1}_W\|^2 - \mu_\Gamma(W)^2 = \mu_\Gamma(W) - \mu_\Gamma(W)^2 = \mu_\Gamma(W)\mu_\Gamma(W').$$

□

We can now immediately prove (1) in Theorem 3.3.2. Indeed, it follows from the next proposition, which is the analogue for graphs of the Cheeger inequality for manifolds [28] (see also Section 5.4):

Proposition 3.3.6 (Expansion and equidistribution; discrete Cheeger inequality). *Let $\Gamma = (V, E, \mathrm{ep})$ be a connected, non-empty, finite graph without isolated vertices. We have*

$$1 - \varrho_\Gamma \leqslant \boldsymbol{\lambda}_1(\Gamma) \leqslant \Bigl(\frac{2v_+}{v_-^2}\Bigr) h(\Gamma) \tag{3.32}$$

where, as before, we denote

$$v_- = \min_{x \in V} \mathrm{val}(x), \qquad v_+ = \max_{x \in V} \mathrm{val}(x).$$

In particular, if Γ is d-regular, then we have

$$1 - \varrho_\Gamma \leqslant \boldsymbol{\lambda}_1(\Gamma) \leqslant \frac{2}{d} h(\Gamma).$$

Proof. Because of (3.30), we can estimate $\boldsymbol{\lambda}_1(\Gamma)$ from above by

$$\boldsymbol{\lambda}_1(\Gamma) \leqslant \frac{\langle (\mathrm{Id} - M)\varphi, \varphi\rangle}{\langle \varphi, \varphi\rangle}$$

for any suitable function φ orthogonal to 1. Applying Lemma 3.3.5 to a non-empty subset $W \subset V$ with $|W| \leqslant |\Gamma|/2$ such that $h(\Gamma) = |\mathcal{E}(W)|/|W|$, we get

$$\boldsymbol{\lambda}_1(\Gamma) \leqslant \frac{|\mathcal{E}(W)|}{N} \frac{1}{\|\varphi\|^2} = \frac{1}{N} \frac{|\mathcal{E}(W)|}{\mu_\Gamma(W)\mu_\Gamma(W')}.$$

We now use (3.14) in order to make the exact ratio $|\mathcal{E}(W)|/|W|$ appear, obtaining

$$N\mu_\Gamma(W)\mu_\Gamma(W') \geqslant v_-|W| \times \frac{v_-}{v_+}\frac{|W'|}{|V|} \geqslant \frac{v_-^2}{2v_+}|W|,$$

and the inequality (3.32) follows. □

Remark 3.3.7. The Cheeger inequality is very often the best way to obtain lower bounds for the expansion constant of a graph (for instance, it will be used in three of the four constructions of expander families in Chapter 4). It is also useful numerically: since $\boldsymbol{\lambda}_1(\Gamma)$ is an eigenvalue of the linear operator $\mathrm{Id} - M$ acting on $L^2(\Gamma)$, which is a finite-dimensional vector space, of dimension $|V|$, the problem of determining $\boldsymbol{\lambda}_1(\Gamma)$ (or indeed ϱ_Γ itself) is a problem *of linear algebra*. Of course, if V has enormous size, it might not be feasible to find all eigenvalues, but the fact that ϱ_Γ is the largest absolute value of any eigenvalue on $L^2_0(\Gamma, \mu_\Gamma)$ also leads to the possibility of applying various approximation algorithms for this specific problem.

We will now investigate the converse of (3.32). We may note already that it can not be a simple relation stating that $\boldsymbol{\lambda}_1$ (or $1 - \varrho$) is of the same order of magnitude as the expansion constant up to constant factors, since for the cycles, we have found in (3.2) that $h(C_m) \asymp 1/m$ for m large, while $1 - \varrho_{C_m} \asymp 1/m^2$ by Example 3.2.26 (2), which is much smaller. However, this is essentially as bad as it can get, as shown by the following bound, which is the discrete analogue of an inequality of Buser in the context of the geometric Cheeger constant [21]:

Proposition 3.3.8 (Discrete Buser inequality). *Let $\Gamma = (V, E, \mathrm{ep})$ be a connected, non-empty, finite graph without isolated vertices. We have*

$$h(\Gamma) \leqslant v_+\sqrt{2\,\boldsymbol{\lambda}_1(\Gamma)}. \tag{3.33}$$

We will prove this by following an argument of L. Trevisan [112, Handout 4], which highlights a practical algorithmic interpretation of this inequality. The idea is to study the expansion of sets of the type

$$W_{\varphi,t} = \varphi^{-1}(]-\infty, t]) = \{x \in V \mid \varphi(x) \leqslant t\}$$

for a real-valued function $\varphi : V \longrightarrow \mathbf{R}$ and a real number t, and to show that some of them satisfy

$$\frac{|\mathcal{E}(W_{\varphi,t})|}{|W_{\varphi,t}|} \leqslant v_+ \sqrt{2\,\boldsymbol{\lambda}_1(\Gamma)},$$

while containing at most $|V|/2$ vertices. The idea, to begin with, is to compute the average (over t) of the size of the sets $\mathcal{E}(W_{\varphi,t})$ for a given function, and deduce the existence of sets with certain expansion ratio. The following lemma performs this computation:

Lemma 3.3.9 (Expansion of sublevel sets). *Let $\Gamma = (V, E, \mathrm{ep})$ be a finite non-empty connected graph and let $\varphi : V \longrightarrow \mathbf{R}$ be a real-valued non-constant function on V. Let*

$$a = \min_{x\in V} \varphi(x), \qquad b = \max_{x \in V} \varphi(x),$$

and let $t_0 \in \mathbf{R}$ be such that[2]

$$|W_{\varphi,t}| \leqslant \frac{|V|}{2}$$

if and only if $t < t_0$.

Then for any choice of a probability measure ν on $\mathbf{R}$ supported on $[a,b]$ and without atoms, we can find $t \in \mathbf{R}$ such that either $W = W_{\varphi,t}$ or $W = V - W_{\varphi,t}$ satisfies $|W| \leqslant |V|/2$ and

$$\frac{|\mathcal{E}(W)|}{|W|} \leqslant \frac{A}{B}$$

where

$$A = \frac{1}{2} \sum_{x,y \in V} a(x,y)\nu([\varphi(x), \varphi(y)]),$$

$$B = \sum_{x \in V} \nu([t_0, \varphi(x)])$$

using the convention that $\nu([a,b]) = \nu([\min(a,b), \max(a,b)])$.

Proof. We denote $W_t = W_{\varphi,t}$ for simplicity. An edge α with $\mathrm{ep}(\alpha) = \{x,y\}$ is in $\mathcal{E}(W_t)$ if and only if t lies in the interval I_α between $\varphi(x)$ and $\varphi(y)$ where the largest is excluded, i.e., $I_\alpha = [\min(\varphi(x), \varphi(y)), \max(\varphi(x), \varphi(y)[$. Thus we may compute the average of $|\mathcal{E}(W_t)|$ as

$$\begin{aligned}\int_{\mathbf{R}} |\mathcal{E}(W_t)| d\nu(t) &= \sum_{\alpha \in E} \nu\{t \mid t \text{ is in the interval } I_\alpha\} \\ &= \sum_{\alpha \in E} \nu(I_\alpha) = \frac{1}{2} \sum_{x,y \in V} a(x,y)\nu([\varphi(x), \varphi(y)]) = A,\end{aligned}$$

since ν has no atom.

We want to compare this with the number of elements of W_t, or rather with the minimum $\min(|W_t|, |V - W_t|) \leqslant |V|/2$ (with the idea of using either W_t or $V - W_t$ to test the expansion constant).

[2] This means that t_0 is a "median" of the values of φ.

Since the size of W_t is non-decreasing as a function of t, a real number t_0 such that $|W_t| \leqslant |V|/2$ if and only if $t < t_0$ exists. Then (again using the fact that ν has no atoms) we have

$$\begin{aligned}\int_{\mathbf{R}} \min(|W_t|, |V - W_t|) d\nu(t) &= \int_{t<t_0} |W_t| d\nu(t) + \int_{t\geqslant t_0} |V - W_t| d\nu(t)\\ &= \sum_{x\in V} \nu\{t \mid \varphi(x) \leqslant t < t_0\} + \sum_{x\in V} \nu\{t \mid t_0 \leqslant t \leqslant \varphi(x)\}\\ &= \sum_{x\in V} \nu([t_0, \varphi(x)]) = B.\end{aligned}$$

We now argue simply that since

$$\int_{\mathbf{R}} \Big(B|\mathcal{E}(W_t)| - A\min(|W_t|, |V - W_t|)\Big) d\nu(t) = 0,$$

there must exist some $t \in [a, b]$ for which

$$B|\mathcal{E}(W_t)| - A\min(|W_t|, |V - W_t|) \leqslant 0,$$

which is the desired conclusion! □

We are now led to an attempt to select a measure ν and then find a function φ to minimize the ratio A/B. The most natural-looking choice seems to be the uniform probability measure on $[a, b]$, with $d\nu(t) = dt/(b - a)$. In this case, we get

$$A = \frac{1}{2}\sum_{x,y\in V} a(x, y)\frac{|\varphi(x) - \varphi(y)|}{b - a}, \qquad B = \sum_{x\in V} \frac{|\varphi(x) - t_0|}{b - a}, \tag{3.34}$$

and the problem looks similar, in a rather more L^1-ish sense, to the computation of $\boldsymbol{\lambda}_1$ using the minimization characterization (3.30). However, because the L^1-norm is much less flexible and accessible than the L^2-norm, this does not seem easy to work out (as mentioned by Trevisan [113]; see Example 3.3.10 below for an instance of this, but also Proposition 3.5.8 for a case where this is sharper than Buser's inequality (3.33)). So we use instead, as in [112], the measure ν defined by

$$d\nu(t) = \frac{1}{S}|t - t_0| dt,$$

where S is the normalizing factor that makes this a probability measure on $[a, b]$. We have then

$$\nu([t_0, \varphi(x)]) = \frac{1}{2S}|\varphi(x) - t_0|^2$$

for all x and a second's thought shows that

$$\nu([\varphi(x), \varphi(y)]) \leqslant \frac{1}{2S}|\varphi(x) - \varphi(y)| \times (|\varphi(x) - t_0| + |\varphi(y) - t_0|).$$

Hence we find in this way a set W for which

$$h(\Gamma) \leqslant \frac{|\mathcal{E}(W)|}{|W|} \leqslant \frac{\widetilde{A}}{\widetilde{B}}$$

where

$$\widetilde{A}=\frac{1}{2}\sum_{x,y\in V}a(x,y)\{|\varphi(x)-t_0|+|\varphi(y)-t_0|\}|\varphi(x)-\varphi(y)|,$$

$$\widetilde{B}=\sum_{x\in V}|\varphi(x)-t_0|^2.$$

We can now estimate further in terms of quantities related to M. First, we write

$$\widetilde{B}=\sum_{x\in V}|\varphi(x)-t_0|^2\geqslant\frac{1}{v_+}\sum_{x\in V}\mathrm{val}(x)|\varphi(x)-t_0|^2=\frac{N}{v_+}\|\varphi-t_0\|^2$$

while, by the Cauchy-Schwarz inequality and the formulas (3.22) and (3.23), we have

$$\begin{aligned}(\widetilde{A})^2&\leqslant\Big(\frac{1}{2}\sum_{x,y}a(x,y)|\varphi(x)-\varphi(y)|^2\Big)\Big(\frac{1}{2}\sum_{x,y}a(x,y)\{|\varphi(x)-t_0|+|\varphi(y)-t_0|\}^2\Big)\\&=N\langle(\mathrm{Id}-M)\varphi,\varphi\rangle\times N\langle(\mathrm{Id}+M)|\varphi-t_0|,|\varphi-t_0|\rangle.\end{aligned}$$

Since $\|\mathrm{Id}+M\|\leqslant 2$, we obtain

$$\frac{\widetilde{A}}{\widetilde{B}}\leqslant v_+\Big(\frac{2\langle(\mathrm{Id}-M)\varphi,\varphi\rangle}{\|\varphi-t_0\|^2}\Big)^{1/2}.$$

We finally select φ to be an eigenfunction of $\mathrm{Id}-M$ with eigenvalue λ_1. Since φ is orthogonal to the constants, it is the orthogonal projection of $\varphi-t_0$ to the orthogonal complement of the constants, so $\|\varphi-t_0\|\geqslant\|\varphi\|$, and we get the inequality

$$h(\Gamma)\leqslant v_+\sqrt{2\lambda_1}$$

(note that there always exists a real-valued eigenfunction of $\mathrm{Id}-M$, since the real and imaginary parts of an eigenfunction φ are still eigenfunctions with the same eigenvalue, and one at least must be non-zero if $\varphi\neq 0$...) This finishes the proof of the discrete Buser inequality.

Example 3.3.10 (The cycles again). Let $\Gamma=C_m$ with $m\geqslant 2$. In Example 3.2.26 (2), we have shown that $\boldsymbol{\lambda}_1(C_m)=1-\cos(2\pi/m)\sim(2\pi^2)/m^2$ as $m\to+\infty$. A real-valued $\boldsymbol{\lambda}_1$-eigenfunction is given by

$$\varphi(x)=\mathrm{Re}\Big(e\Big(\frac{x}{m}\Big)\Big)=\cos\Big(\frac{2\pi x}{m}\Big)$$

for $x\in\mathbf{Z}/m\mathbf{Z}$. It follows that $W_0=\{x\mid\varphi(x)\leqslant 0\}$ is roughly the image modulo m of the integers between $m/4$ and $3m/4$, which we used in Example 3.1.3 to lead to the expansion constant $h(C_m)\sim 4/m$.

If we assume that m is even for simplicity, the median is $t_0=0$, and the application of Lemma 3.3.9 for this function, with the uniform probability measure, shows that the existence of some set W with

$$h(C_m)\leqslant\frac{|\mathcal{E}(W)|}{|W|}\leqslant\frac{A}{B}$$

where, spelling out (3.34), we have

$$A=\sum_{0\leqslant x\leqslant m-1}\Big|\cos\Big(\frac{2\pi x}{m}\Big)-\cos\Big(\frac{2\pi(x+1)}{m}\Big)\Big|,$$

$$B=\sum_{0\leqslant x\leqslant m-1}\Big|\cos\Big(\frac{2\pi x}{m}\Big)\Big|.$$

It is elementary (looking at the graph of the cosine) that A converges to 4 as m tends to infinity, while $B \sim \frac{2}{\pi}m$. Thus the bound $h(C_m) \leqslant A/B \sim 2\pi/m$ is of the right order of magnitude in that case.

We can now also conclude the proof of part (2) in Theorem 3.3.2. Given a family (Γ_i) of expanders, we see from the discrete Buser inequality that the relaxed graph satisfy

$$v\sqrt{2\,\boldsymbol{\lambda}_1(\widetilde{\Gamma}_i)} \geqslant h(\widetilde{\Gamma}_i) = h(\Gamma_i).$$

This shows that the normalized spectral gap is bounded away from zero. Hence it is now enough to prove that $\widetilde{\Gamma}_i$ can not have an eigenvalue too close to -1. But the definition of $\widetilde{\Gamma}_i$ with its added loops leads to the formula

$$\begin{aligned}\langle(\mathrm{Id} + \widetilde{M}_i)\varphi, \varphi\rangle &= \frac{1}{2\widetilde{N}_i}\sum_{x,y\in V_i} \widetilde{a}(x,y)|\varphi(x)+\varphi(y)|^2\\ &= \frac{1}{2\widetilde{N}_i}\Bigl(\sum_{x,y\in V_i} a(x,y)|\varphi(x)+\varphi(y)|^2 + 4\sum_{x\in V_i}|\varphi(x)|^2\Bigr)\end{aligned}$$

and since $\widetilde{N}_i = N_i + |V_i| \leqslant 2N_i$ and $\mathrm{val}(x) \leqslant v_+ \leqslant v$, we get by positivity

$$\langle(\mathrm{Id} + \widetilde{M}_i)\varphi, \varphi\rangle \geqslant \frac{1}{N}\sum_{x\in V_i}|\varphi(x)|^2 \geqslant \frac{1}{v}\|\varphi\|^2,$$

which implies that $\widetilde{M}_i$ has no eigenvalue $< -1 + v^{-1}$. Hence we derive

$$\varrho_{\widetilde{\Gamma}_i} \leqslant 1 - \min\Bigl(\frac{h^2}{2v}, \frac{1}{v}\Bigr) < 1$$

for all i, giving equidistribution parameters of the relaxed graphs in terms of the expansion parameters (h, v) of (Γ_i). (Typically, $h^2/2$ is less than 1, of course, so we can replace this expression by $1 - h^2/(2v)$.)

Example 3.3.11 (Expanders, but not absolute expanders). It is clear that there is a large extent of flexibility in adding loops here and there to expanders in order to obtain absolute expanders. However, too little would not be enough. Indeed, assuming some results on the existence of expanders (which will follow from the later sections), we can give some easy examples of families of graphs which are expanders, but not absolute expanders.

For this, we start with any sequence (Γ_n) of *bipartite* expanders (whose existence follows from Chapter 4, either from Theorem 4.1.1 in Section 4.1, or Theorem 4.2.5 in Section 4.2 or Theorem 4.3.1 in Section 4.3, or Corollary 4.3.4...). Then, for each n, we attach a *single* loop at some (arbitrarily chosen) vertex x_n of Γ_n, obtaining a new sequence (Γ'_n) of non-bipartite graphs. Since attaching loops does not change the expansion constant (and attaching a single loop barely increases the maximal valency!), this family is still a family of expanders. Intuitively, adding this puny loop should not change the equidistribution constant very much, and it is easy to find a lower bound using an upper-bound for $\boldsymbol{\mu}_1(\Gamma_n)$ and its characterization

$$\boldsymbol{\mu}_1(\Gamma) = \min_{0\neq\varphi\in L^2(\Gamma)} \frac{\langle(\mathrm{Id} + M)\varphi, \varphi\rangle}{\langle\varphi, \varphi\rangle}$$

for a non-bipartite graph, so there is no condition required about φ, in the absence of an eigenfunction of eigenvalue -1 of M.

For Γ_n, the function $\varepsilon_\pm$ defined in Proposition 3.2.18 minimizes this expression. For Γ'_n, it is natural enough to expect that it will also be close to the minimum. We have, with obvious notation, $N'_n = N_n + 1$ and

$$\langle (\mathrm{Id} + M'_n)\varphi, \varphi\rangle = \frac{1}{2N'_n} \sum_{x,y\in V_n} a(x,y)(\varepsilon_\pm(x) + \varepsilon_\pm(y))^2,$$

which only differs from the corresponding quantity for Γ_n by having a non-zero term for $x = y = x_n$, which is equal to $\frac{2}{N'_n}$. Since $\varepsilon_\pm$ is in the kernel of $\mathrm{Id} + M_n$, this gives

$$\langle (\mathrm{Id} + M'_n)\varepsilon_\pm, \varepsilon_\pm\rangle = \frac{2}{N'_n} = \frac{2}{N_n + 1}.$$

On the other hand, since $|\varepsilon_\pm(x)| = 1$, the norm squared of $\varepsilon_\pm$ is still one, and we get

$$\lim_{n\to+\infty} \boldsymbol{\mu}_1(\Gamma_n) = 0$$

since $N_n \geqslant |V_n| \to +\infty$. Hence the graphs (Γ'_n) are *not* absolute expanders.

Remark 3.3.12 (Trivial lower bound). From Lemma 3.1.4 for $h(\Gamma)$ and Proposition 3.3.8, we see that there is a universal “trivial” lower bound

$$\boldsymbol{\lambda}_1(\Gamma) \geqslant \frac{1}{2v_+^2}\frac{1}{|\Gamma|^2} \tag{3.35}$$

for a finite connected graph Γ without isolated vertex. The example of the cycles C_m with $v_+ = 2$ and $\boldsymbol{\lambda}_1(C_m) \asymp m^{-2} = |C_m|^{-2}$ shows that the order of magnitude can not be improved.

Exercise 3.3.13. Let $\Gamma = (V, E)$ be a finite simple graph. The *chromatic number* χ_Γ is the smallest integer $k \geqslant 0$ such that there is a k-coloring of V where no adjacent vertices have the same color (i.e., such that there is a function $f\colon V \to \{1, \ldots, k\}$ such that $f(x) \neq f(y)$ whenever x and y are connected by an edge). The *independence number* i_Γ is the largest $k \geqslant 0$ such that there exists $Y \subset V$ with the property that elements of Y are never connected.

(1) Show that $\chi_\Gamma i_\Gamma \geqslant |\Gamma|$.

(2) If Γ is d-regular with $d \geqslant 2$, then show that $i_\Gamma \leqslant \varrho_\Gamma|\Gamma|$.

Finally, we state a result that is often very useful (and usually called the “expander mixing lemma”):

Proposition 3.3.14. *Let $\Gamma = (V, E, \mathrm{ep})$ be a finite graph with no isolated vertices. For any subsets V_1 and V_2 of V, we have*

$$\left|\frac{|\mathcal{E}(V_1, V_2)|}{N} - \mu_\Gamma(V_1)\mu_\Gamma(V_2)\right| \leqslant \tilde{\varrho}_\Gamma\sqrt{\mu_\Gamma(V_1)\mu_\Gamma(V_2)},$$

where $\tilde{\varrho}_\Gamma$ is the spectral radius of M restricted to the orthogonal of the constant functions in $L^2(\Gamma, \mu_\Gamma)$. In particular, if Γ is connected and not bipartite, we have $\tilde{\varrho}_\Gamma = \varrho_\Gamma$, and if Γ is d-regular for some $d \geqslant 2$, then we have

$$\left||\mathcal{E}(V_1, V_2)| - \frac{d|V_1||V_2|}{|V|}\right| \leqslant d\tilde{\varrho}_\Gamma\sqrt{|V_1||V_2|}.$$

Proof. For $i = 1$, 2, let φ_i be the characteristic function of V_i. Since these are real-valued, by (3.24), we have

$$\langle M\varphi_1, \varphi_2\rangle = \frac{1}{N}\sum_{x,y\in V} a(x,y)\varphi_1(y)\varphi_2(x)$$

and by definition of $a(x, y)$, this is equal to $|\mathcal{E}(V_1, V_2)|/N$.

On the other hand, we write $\varphi_i = \langle \varphi_i, 1\rangle + \varphi_{i,0}$, where $\varphi_{i,0}$ is orthogonal to the constants. By orthogonality of the eigenvectors, it follows that

$$\langle M\varphi_1, \varphi_2\rangle = \langle \varphi_1, 1\rangle\langle \varphi_2, 1\rangle + \langle M\varphi_{1,0}, \varphi_{2,0}\rangle.$$

The first term is equal to $\mu(V_1)\mu(V_2)$, whereas the second satisfies

$$|\langle M\varphi_{1,0}, \varphi_{2,0}\rangle| \leqslant \tilde{\varrho}_\Gamma \|\varphi_{1,0}\|\|\varphi_{2,0}\| \leqslant \tilde{\varrho}_\Gamma \|\varphi_1\|\|\varphi_2\| = \tilde{\varrho}_\Gamma \sqrt{\mu(V_1)\mu(V_2)}.$$

Comparing this with the first formula gives the desired statement. If Γ is d-regular, then $\mu_\Gamma(V_i) = |V_i|/|V|$ and $N = d|V|$, hence the second inequality follows. □

This result gives a good idea of the virtues of expander graphs: if $\tilde{\varrho}_\Gamma$ is relatively small, but the sets V_1 and V_2 are pretty large, then we obtain a very precise estimate on the size of $\mathcal{E}(V_1, V_2)$. The result also fits with the often-stated philosophy that an expander behaves like a random graph in many ways: indeed, if we consider a random graph with vertex set V and edges added independently between each pair of vertices, with the same probability for each edge, adjusted so that the average degree is d, then it is elementary that the expected value of $|\mathcal{E}(V_1, V_2)|$ is $d|V_1||V_2|/|V|$.

3.4. The discrete Laplace operator

In the course of Section 3.2, we have in fact seen that the spectral gap of a connected graph controls the expansion constant. This leads to a characterization of expanders using only the operator $\mathrm{Id} - M$. We introduced this operator using the random walks on a graph, but it may be defined directly without referring to these ideas. It then acquires a new name:

Definition 3.4.1. Let $\Gamma = (V, E, \mathrm{ep})$ be a finite graph. The *normalized Laplace operator* of Γ, denoted $\boldsymbol{\Delta}_\Gamma$, is the linear operator

$$\boldsymbol{\Delta}_\Gamma \left\{ \begin{array}{ccc} L^2(\Gamma) & \longrightarrow & L^2(\Gamma) \\ \varphi & \mapsto & (\mathrm{Id} - M)\varphi \end{array} \right.$$

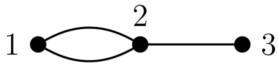
where M is the Markov operator of Γ. If Γ is d-regular for some $d \geqslant 1$, then the *Laplace operator* of Γ is defined by $\underline{\Delta}_\Gamma = d\boldsymbol{\Delta}_\Gamma$, and its spectral gap $\underline{\lambda}_1(\Gamma)$ is its smallest non-zero eigenvalue. It is equal to $d\,\boldsymbol{\lambda}_1(\Gamma_1)$.

Remark 3.4.2. In many sources (e.g., [78, §4.2]), a Laplace operator is defined for an arbitrary finite graph by $\underline{\Delta}_\Gamma = \mathrm{val} - A_\Gamma$, where val is the operator of multiplication by $x \mapsto \mathrm{val}(x)$, and A_Γ is the adjacency operator. In other words, we have

$$\underline{\Delta}_\Gamma \varphi(x) = \mathrm{val}(x)\varphi(x) - \sum_{y \in V} a(x, y)\varphi(y). \tag{3.36}$$

However, when the valency is not constant, there isn't a very clear relation between the spectrum of $\boldsymbol{\Delta}_\Gamma$ and that of $\underline{\Delta}_\Gamma$. For instance, it is not easy to translate the formula (3.22) for the combinatorial Laplace operator (altough see Exercise 3.4.5 for a similar formula). But one can still prove that the smallest non-zero eigenvalue of $\underline{\Delta}_\Gamma$ satisfies inequalities similar to that of $\mathrm{Id} - M$, taking the renormalization into account (see [78, Prop. 4.2.4, 4.2.5] and Exercise 3.4.5 for the easier one of these). For general graphs, we will use the normalized Laplace operator in this book.

Here is a simple random example. For the graph below



the matrices representing M_Γ and A_Γ in the basis of characteristic functions of single points are given, respectively, by

$$\begin{pmatrix} 0 & 2/3 & 0 \\ 1 & 0 & 1 \\ 0 & 1/3 & 0 \end{pmatrix} \qquad \begin{pmatrix} 0 & 2 & 0 \\ 2 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

and the matrices representing $\boldsymbol{\Delta}_\Gamma$ and $\underline{\Delta}_\Gamma$ are

$$\begin{pmatrix} 1 & -2/3 & 0 \\ -1 & 1 & -1 \\ 0 & -1/3 & 1 \end{pmatrix} \qquad \begin{pmatrix} 2 & -2 & 0 \\ -2 & 3 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

The characteristic polynomials of M_Γ, A_Γ, $\boldsymbol{\Delta}_\Gamma$ and $\underline{\Delta}_\Gamma$ are, respectively

$$X(X-1)(X+1), \qquad X(X^2-5), \qquad X(X-1)(X-2), \qquad X(X^2-6X+6)$$

(as products of irreducible factors over $\mathbf{Q}$).

Here is a summary of the results of the previous section in terms of the combinatorial Laplace operator, for regular graphs.

Proposition 3.4.3 (Properties of $\underline{\Delta}_\Gamma$). *Let $\Gamma = (V, E, \mathrm{ep})$ be a finite connected d-regular graph without isolated vertex.*

(1) *The Laplace operator is self-adjoint and non-negative; its kernel is one-dimensional and spanned by the constant functions. Moreover we have*

$$\langle \underline{\Delta}_\Gamma \varphi, \varphi \rangle = \frac{1}{2|V|} \sum_{x,y \in V} a(x,y) |\varphi(x) - \varphi(y)|^2$$

for all $\varphi \in L^2(\Gamma)$.

(2) *We have*

$$\underline{\lambda}_1(\Gamma) = \min_{\substack{\varphi \in L^2(\Gamma) \\ \langle \varphi, 1 \rangle = 0}} \frac{\langle \underline{\Delta}_\Gamma \varphi, \varphi \rangle}{\langle \varphi, \varphi \rangle}$$

and

$$\frac{\underline{\lambda}_1(\Gamma)}{2} \leqslant h(\Gamma) \leqslant \sqrt{2d\, \underline{\lambda}_1(\Gamma)}. \tag{3.37}$$

These are immediate consequences of the previous discussion. Similarly, we state for completeness the characterization of expander graphs in terms of $\boldsymbol{\lambda}_1(\Gamma)$ and $\underline{\lambda}_1(\Gamma)$.

Theorem 3.4.4 (Spectral definition of expanders). *Let $(\Gamma_i)_{i \in I}$ be a family of connected finite graphs with $|\Gamma_i| \to +\infty$ and bounded valency $\max_i \max_x \mathrm{val}(x) \leqslant v$. Then (Γ_i) is an expander family if and only if there exists $\lambda > 0$ such that*

$$\boldsymbol{\lambda}_1(\Gamma_i) \geqslant \lambda > 0$$

for all $i \in I$.

If each Γ_i is d-regular for a fixed $d \geqslant 3$, then $(\Gamma_i)_{i \in I}$ is an expander family if and only if there exists $\lambda' > 0$ such that

$$\underline{\lambda}_1(\Gamma_i) \geqslant \lambda' > 0$$

for all $i \in I$.

We call (λ, v), or (λ', d), the spectral expansion parameters *of the family. For d-regular graphs, one can take $\lambda' = d\lambda$.*

Exercise 3.4.5 (The general Laplace operator). Let $\Gamma = (V, E, \mathrm{ep})$ be a finite connected graph without isolated vertex. In addition to the space of functions on the vertices, let $L^2(E)$ be the space of complex-valued functions on E with the inner-product

$$\langle f_1, f_2\rangle_E = \frac{1}{2N}\sum_{\alpha\in E} |\,\mathrm{ep}(\alpha)|f_1(\alpha)\overline{f_2(\alpha)}.$$

An *orientation* of Γ is the data of two maps

$$b, e \,:\, E \to V$$

such that $\mathrm{ep}(\alpha) = \{b(\alpha), e(\alpha)\}$ for all $\alpha \in E$ (in other words, if α has two extremities, a choice of a "beginning" $b(\alpha)$ and an "end" $e(\alpha)$ of the edge). Given such an orientation, we define a linear map

$$d \,:\, \begin{cases} L^2(\Gamma) & \longrightarrow & L^2(E) \\ \varphi & \mapsto & d\varphi \end{cases}$$

where

$$d\varphi(\alpha) = \varphi(b(\alpha)) - \varphi(e(\alpha)).$$

(1) Show that for any φ_1, $\varphi_2 \in L^2(\Gamma)$, we have

$$\langle \underline{\Delta}_\Gamma\varphi_1, \varphi_2\rangle = \langle d\varphi_1, d\varphi_2\rangle_E,$$

where $\underline{\Delta}_\Gamma$ is the general Laplace operator given by (3.36).

(2) Deduce that the smallest non-zero eigenvalue of the Laplace operator is $\leqslant 2h(\Gamma)$.

The definition of expanders using the Laplace operator is qualitatively equivalent to that based on the expansion constant, and choosing one instead of the other may be a matter of personal taste. In concrete applications, on the other hand, it may well be the case that one requires that a family of graph satisfy specifically one of the two conditions (or three, if random walks are considered as slightly different). Even then, if the actual values of the expansion parameters (λ, v) or (h, v) are not important, there is no problem in using either definition.

But it can very well happen that one wishes to have expanders according to, say, the spectral definition, and that the explicit value $\lambda > 0$ of the spectral gap plays a role in the results (for instance, this matters enormously for applications of expander graphs involving sieve methods in number theory, as we will sketch in Section 5.3). In such cases, starting from the "wrong" definition and translating the parameters from the expansion constant to the spectral gap might lead to serious loss of precision, since the order of magnitude of $h(\Gamma)$ and $\underline{\lambda}_1(\Gamma)$ might differ quite significantly.

To give an example: suppose one requires the spectral gap for a sequence of d'-regular graphs (Γ'_n), which is obtained by "perturbation", as in Corollary 3.1.18, of a family of d-regular graphs (Γ_n). If one knows a lower-bound for the spectral gap of (Γ_n), we already know how to deduce one for Γ'_n, namely

$$\underline{\lambda}_1(\Gamma'_n) \geqslant \frac{h(\Gamma'_n)^2}{2d'} \geqslant c\frac{h(\Gamma_n)^2}{2d'} \geqslant \frac{c}{8d'}\underline{\lambda}_1(\Gamma_n)^2,$$

where $c > 0$ is given by Corollary 3.1.18. If the spectral gap $\underline{\lambda}_1(\Gamma_n)$ is fairly small, this is a significant loss, in comparison with the statement of Corollary 3.1.18 for the expansion constants. This suggests that it might be useful to look for an analogue of this result for the spectral gap, that does not involve any comparison with the expansion constant.

For the spectral gap of the normalized Laplace operator, there are a number of results and techniques towards this goal, as explained in [76, Ch. 13]. We will apply the following bound (which corresponds to [76, Th. 13.23] and is due to Diaconis and Saloff-Coste) in the special case of Cayley graphs (it is a bit simpler then, but the general case is also instructive).

Proposition 3.4.6 (Perturbing the Laplace operator). *Let $\Gamma = (V, E, \mathrm{ep})$ be a a finite connected graph with at least two vertices, and let $\Gamma' = (V, E', \mathrm{ep}')$ be another connected graph with the same vertex set.*

For each pair (x, y) of distinct vertices, let $\gamma_{x,y}$ be a path in Γ', of length $\ell(x, y) \geqslant 1$, from x to y. For each pair of distinct vertices $(s, t) \in V \times V$, let then $\mathcal{A}_{s,t}$ be the set of $(x, y) \in V \times V$ such that the path $\gamma_{x,y}$ passes successively by s and t, *i.e., such that there exists some i, $0 \leqslant i < \ell(x, y)$, with*

$$\varphi(\gamma_{x,y}(i)) = s, \qquad \varphi(\gamma_{x,y}(i + 1)) = t.$$

We then have

$$\boldsymbol{\lambda}_1(\Gamma') \geqslant \frac{1}{c_1 c_2} \boldsymbol{\lambda}_1(\Gamma)$$

where

$$c_1 = \max_{x \in V} \frac{\mu_{\Gamma'}(x)}{\mu_\Gamma(x)} = \max_{x \in V} \frac{N \operatorname{val}_{\Gamma'}(x)}{N' \operatorname{val}_\Gamma(x)},$$

$$c_2 = \frac{N'}{N} \max_{\substack{(s,t) \in V \times V \\ a'(s,t) \neq 0}} \frac{1}{a'(s, t)} \sum_{(x,y) \in \mathcal{A}_{s,t}} \ell(x, y) a(x, y). \tag{3.38}$$

Example 3.4.7. The quantity c_2 which appears in this result is not always straightforward to estimate, since one has to be careful to pick up paths between the vertices which do not go too often through the same edge.

Proof. The basic idea is to show that

$$\langle (\mathrm{Id} - M)\varphi, \varphi \rangle_\Gamma \leqslant c_2 \langle (\mathrm{Id} - M')\varphi, \varphi \rangle_{\Gamma'} \tag{3.39}$$

for each $\varphi \in L^2(\Gamma)$ (note that the underlying vector spaces of $L^2(\Gamma)$ and $L^2(\Gamma')$ coincide; only the inner-product changes). Since we have

$$\|\varphi\|_\Gamma^2 \geqslant c_1^{-1} \|\varphi\|_{\Gamma'}^2,$$

we see using (3.31) that such a bound immediately gives the stated inequality.

To prove (3.39), we begin again with (3.22):

$$\langle (\mathrm{Id} - M)\varphi, \varphi \rangle_\Gamma = \frac{1}{2N} \sum_{x,y \in V} a(x, y) |\varphi(x) - \varphi(y)|^2.$$

The non-zero terms are those corresponding to adjacent vertices in Γ. To introduce the edges of Γ' in the formula, we write the difference $\varphi(x) - \varphi(y)$ as a telescopic sum of differences along the successive vertices of the path $\gamma_{x,y}$:

$$\varphi(x) - \varphi(y) = \sum_{i=0}^{\ell(x,y)-1} \{\varphi(\gamma_{x,y}(i + 1)) - \varphi(\gamma_{x,y}(i))\},$$

and by the Cauchy-Schwarz inequality, we get

$$|\varphi(x) - \varphi(y)|^2 \leqslant \ell(x, y) \sum_{i=0}^{\ell(x,y)-1} |\varphi(\gamma_{x,y}(i + 1)) - \varphi(\gamma_{x,y}(i))|^2,$$

where the successive differences are between points which are, by definition, adjacent in Γ'. And from then on, we basically just gather things up as they flow: we write

$$\sum_{x,y\in V} a(x,y)|\varphi(x)-\varphi(y)|^2 \leqslant \sum_{x,y\in V} a(x,y)\ell(x,y)\sum_{i=0}^{\ell(x,y)-1} |\varphi(\gamma_{x,y}(i+1))-\varphi(\gamma_{x,y}(i))|^2$$
$$= \sum_{s,t\in V} a'(s,t)\beta(s,t)|\varphi(t)-\varphi(s)|^2$$

with $\beta(s,t)=0$ unless $a'(s,t)=0$ and otherwise, by definition, we have

$$\beta(s,t)=\frac{1}{a'(s,t)}\sum_{(x,y)\in\mathcal{A}_{s,t}} a(x,y)\ell(x,y).$$

This leads to

$$\langle(\mathrm{Id}-M)\varphi,\varphi\rangle_\Gamma \leqslant \frac{N'}{N}(\max_{s,t}\beta(s,t))\langle(\mathrm{Id}-M')\varphi,\varphi\rangle_{\Gamma'},$$

which is the same as (3.39). □

As a final remark, the spectral theory of graphs is a very useful and powerful tool in graph theory, well beyond simply giving a characterization of expansion. It is especially interesting in concrete applications because it is algorithmically very manageable to compute eigenvalues of the Markov operator or of the discrete Laplace operator, even for rather large graphs (because it is a problem of linear algebra). Hence any problem that can be reduced (even if only approximately) to spectral properties can be studied quite deeply. Examples are given in Trevisan's notes [112] on spectral partitioning. Another illustration is the paper of Varshney, Chen, Paniagua, Hall and Chklovskii [117] on the graph of the nervous system of the worm *c. elegans* (see Figure 1.2), which discusses (among other things) some of the spectral properties of this graph.

3.5. Expansion of Cayley graphs

When we specialize the general definitions and results of the previous sections to the case of a Cayley graph, we obtain group-theoretic reformulation of the definitions, which are as follows:

(1) Let G be a finite group, and $S\subset G$ is a non-empty[3] symmetric generating set. For the Cayley graph $\Gamma=\mathcal{C}(G,S)$, we have

$$h(\Gamma)=\min_{\substack{\emptyset\neq W\subset G\\ |W|\leqslant|G|/2}}\frac{|\mathcal{E}(W)|}{|W|}$$

with

$$|\mathcal{E}(W)|=|\{(g,s)\in W\times S\mid gs\notin W\}|$$

(a bijection from $\mathcal{E}(W)$ and the set on the right is $(g,s)\mapsto\{g,gs\}\in E_\Gamma$).

(2) The space $L^2(\Gamma,\mu_\Gamma)$ coincides with the space $L^2(G)$ of complex-valued functions on G, with the inner-product corresponding to the uniform probability measure on G, namely

$$\langle\varphi_1,\varphi_2\rangle=\frac{1}{|G|}\sum_{g\in G}\varphi_1(g)\overline{\varphi_2(g)}$$

[3] It could only be empty if G is trivial.

for φ_1 and φ_2 in $L^2(G)$. The Markov averaging operator is given by

$$M\varphi(g) = \frac{1}{|S|}\sum_{s\in S}\varphi(gs),$$

for $\varphi \in L^2(G)$ and $g \in G$. Therefore we have

$$\underline{\Delta}_\Gamma\varphi(g) = |S|\varphi(g) - \sum_{s\in S}\varphi(gs),$$

$$\langle\underline{\Delta}_\Gamma\varphi, \varphi\rangle = \frac{1}{2|G|}\sum_{\substack{g\in G\\ s\in S}}|\varphi(gs) - \varphi(g)|^2$$

for all $\varphi \in L^2(G)$ and, as usual, the minimization formula

$$\underline{\lambda}_1(\Gamma) = |S|\,\boldsymbol{\lambda}_1(\Gamma) = \min_{\varphi\perp 1}\frac{\langle\underline{\Delta}_\Gamma\varphi,\varphi\rangle}{\|\varphi\|^2}.$$

The most important feature distinguishing Cayley graphs from “general” regular graphs, is their symmetry (recall that G acts by graph automorphisms on Γ, see Proposition 2.3.8). In particular, “every vertex looks the same”. This has important consequences when applying Proposition 3.4.6 for Cayley graphs. Indeed, we obtain the following version of this result:

Proposition 3.5.1 (Perturbing Cayley graphs). *Let G be a finite group, S, $S' \subset G$ two non-empty finite symmetric generating sets of G, and denote $\Gamma = \mathcal{C}(G, S)$, $\Gamma' = \mathcal{C}(G, S')$ the associated Cayley graphs. We then have*

$$\boldsymbol{\lambda}_1(\Gamma') \geqslant c^{-1}\,\boldsymbol{\lambda}_1(\Gamma)$$

where $c = |S'|\max_{s\in S}(\ell_{S'}(s)^2)$.

Proof. We apply Proposition 3.4.6 to the two Cayley graphs. The quantity c_1 of the proposition is equal to 1, and for any x, $y \in V$, we take a path $\gamma_{x,y}$ in Γ' obtained by concatenating to x a path representing an expression

$$x^{-1}y = w_1\cdots w_m, \qquad w_i \in S', \tag{3.40}$$

as a word in the generators from S' (this exploits the homogeneity of the Cayley graphs). Thus we have $\ell(x, y) = m = \ell_{S'}(x^{-1}y)$.

We can then estimate c_2. Two elements g, $h \in G$ are joined by an edge of Γ' if $h = gs'$ for some $s' \in S'$. Similarly, x, y are joined by an edge of Γ (i.e., have $a(x, y) \neq 0$) if and only if $y = xs$ for some $s \in S$. The pair (x, y) belongs to $\mathcal{A}_{g,h}$ if the edge joining g to h appears in the path $\gamma_{x,y}$. In terms of the decomposition (3.40) of $x^{-1}y = s$, this happens exactly when $w_i = s'$ for some i, with

$$x = g(w_1\cdots w_{i-1})^{-1}, \qquad y = h(w_{i+1}\cdots w_m).$$

For each $s \in S$ we get therefore as many elements in $\mathcal{A}_{g,h}$ as there are occurrences of $s' = g^{-1}h$ in the S'-decomposition of s, say $\ell_{g^{-1}h}(s) \geqslant 0$ times, and we obtain

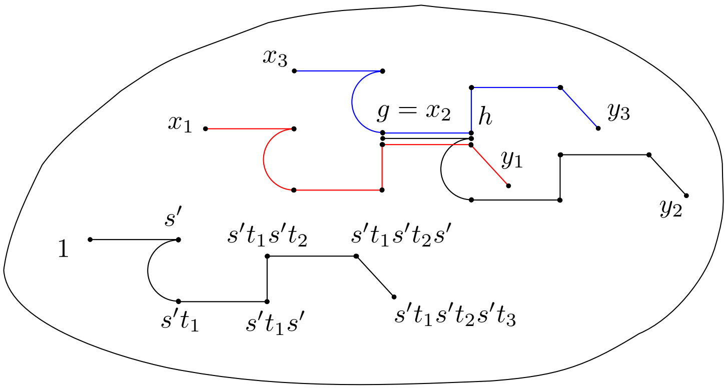
$$\sum_{(x,y)\in\mathcal{A}_{g,h}}\ell(x, y)a(x, y) \leqslant \sum_{s\in S}\ell_{S'}(s)\ell_{g^{-1}h}(s) \leqslant |S|\ell_{S'}(s)^2,$$

for all g and h. Referring to (3.38), this is precisely the desired formula since

$$\frac{N'}{N} = \frac{|S'|}{|S|}.$$

□

Remark 3.5.2. We illustrate the last computation: here $s = s't_1s't_2s't_3$, and we "show" the three paths from x_i to y_i on which the edge $\{g, h\} = \{g, gs'\}$ lies.



There is now a rather striking corollary of this, which illustrates how special Cayley graphs are: the diameter gives a rather good control of the spectral gap, and hence of the expansion constant!

Corollary 3.5.3 (Bounding the spectral gap from the diameter). *Let G be a finite group, $S \subset G$ a non-empty finite symmetric generating set of G. For the Cayley graph $\Gamma = \mathcal{C}(G, S)$, we have*

$$\underline{\lambda}_1(\Gamma) \geqslant \frac{1}{\operatorname{diam}(\Gamma)^2},$$

and hence

$$h(\Gamma) \geqslant \frac{1}{2\operatorname{diam}(\Gamma)^2}.$$

Proof. The last inequality follows from the first and from (3.37), and the first is equivalent to

$$\boldsymbol{\lambda}_1(\Gamma) \geqslant \frac{1}{|S|\operatorname{diam}(\Gamma)^2}.$$

We prove this lower bound by comparing Γ to the "perturbed" Laplace operator of $\widetilde{\Gamma} = \mathcal{C}(G, T)$ with $T = G$. More precisely we take $(\widetilde{\Gamma}, \Gamma)$ for (Γ, Γ') in Proposition 3.5.1 (with apologies for the possible confusion); we have $\ell_S(g) \leqslant \operatorname{diam}(\Gamma)$ for all generators $g \in T$, and therefore we get

$$\boldsymbol{\lambda}_1(\Gamma) \geqslant c^{-1}\,\boldsymbol{\lambda}_1(\widetilde{\Gamma})$$

with $c = |S|(\operatorname{diam}(\Gamma))^2$. But $\boldsymbol{\lambda}_1(\widetilde{\Gamma}) = 1$, which finishes the proof! Indeed, $\widetilde{\Gamma}$ is a complete graph with vertex set G, with an additional loop inserted at each vertex. Therefore, if we start from $X_0 = 1$, a random walk on $\widetilde{\Gamma}$ is exactly uniformly distributed at each step X_n for $n \geqslant 1$, which means that the Markov operator of $\widetilde{\Gamma}$ has no non-zero eigenvalue different from 1.

Analytically, we see this by noting that we have

$$\widetilde{M}\varphi(x) = \frac{1}{|G|}\sum_{g \in G}\varphi(xg) = \frac{1}{|G|}\sum_{g \in G}\varphi(g),$$

i.e., $\widetilde{M}$ is the orthogonal projection on the constants, and therefore its only eigenvalues are 0 and 1. □

Example 3.5.4. (1) The result is essentially sharp. Consider indeed the cycles $C_m = \mathcal{C}(\mathbf{Z}/m\mathbf{Z}, \{\pm 1\})$, for which we have $\mathrm{diam}(C_m) \sim m/2$ and

$$\underset{\sim}{\lambda}_1(\Gamma_m) \sim \frac{4\pi^2}{m^2} \sim \frac{\pi^2}{(\mathrm{diam}\, C_m)^2}$$

as $m \to +\infty$ by Example 3.2.26 (2) (taking into account that C_m is 2-regular.)

(2) An immediate consequence of Corollary 3.5.3, is an explicit uniform lower bound for the spectral gap of all Cayley graphs of finite groups of order bounded by some absolute constant, namely

$$\boldsymbol{\lambda}_1(\mathcal{C}(G,S)) \geqslant \frac{1}{|S||G|^2} \geqslant \frac{1}{|G|^3}. \tag{3.41}$$

This is slightly better than the general lower bound (3.35). This remark will be useful in Chapter 6, as it shows that – even when one wishes to get explicit estimates – one can always restrict attention to large enough groups in a family when one attempts to prove that they are expanders.

Using Corollary 3.5.3, we will see that it is relatively easy to produce explicitly some families of Cayley graphs which are "almost" expanders, in the following quantitative sense.

Definition 3.5.5 (Esperantist graphs). A family $(\Gamma_i)_{i\in I}$ of finite non-empty connected graphs $\Gamma_i = (V_i, E_i, \mathrm{ep})$ is an *esperantist family* if there exist constants $v \geqslant 1$, $c > 0$, $A \geqslant 0$, independent of i, such that the number of vertices $|V_i|$ tends to infinity, the maximal valency of Γ_i is at most v for all i, and the expansion constant satisfies

$$h(\Gamma_i) \geqslant \frac{c}{(\log 2|\Gamma_i|)^A}.$$

The point of this definition is that some applications of graph expansion turn out to require less than "full" expanders, and in particular are quite contented with taking an esperantist family as input (see the discussion in Section 5.5). And there are quite simple sequences of Cayley graphs which are esperantist but not expanders.

From Corollary 3.5.3, we see that, for Cayley graphs, the esperantist condition has not only an equivalent formulation in terms of spectral gap, but also one in terms of diameter growth.

Proposition 3.5.6 (Forms of esperantism). *A family $(\Gamma_i)_{i\in I}$ of finite non-empty Cayley graphs $\Gamma_i = \mathcal{C}(G_i, S_i)$, with $|G_i|$ tending to infinity and $|S_i| \leqslant v$ for all i, is an esperantist family if and only if one of the two properties below hold:*

(1) *For some $c > 0$ and $A \geqslant 1$, we have*

$$\mathrm{diam}(\Gamma_i) \leqslant c(\log 2|\Gamma_i|)^A.$$

(2) *For some $c' > 0$ and $A' \geqslant 0$, we have*

$$\underset{\sim}{\lambda}_1(\Gamma_i) \geqslant c'(\log 2|\Gamma_i|)^{-A}.$$

Proof. Supppose that (1) holds. Then by Corollary 3.5.3, we obtain

$$h(\Gamma_i) \geqslant \frac{1}{2\,\mathrm{diam}(\Gamma_i)^2} \geqslant \frac{1}{2C^2}\frac{1}{(\log 2|\Gamma_i|)^{2A}}.$$

Conversely, by Proposition 3.1.5, we have

$$\mathrm{diam}(\Gamma_i) \leqslant 2\frac{\log \frac{|\Gamma_i|}{2}}{\log\Bigl(1 + \frac{h(\Gamma_i)}{v}\Bigr)} + 3,$$

and for an esperantist family we can apply

$$\log(1+x) \geqslant \min\Bigl(\frac{x}{2}, \log(2)\Bigr)$$

for $x \geqslant 0$ to obtain

$$\log\Bigl(1+\tfrac{h(\Gamma_i)}{v}\Bigr) \geqslant \min\Bigl(\log 2, \frac{1}{2v}h(\Gamma_i)\Bigr) \geqslant \min\Bigl(\log 2, \frac{1}{2v}\frac{c}{(\log 2|\Gamma_i|)^A}\Bigr),$$

and hence

$$\operatorname{diam}(\Gamma_i) \ll (\log|\Gamma_i|)^{A+1},$$

which gives the polylogarithmic growth of the diameter.

As for (2), the equivalence follows immediately from (3.37). □

Example 3.5.7 (Symmetric groups as an esperantist family). Using these results, we can already exhibit an explicit esperantist family. Let $G_n = \mathcal{C}(\mathfrak{S}_n, S_n)$ for $n \geqslant 3$ as in Example 2.3.2, where we see again $\mathfrak{S}_n$ as acting on $\mathbf{Z}/n\mathbf{Z}$. We already know that $\operatorname{diam}(G_n) \ll n^2$ (by Exercise 2.3.5), and therefore we derive

$$\underline{\lambda}_1(G_n) \gg \frac{1}{n^4} \gg \frac{1}{(\log|G_n|)^4}$$

by Corollary 3.5.3, which proves the esperantist property. We also know (Exercise 3.1.14) that (G_n) is not an expander, since $h(G_n) \ll n^{-1}$. In fact, we can obtain a better result:

Proposition 3.5.8. *For the graphs (G_n) above, we have*

$$h(\mathcal{C}(\mathfrak{S}_n, S_n)) \ll \frac{1}{n^2}, \qquad \boldsymbol{\lambda}_1(\mathcal{C}(\mathfrak{S}_n, S_n)) \asymp \frac{1}{n^3},$$

for $n \geqslant 3$.

Proof. The random walk on these Cayley graphs is analyzed by Diaconis and Saloff-Coste in [32, §3, Ex. 1]. We will first show that

$$\boldsymbol{\lambda}_1(G_n) \ll n^{-3},$$

while the corresponding lower-bound $\boldsymbol{\lambda}_1(G_n) \gg n^{-3}$ is proved in [32], and we defer it to an exercise below. We then use the argument underlying the proof of the discrete Buser inequality to show that the bound $n^{-3/2}$ that follows directly from Proposition 3.3.8 can be improved to

$$h(G_n) \ll n^{-2}.$$

To get an upper bound for $\boldsymbol{\lambda}_1(G_n)$, we use a specific ad-hoc test function φ in the characterization (3.30). The goal is to have φ be "almost" invariant under multiplication by τ and by $\sigma_n^{\pm 1}$. Since $\sigma_n^{\pm 1}$ is a "circular" shift, it is therefore tempting to use a function defined using the cyclic ordering of $\mathbf{Z}/n\mathbf{Z}$. Thus the definition

$$\varphi(\sigma) = \text{the cyclic distance between } \sigma^{-1}(1) \text{ and } \sigma^{-1}(2) = d_{C_n}(\sigma^{-1}(1), \sigma^{-1}(2))$$

(using the distance on the cycle C_n, which has the same vertex set $\mathbf{Z}/n\mathbf{Z}$) may seem to have a good chance. Indeed, we have $\varphi(\sigma\sigma_n) = \varphi(\sigma\sigma_n^{-1}) = \varphi(\sigma)$ for all $x \in \mathbf{Z}/n\mathbf{Z}$, and therefore

$$\langle(\mathrm{Id} - M)\varphi, \varphi\rangle = \frac{1}{6|G_n|}\sum_{\sigma\in\mathfrak{S}_n}(\varphi(\sigma\tau) - \varphi(\sigma))^2.$$

The difference $\varphi(\sigma\tau) - \varphi(\sigma)$ is at most 1 in absolute value, and takes this value *only if* one of $\sigma^{-1}(1)$ or $\sigma^{-1}(2)$ is equal to 1 or 2, as a few minutes thoughts will convince the reader (but all such permutations do not contribute necessarily). There are at most $4(n-1)!$ permutations σ such that

$$\sigma^{-1}(1) \in \{1,2\} \text{ or } \sigma^{-1}(2) \in \{1,2\},$$

and hence we get

$$\langle (\mathrm{Id} - M)\varphi, \varphi\rangle \leqslant \frac{2}{3n}. \tag{3.42}$$

On the other hand, we have

$$\|\varphi - \langle \varphi, 1\rangle\|^2 = \frac{1}{|\mathfrak{S}_n|} \sum_{\sigma\in\mathfrak{S}_n} \varphi(\sigma)^2 - \Bigl(\frac{1}{|\mathfrak{S}_n|} \sum_{\sigma\in\mathfrak{S}_n} \varphi(\sigma)\Bigr)^2,$$

which we evaluate by using the distribution of values of φ. It is intuitively clear that the probability that $\varphi(\sigma)$ take any of its permitted values (integers between 1 and $n/2$) should be roughly the same. In fact, this holds exactly for n odd, leading in that case to

$$\frac{1}{|\mathfrak{S}_n|} \sum_{\sigma\in\mathfrak{S}_n} \varphi(\sigma)^2 = \frac{1}{|\mathfrak{S}_n|} \sum_{1\leqslant j\leqslant n/2} j^2 |\{\sigma \mid \varphi(\sigma) = j\}| = \frac{2}{n} \sum_{1\leqslant j\leqslant n/2} j^2 \sim \frac{n^2}{12}$$

for $n \to +\infty$. Similarly

$$\frac{1}{|\mathfrak{S}_n|} \sum_{\sigma\in\mathfrak{S}_n} \varphi(\sigma) = \frac{2}{n} \sum_{1\leqslant j\leqslant n/2} j \sim \frac{n}{4},$$

and hence

$$\|\varphi - \langle \varphi, 1\rangle\|^2 \sim \frac{n^2}{48}.$$

For n even, one checks that the cyclic distance $n/2$ is represented half as often as the others, but that this leads to the same asymptotic. The conclusion is that

$$\boldsymbol{\lambda}_1(\Gamma_1) \leqslant \frac{\langle (\mathrm{Id} - M)\varphi, \varphi\rangle}{\|\varphi - \langle \varphi, 1\rangle\|^2} \leqslant \frac{32}{n^3} + o(1)$$

as $n \to +\infty$.

We now apply (3.34) to estimate $h(G_n)$ (in other words, Lemma 3.3.9 with the uniform probability measure on $[1, \lfloor \frac{n}{2} \rfloor]$, which is the interval $[\min\varphi, \max\varphi]$). The median t_0 is roughly $n/4$, and by translating the estimate to our situation, we obtain

$$h(G_n) \leqslant \frac{A}{B}$$

with

$$A = \frac{1}{2} \sum_{\sigma\in\mathfrak{S}_n} |\varphi(\sigma) - \varphi(\sigma\tau)|, \qquad B = \sum_{\sigma\in\mathfrak{S}_n} |\varphi(\sigma) - t_0|.$$

As we have seen, $|\varphi(\sigma) - \varphi(\sigma\tau)|$ is either 0 or 1, and therefore we get

$$A \leqslant \frac{2|\mathfrak{S}_n|}{3n},$$

just as in (3.42). We estimate B by summing according to the values of φ, in the manner used for the computation of the norm $\|\varphi - \langle \varphi, 1\rangle\|^2$. This leads to

$$B \sim \frac{n|\mathfrak{S}_n|}{8},$$

for $n \to +\infty$, and therefore

$$h(G_n) \leqslant \frac{A}{B} \ll \frac{1}{n^2},$$

as claimed. □

Exercise 3.5.9 (A comparison of Cayley graphs). Let $\widetilde{G}_n = \mathcal{C}(\mathfrak{S}_n, T_n)$ where T_n is the set of all transpositions in $\mathfrak{S}_n$.

(1) Show that

$$\boldsymbol{\lambda}_1(\widetilde{G}_n) = \frac{2}{n}.$$

(2) Deduce by comparison that

$$\boldsymbol{\lambda}_1(G_n) \gg n^{-3}.$$

3.6. Matchings

This section can be skipped in a first reading. It will only be used later in the construction of Ramanujan graphs following Marcus, Spielman and Srivastava (see Section 4.2).

Let $\Gamma = (V, E, \mathrm{ep})$ be a finite graph. Given a permutation σ of the set V of vertices of Γ, we say that σ *is realized in* Γ if, for $x \in V$ such that $\sigma(x) \neq x$, there is an edge in Γ joining x and $\sigma(x)$. This means equivalently that, when writing σ as a product of disjoint cycles c, the graph Γ contains as subgraphs a cycle with the vertices corresponding to each such cycle. (In other words, for any vertex x_0, if the cycle of σ starting from x_0 is

$$x_0 \mapsto x_1 = \sigma(x_0) \mapsto \cdots \mapsto x_{k-1} = \sigma^{k-1}(x_0) \mapsto x_0 = \sigma^k(x_0),$$

where $(x_0, \ldots, x_{k-1})$ are distinct, then there are edges in Γ with extremities $\{x_0, x_1\}$, ..., $\{x_{k-1}, x_0\}$.)

Example 3.6.1. Assume that $V = \{1, \ldots, n\}$ with $n \geqslant 3$. The permutation $\sigma = (1\ 2\ 3)$ (i.e., the 3-cycle $1 \mapsto 2 \mapsto 3 \mapsto 1$) is realized in Γ if and only if the vertices $\{1, 2, 3\}$ form a triangle.

The interest of realizable permutations is the following elementary fact:

Lemma 3.6.2. *Assume that Γ is a simple graph. Let A_Γ be the adjacency operator of Γ. We have*

$$\det(X - A_\Gamma) = \sum_{\sigma \text{ realized}} \varepsilon(\sigma)(-1)^{|V| - f(\sigma)} X^{f(\sigma)},$$

where $f(\sigma)$ is the number of fixed points of σ.

Proof. The matrix of A_Γ in the basis of characteristic functions of vertices of Γ is the adjacency matrix $(a(x, y))$ of Γ. Expanding the determinant of this matrix, we obtain

$$\det(X - A_\Gamma) = \sum_{\sigma} \varepsilon(\sigma) X^{f(\sigma)} (-1)^{|V| - f(\sigma)} g(\sigma),$$

where the sum is over all permutations of V and

$$g(\sigma) = \prod_{\sigma(x) \neq x} a(x, \sigma(x)).$$

By the very definition, we see that $g(\sigma)$ is non-zero if and only if σ is realizable in Γ, and in that case, we have $g(\sigma) = 1$ because Γ is a simple graph. The formula follows. □

An important special case concerns those permutations σ that are involutions, i.e., such that $\sigma^2 = 1$. This means that σ is a product of transpositions exchanging disjoint pairs of vertices, and σ is realized if and only if each such pair is joined by an edge. Assume that this is the case. Let $V' \subset V$ be the set of $x \in V$ such that $\sigma(x) \neq x$, and E' a set of edges joining x to $\sigma(x)$ for all $x \in V'$. Then (V', E', ep) is a 1-regular subgraph in Γ. If we assume furthermore that Γ is a simple graph, then conversely, from any 1-regular subgraph Γ' of Γ, we can construct the permutation σ which is the product of the transpositions exchanging the vertices of the edges of Γ', and this permutation is an involution that is realized in Γ.

Definition 3.6.3. Let $\Gamma = (V, E, \mathrm{ep})$ be a graph. A *matching* of Γ is a 1-regular subgraph Γ' of Γ. If the vertex set of Γ' is equal to V, then Γ' is called a *perfect matching.*

The remarks preceding the definition show that if Γ is a finite simple graph, there is a bijection between the set of realized involutions (among the permutations of the vertices) and matchings in Γ.

Definition 3.6.4. Let $\Gamma = (V, E, \mathrm{ep})$ be a finite graph. The *matching polynomial* of Γ is the polynomial

$$p(\Gamma) = \sum_{\Gamma' \text{ matching in } \Gamma} (-1)^{|E'|} X^{|\Gamma| - |\Gamma'|} \in \mathbf{Z}[X],$$

where E' is the set of edges of Γ'.

Remark 3.6.5. (1) Since Γ has a unique matching with 0 vertices (the empty subgraph), the matching polynomial is monic of degree $|\Gamma|$.

(2) If Γ is a finite simple graph, then we can also write

$$p(\Gamma) = \sum_{\substack{\sigma^2 = 1 \\ \sigma \text{ realized}}} \varepsilon(\sigma) X^{f(\sigma)}$$

since each matching Γ' corresponds to a unique realized involution, with E' of cardinality the number of disjoint transpositions defining σ and $\Gamma - \Gamma'$ the set of fixed points of σ.

The following lemma is now quite easy:

Lemma 3.6.6. *Let $T = (V, E)$ be a finite tree. We have $p(T) = \det(X - A_T)$.*

Proof. Since T is a tree, by definition no cycle of length $\geqslant 3$ can be realized in T, which means that in the expansion of Lemma 3.6.2, only σ involving cycles of length $\leqslant 2$ occur. These are exactly the involutions. For an involution σ, the integer $|V| - f(\sigma)$ is even, hence the formula of Lemma 3.6.2 becomes

$$\det(X - A_T) = \sum_{\substack{\sigma^2 = 1 \\ \sigma \text{ realized}}} \varepsilon(\sigma) X^{f(\sigma)} = p(T)$$

by the remark above. □

Exercise 3.6.7. (1) If Γ is the disjoint union of two graphs Γ_1 and Γ_2, then $p(\Gamma) = p(\Gamma_1)p(\Gamma_2)$.

(2) Let v_0 be a vertex of Γ. Show that

$$p(\Gamma) = p(\Gamma - v_0) - \sum_{v \sim v_0} p(\Gamma - v_0 - v). \tag{3.43}$$

In Section 4.2, a key point in the proof of existence of Ramanujan graphs will be the following remarkable fact about matching polynomials, due to Godsil [45].

Theorem 3.6.8 (Godsil). *Let $d \geqslant 2$ be an integer. Let Γ be a d-regular finite graph without loops. All roots of $p(\Gamma)$ are real numbers, and have absolute value $\leqslant 2\sqrt{d-1}$.*

To prove this theorem, it is enough to prove that $p(\Gamma)$ divides the matching polynomial $p(T)$ of some finite tree T contained in a d-regular tree, since we have $p(T) = \det(X - A_T)$ by Lemma 3.6.6, and this polynomial has only real roots of absolute value $\leqslant 2\sqrt{d-1}$ by Lemma 3.2.35.

We may assume that Γ is not empty, and that it is connected (using Exercise 3.6.7). Let v_0 be a fixed vertex of Γ. We consider the path graph $\varpi_{v_0}(\Gamma)$ starting at v_0 (Example 2.2.16). Since this is a subgraph of the universal cover $\widehat{\Gamma}_{v_0}$ of Γ, which is a d-regular tree (Proposition 2.2.14), the following proposition establishes the desired statement:

Proposition 3.6.9 (Godsil). *Let Γ be a finite, non-empty, simple connected graph and v_0 a vertex of Γ. The matching polynomial $p(\Gamma)$ divides the matching polynomial $p(\varpi_{v_0}(\Gamma))$.*

Sketch of proof. We will only sketch the proof, leaving to the reader the exercise of checking a number of identities concerning path graphs; drawing on paper even a small graph and the corresponding path trees will clarify these identities.

The key formula is

$$\frac{p(\Gamma)}{p(\Gamma - v_0)} = \frac{p(\varpi_{v_0}(\Gamma))}{p(\varpi_{v_0}(\Gamma) - v_0)}. \tag{3.44}$$

Indeed, assuming it is correct, we can proceed by induction on $|\Gamma|$. If $|\Gamma| \leqslant 2$, the desired result can be checked directly. Suppose it holds for graphs with $|\Gamma| - 1$ vertices. It suffices then to note that $\varpi_{v_0}(\Gamma) - v_0$ is a disconnected graph that contains $\varpi_{v_0}(\Gamma - v_0)$ as a connected component. Then $p(\varpi_{v_0}(\Gamma) - v_0)$ is therefore divisible by $\varpi_{v_0}(\Gamma - v_0)$ (see Exercise 3.6.7), and writing the formula in the form

$$p(\Gamma) = p(\Gamma - v_0)\frac{p(\varpi_{v_0}(\Gamma))}{p(\varpi_{v_0}(\Gamma) - v_0)},$$

the divisibility follows for Γ.

The proof of (3.44) proceeds also by induction on $|\Gamma|$. If $|\Gamma| \leqslant 2$, one can check the result directly again (e.g., Γ is then a tree, but $\varpi_{v_0}(\Gamma)$ is isomorphic to Γ in that case). Assume now that $|\Gamma| \geqslant 3$ and that the formula holds for graphs Γ' with $|\Gamma'| \leqslant |\Gamma| - 1$.

We start from (3.43), namely

$$p(\Gamma) = p(\Gamma - v_0) - \sum_{v \sim v_0} p(\Gamma - v_0 - v),$$

where v runs over vertices adjacent to v_0. Dividing by $p(\Gamma - v_0)$ and then using induction for the graph $\Gamma - v_0$ (and the vertices v), we get

$$\frac{p(\Gamma)}{p(\Gamma - v_0)} = X - \sum_{v \sim v_0} \frac{p(\varpi_v(\Gamma - v_0) - v)}{p(\varpi_v(\Gamma - v_0))}. \tag{3.45}$$

Let v be adjacent to v_0. By applying Exercise 2.2.16, the graph $\varpi_v(\Gamma - v_0) - v$ can be identified with a disjoint union of $\varpi_y(\Gamma - v_0 - v)$ over all y adjacent to v, except v_0. On the other hand, the graph $\varpi_{v_0}(\Gamma) - v_0 - v_0 v$ (where $v_0 v$ is the vertex of the path graph corresponding to the edge from v_0 to v) is (isomorphic to) the disjoint union of the same graphs, and of $\varpi_y(\Gamma - v_0)$ for y adjacent to v_0, but distinct from v, again by Exercise 2.2.16. Combining these, we get

$$p(\varpi_{v_0}(\Gamma) - v_0 - v_0 v) = p(\varpi_v(\Gamma - v_0) - v) \prod_{\substack{y \sim v_0 \\ y \neq v}} p(\varpi_y(\Gamma - v_0)).$$

Once more using Exercise 2.2.16, $\varpi_{v_0}(\Gamma) \boldsymbol{-} v_0$ is the disjoint union over $v \sim v_0$ of $\varpi_v(\Gamma \boldsymbol{-} v_0)$, so

$$p(\varpi_{v_0}(\Gamma) \boldsymbol{-} v_0) = \prod_{v \sim v_0} p(\varpi_v(\Gamma \boldsymbol{-} v_0)).$$

Taking the quotients, we obtain

$$\frac{p(\varpi_{v_0}(\Gamma) \boldsymbol{-} v_0 \boldsymbol{-} v_0 v)}{p(\varpi_{v_0}(\Gamma) \boldsymbol{-} v_0)} = \frac{p(\varpi_v(\Gamma \boldsymbol{-} v_0) \boldsymbol{-} v)}{p(\varpi_v(\Gamma \boldsymbol{-} v_0))},$$

and (3.45) becomes

$$\frac{p(\Gamma)}{p(\Gamma \boldsymbol{-} v_0)} = X - \sum_{v \sim v_0} \frac{p(\varpi_{v_0}(\Gamma) \boldsymbol{-} v_0 \boldsymbol{-} v_0 v)}{p(\varpi_{v_0}(\Gamma) \boldsymbol{-} v_0)}.$$

Since $v_0 v$ runs over all vertices of $\varpi_{v_0}(\Gamma)$ adjacent to v_0, applying (3.43) to the path graph shows that

$$\frac{p(\Gamma)}{p(\Gamma \boldsymbol{-} v_0)} = \frac{p(\varpi_{v_0}(\Gamma))}{p(\varpi_{v_0}(\Gamma) \boldsymbol{-} v_0)}$$

as claimed. □

CHAPTER 4

Expanders exist

We will discuss in this chapter some rather different constructions of expanders. The variety of techniques involved is remarkable, especially so since we are far from exhaustive: we will not discuss, for instance, the zig-zag product of Reingold, Vadhan and Wigderson (see for instance [54, §9]), nor the Gabber-Galil construction (see [54, Th. 8.2]), nor the original Ramanujan graphs of Lubotzky, Phillips and Sarnak and Margulis (see [82] and [84], or the books [31], [78] or [101]). The recent book [26, Ch. 8–9] of Ceccherini-Silberstein, Scarabotti and Tolli discusses some of these.

4.1. Probabilistic existence of expanders

In this first section, we establish, using probabilistic arguments, the existence of expander families. This is the same technique that was used originally by Barzdin and Kolmogorov and by Pinsker [95, Lemma 1]. It turns out, in fact, that for many models of random graphs, there is a high probability that they are expanders, in the sense that there is a positive lower bound for the Cheeger constant, valid with high probability.

We will use a standard model to prove the result. To begin with, we construct bipartite expanders. Fix some integer $d \geqslant 3$. For any fixed $n \geqslant 1$ and any d-tuple $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$ of permutations of $\{1, \ldots, n\}$, we define a graph $\Gamma_{\boldsymbol{\sigma}}$ with vertex set

$$V = \{(i, 0) \mid 1 \leqslant i \leqslant n\} \cup \{(i, 1) \mid 1 \leqslant i \leqslant n\} = V_0 \cup V_1,$$

(independent of σ) and with edges joining $(i, 0)$ to $(\sigma_j(i), 1)$ for $1 \leqslant j \leqslant d$: formally, we take

$$E_{\boldsymbol{\sigma}} = \{(i, \sigma_j(i)) \mid 1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant d\},$$

and $\mathrm{ep}((i, \sigma_j(i))) = \{(i, 0), (\sigma_j(i), 1)\}$. These graphs are bipartite and d-regular, and they may have multiple edges.

We view these graphs as *random graphs* by thinking of the permutations σ_i as taken independently and uniformly at random in $\mathfrak{S}_n$. Thus the *probability that the graphs* Γ_σ *satisfy a property* $\mathcal{P}(\Gamma)$ of graphs, denoted $\mathbf{P}(\Gamma_\sigma \text{ has } \mathcal{P})$, is simply

$$\mathbf{P}(\Gamma_\sigma \text{ has } \mathcal{P}) = \frac{1}{|\mathfrak{S}_n|^d} |\{\boldsymbol{\sigma} \in \mathfrak{S}_n^d \mid \Gamma_\sigma \text{ has } \mathcal{P}\}| = \frac{1}{(n!)^d} |\{\boldsymbol{\sigma} \in \mathfrak{S}_n^d \mid \Gamma_\sigma \text{ has } \mathcal{P}\}|.$$

We will prove:

Theorem 4.1.1. *Fix $d \geqslant 3$. There exists $h_d > 0$ such that*

$$\lim_{n \to +\infty} \mathbf{P}(h(\Gamma_{\boldsymbol{\sigma}}) < h_d) = 0.$$

In particular, for all n large enough, some Γ_σ satisfies $h(\Gamma_\sigma) \geqslant h_d$.

Remark 4.1.2. Here is one justification for hoping that such a result could be true. Recall that we suggested at the end of Section 3.1 that a possible way of constructing expanders would be to start with the finite trees $T_{d,k}$ of depth $k \geqslant 1$ with $d \geqslant 3$ fixed and $k \to +\infty$, and attempt to add some edges connecting the leaves of the tree to vertices "in the core" of the tree, and in particular to vertices on other branches from the root. Some elementary attempts of defining a family of edges of this type turn out to fail – either because the resulting graphs are again too easily disconnected, or because they seem hard to analyze. But these attempts might suggest that the best chance is to "throw edges at random". However, at this point, one can also simply decide that *all* edges should be placed randomly, to avoid dealing with two types of edges. This might naturally lead to the graphs of the type we consider here.

Proof. Since the graphs we construct are bipartite, we can use Lemma 3.1.16, which shows that

$$\mathbf{P}(h(\Gamma_{\boldsymbol{\sigma}}) < h_d) \leqslant \mathbf{P}(\widehat{h}(\Gamma_{\sigma}) < 1 + 2h_d)$$

where $\widehat{h}$ is defined in (3.6).

A simple symmetry argument (left to the reader) shows that, for any any $\delta > 0$, we have

$$\mathbf{P}(\widehat{h}(\Gamma_{\sigma}) < 1 + \delta) = \mathbf{P}\Big(\min_{\substack{W \subset V_0 \\ 1 \leqslant |W| \leqslant |V_0|/2}} \frac{|\partial_{\boldsymbol{\sigma}} W|}{|W|} < 1 + \delta\Big)$$

where we indicate with the subscript $\boldsymbol{\sigma}$ that the boundary is, of course, taken in the sense of $\Gamma_{\boldsymbol{\sigma}}$ (compare with (3.6): this means we only need to consider expansion of vertices in the "input" set V_0). Let p_n denote the right-hand side of this inequality.

We next write the simplest upper-bound for p_n, taking advantage of the fact that the vertex set is independent of $\boldsymbol{\sigma}$:

$$p_n \leqslant \sum_{\substack{W \subset V_i \\ 1 \leqslant |W| \leqslant n/2}} \mathbf{P}\Big(|\partial_{\boldsymbol{\sigma}} W| < (1+\delta)|W|\Big).$$

Again for symmetry reasons, the probability $\mathbf{P}\big(|\partial_{\boldsymbol{\sigma}} W| < (1+\delta)|W|\big)$ depends only on $|W|$, because any subset of size ℓ in V_0 is, for this model of random graphs, equivalent to $\{1, \ldots, \ell\}$. Hence

$$p_n \leqslant \sum_{1 \leqslant \ell \leqslant n/2} \binom{n}{\ell} \mathbf{P}\Big(|\partial_{\boldsymbol{\sigma}}\{1, \ldots, \ell\}| < (1+\delta)\ell\Big).$$

Since there are edges joining $(0, j)$ to $(1, \sigma_1(j))$ for all j, and σ_1 is a bijection, we always have

$$|\partial_{\boldsymbol{\sigma}}\{1, \ldots, \ell\}| \geqslant \ell.$$

Hence, if $|\partial_{\boldsymbol{\sigma}}\{1, \ldots, \ell\}| < (1+\delta)\ell$ for some $\boldsymbol{\sigma}$, it is necessary that at most $\delta\ell$ of the values $\sigma_2(j)$ for $1 \leqslant j \leqslant \ell$ be outside the set

$$I_\ell = \{\sigma_1(1), \ldots, \sigma_1(\ell)\},$$

and similarly for σ_3, …, σ_d, since the occurrence of any of these events would imply $|\partial_{\boldsymbol{\sigma}}\{1, \ldots, \ell\}| \geqslant (1+\delta)\ell$. These events are independent, since $\sigma_2, \ldots, \sigma_d$ are independently chosen in the symmetric group. We only consider σ_2 and σ_3, and find that the probability that they both occur is at most

$$\sum_{|E| = \lfloor \delta\ell \rfloor} \mathbf{P}\Big(\sigma_2(\{1, \ldots, \ell\}) \subset I_\ell \cup E\Big)^2,$$

since for each of the two independent events the probability is the same. Again because of symmetry, this quantity is at most

$$\binom{n}{\lfloor \delta\ell \rfloor} \mathbf{P}\Big(\sigma_2(\{1, \ldots, \ell\}) \subset I_\ell \cup \{\ell+1, \ldots, \ell + \lfloor \delta\ell \rfloor\}\Big)^2.$$

We have furthermore

$$\mathbf{P}\Bigl(\sigma_2(\{1,\ldots,\ell\})\subset I_\ell\cup\{\ell+1,\ldots,\ell+\lfloor\delta\ell\rfloor\}\Bigr)\\ \leqslant \frac{1}{n!}(\ell+\lfloor\delta\ell\rfloor)(\ell+\lfloor\delta\ell\rfloor-1)\cdots(1+\lfloor\delta\ell\rfloor)(n-\ell)!$$

which expresses the fact that if σ is a permutation with $\sigma(\{1,\ldots,\ell\})\subset I_\ell\cup E$, we have $\ell+\lfloor\delta\ell\rfloor$ possibilities for $\sigma(1)$, one less for $\sigma(2)$, and so on, until $1+\lfloor\delta\ell\rfloor$ possible values of $\sigma(\ell)$, and then finally $(n-\ell)!$ possibilities for the remaining values of σ.

We are thus left with the estimate

$$\begin{aligned} p_n &\leqslant \sum_{1\leqslant\ell\leqslant n/2}\binom{n}{\ell}\binom{n}{\lfloor\delta\ell\rfloor}\Bigl(\frac{(\ell+\lfloor\delta\ell\rfloor)!}{\lfloor\delta\ell\rfloor!}\Bigr)^2\Bigl(\frac{(n-\ell)!}{n!}\Bigr)^2\\ &= \sum_{1\leqslant\ell\leqslant n/2}\frac{(n-\ell)!((\ell+\lfloor\delta\ell\rfloor)!)^2}{\ell!(n-\lfloor\delta\ell\rfloor)!(\lfloor\delta\ell\rfloor)!^3}\end{aligned}$$

At this point, many treatments of the question just ask the reader to complete the estimation of this sum using the Stirling Formula, to check that it tends to 0 when δ is fixed and sufficiently small. This is a good exercise certainly, but we will give some details, explaining *why* one can quickly convince oneself that this should be indeed small.

There are two different arguments. When ℓ is fixed observe that

$$\frac{(n-\ell)!((\ell+\lfloor\delta\ell\rfloor)!)^2}{\ell!(n-\lfloor\delta\ell\rfloor)!((\lfloor\delta\ell\rfloor)!)^3}\leqslant C_\ell\frac{1}{(n-\lfloor\delta\ell\rfloor)\cdots(n-\ell+1)}$$

for some constant $C_\ell\leqslant((2\ell)!)^2$, which shows that the contribution of ℓ bounded (or even growing slowly with n) tends to zero.

For ℓ large, we use the Stirling Formula to estimate the logarithm of each term

$$\log\Bigl(\frac{(n-\ell)!((\ell+\lfloor\delta\ell\rfloor)!)^2}{\ell!(n-\lfloor\delta\ell\rfloor)!(\lfloor\delta\ell\rfloor)!^3}\Bigr)\tag{4.1}$$

in the sum. This has every chance to succeed, because of the following observation: the Stirling formula can be written

$$\log(k!)=A_k-B_k+C_k+O(1),$$

for $k\geqslant 1$, where each of the three terms is positive and tends to infinity with k, but they have different orders of magnitude, namely $A_k=k\log k$, $B_k=k$ and $C_k=\frac{1}{2}\log(2\pi k)$. In the expression (4.1), the contribution of the second terms B_k appearing in each factorial cancels out: the denominator gives

$$\ell+(n-\lfloor\delta\ell\rfloor)+3\lfloor\delta\ell\rfloor=n+\ell+2\lfloor\delta\ell\rfloor,$$

as does the numerator. We can therefore expect that (4.1) will tend to 0 if the contribution of the largest part of the factorial, namely A_k, goes to $-\infty$. And here the key point is that if δ is small, then the decay of the A_k-term from $(n-\lfloor\delta\ell\rfloor)!$ in the denominator will be much faster than the growth from $(n-\ell)!$ (stated another way: for $\delta=0$, the numerator and denominator balance perfectly, and both can be expected to become smaller as δ increases a bit, but the denominator will do so much more slowly; note however that if, say, $\delta=1$, the quantity (4.1) is $((2\ell)!)^2/(\ell!)^4$, which will definitely go to infinity, so one must be careful.)

The contribution of the A_k terms in the Stirling formula is

$$(n-\ell)\log(n-\ell)+2(1+\delta)\ell\log((1+\delta)\ell)\\ -\ell\log\ell-(n-\delta\ell)\log(n-\delta\ell)-3\delta\ell\log(\delta\ell)$$

(disregarding the difference between $\lfloor\delta\ell\rfloor$ and $\delta\ell$ for simplicity). The first and fourth term together are (roughly) equal to

$$n\log\Bigl(\frac{1-\ell/n}{1-\delta\ell/n}\Bigr)-\ell\log(n-\ell)+\delta\ell\log(n-\delta\ell)\approx(\delta-1)\ell+(\delta-1)\ell\log(n). \tag{4.2}$$

The second and third term contribute

$$2(1+\delta)\ell\log((1+\delta)\ell)-\ell\log\ell\approx 2\delta(1+\delta)\ell+(1+2\delta)\ell\log(\ell).$$

Together with the fifth term, this becomes

$$\approx\Bigl((\delta-1)+2\delta(1+\delta)-3\delta\log(\delta)\Bigr)\ell+(1-\delta)\ell\log(\ell). \tag{4.3}$$

If we sum (4.2) and (4.3), the leading contribution (in view of the bound $n\geqslant\ell/2$) becomes

$$(\delta-1)\ell\log(n)+(1-\delta)\ell\log(\ell)\leqslant\log(2)(\delta-1)\ell.$$

Since this is negative, the result will follow... □

4.2. Ramanujan graphs

The definition of an expander family exhibits the remarkable feature of being quantitative in some sense (it refers to quantitative properties of the expansion constant) and qualitative in another (it asks for the existence of *some* positive lower bounds for the expansion constants). In applications, as we will see in Chapter 5, it happens frequently however that the value of this lower bound plays a role (in the random walk definition, this is obviously related to the speed of convergence to a uniform measure). It is natural to ask if the expansion or equidistribution constants can have a meaning, or in a related way, how good can equidistribution be in the best possible world.

Although (to the author's knowledge) the values and limits of the expansion constant for expander families do not have any special property or interpretation, it turns out that the equidistribution parameters in the random walk interpretation (Definition 3.3.1) can have some meaning, and in particular that it is natural to consider optimal cases: these are known as *Ramanujan graphs.*

Definition 4.2.1. Let $d\geqslant 2$ be an integer. A d-regular connected finite graph Γ is called a *Ramanujan graph* if all the eigenvalues λ of the Markov operator M of Γ satisfy either $\lambda\in\{-1,1\}$ or $|\lambda|\leqslant\frac{2\sqrt{d-1}}{d}$, or in other words, if $\varrho_\Gamma\leqslant\frac{2\sqrt{d-1}}{d}$.

To understand why this definition is not arbitrary, recall that for any non-empty d-regular graph Γ, the universal cover $\widehat{\Gamma}$ (based at any vertex) is an infinite d-regular tree T_d (Proposition 2.2.14), and that the spectrum of the Markov operator of T_d is contained in the interval

$$\Bigl[-\frac{2\sqrt{d-1}}{d},\frac{2\sqrt{d-1}}{d}\Bigr]$$

(Proposition 3.2.31). Assume that Γ is connected. Since, by Corollary 3.2.20, the eigenvalues ±1 of the Markov operator correspond to either the constant function, or to the characteristic functions of the two parts of a bipartite decomposition (if Γ is bipartite), the definition of a Ramanujan graph therefore means precisely that all other eigenvalues of M must be contained in the spectrum of the universal cover of Γ. Moreover, a result of Alon-Boppana (see, e.g. [54, Th. 5.3] or [101, Prop. 3.2.7]) shows that this is the strongest possible restriction for an infinite family of graphs: if $(\Gamma_i)_{i\in I}$ is any family of d-regular connected graphs with $|\Gamma_i| \to +\infty$, then we have

$$\limsup_{i} \varrho_{\Gamma_i} \geqslant \frac{2\sqrt{d-1}}{d}.$$

Example 4.2.2. (1) Let $d \geqslant 3$. The complete graph K_d is a Ramanujan graph: indeed by Example 3.2.26, we have $\varrho_{K_d} = 1/(d-1) \leqslant 2\sqrt{d-1}/d$.

(2) Let $d \geqslant 3$ and let $K_{d,d}$ be the complete bipartite graph with input set $V_0 = \mathbf{Z}/d\mathbf{Z}$ and output set $V_1 = \mathbf{Z}/d\mathbf{Z}$ (Example 2.1.24, (2)). Then $K_{d,d}$ is also a Ramanujan graph. Indeed, since $K_{d,d}$ is bipartite, both 1 and -1 are eigenvalues of the Markov operator. But also, the kernel of the Markov operator is the space of $f \in L^2(K_{d,d})$ such that

$$\sum_{x\in V_0} f(x) = \sum_{x\in V_1} f(x) = 0,$$

which has codimension 2 in $L^2(K_{d,d})$. This means that 0 is the only eigenvalue of the Markov operator on $L^2_0(K_{d,d})$.

Since Ramanujan graphs are, individually, the best-possible graphs from the point of view of the Markov operator, one can ask if they can form expanders. In other words, does there exist an infinite family of Ramanujan graphs with bounded valency and increasing size? This turns out to be a rather subtle question. The paper where Ramanujan graphs were first defined by Lubotzky, Phillips and Sarnak [82] contains explicit examples of infinite families of d-regular Ramanujan graphs (also discovered independently by Margulis [84], both constructions relying on deep arithmetic input due to Deligne and Drinfeld), but only when $d = p+1$ for some prime number p. This essential restriction was related to the specific arithmetic origin of these graphs. Further examples, always relying on number theory, produced examples with $d = p^\nu \pm 1$ for $\nu \geqslant 1$, always with p prime. Only quite recently have Marcus, Spielman and Srivastava [83] constructed Ramanujan graphs of arbitrary degree:

Theorem 4.2.3. *Let $d \geqslant 3$ be an integer. There exists a family $(\Gamma_i)_{i\geqslant 0}$ of* bipartite *d-regular Ramanujan graphs with $|\Gamma_i| = d2^i$.*

The proof uses a probabilistic argument, but in rather different manner than Section 4.1: the idea is to show that given any starting d-regular bipartite Ramanujan graph Γ, there exists another bipartite Ramanujan graph Γ' with $|\Gamma'| = 2|\Gamma|$ which is a "2-covering" of Γ. This property had been conjectured by Bilu[1] and Linial. Applied inductively, starting with the "trivial" example of the complete bipartite graph $K_{d,d}$ (Example 4.2.2, (2)), the theorem follows. (For generalizations to other coverings, see the paper [49] of Hall, Puder and Sawin.)

The probabilistic ingredient is found in taking a family of 2-coverings Γ' of Γ, and showing that one of them must have the required property. To make this precise, we now define this family.

First, we fix an integer $d \geqslant 2$ and a simple d-regular connected graph without loops $\Gamma = (V, E, \mathrm{ep})$. Let $\mathcal{S}_\Gamma$ be the set of all functions $s \colon E \to \{-1, 1\}$ (called "signings" of the edges of Γ). We view $\mathcal{S}_\Gamma$ as a probability space with the uniform probability measure,

[1] Yonatan Bilu, not Yuri Bilu.

so that the expectation notation $\mathbf{E}(\cdot)$ refers to expectation over $\mathcal{S}_\Gamma$ with respect to the uniform measure, namely

$$\mathbf{E}(f(s)) = \frac{1}{|\mathcal{S}_\Gamma|}\sum_{s\in\mathcal{S}_\Gamma} f(s) = \frac{1}{2^{|E|}}\sum_{s\in\mathcal{S}_\Gamma} f(s)$$

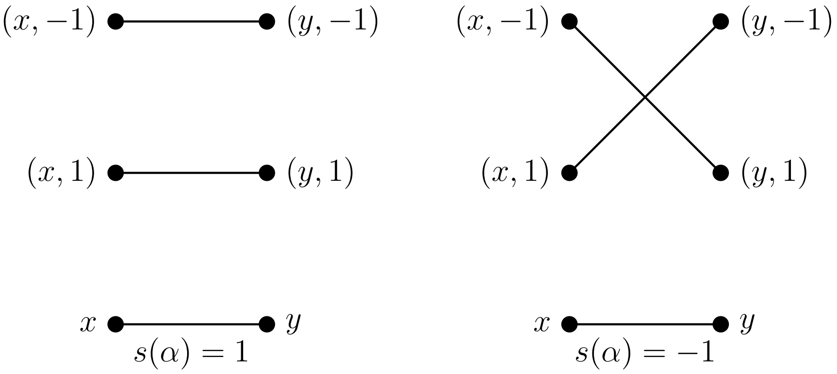
for any function f on $\mathcal{S}$. By definition, this choice of measure implies that the random variables $(s \mapsto s(\alpha))_{\alpha\in E}$ are independent.

Let $s \in \mathcal{S}_\Gamma$. Since Γ is a simple graph, we can also define $s(x, y) = s(\alpha)$ if x and y are adjacent in Γ, where α is the unique edge joining x and y.

We define a simple graph Γ_s with vertex set $V_s = V \times \{-1, 1\}$ (independently of s) by specifying which vertices (x, ε) and (y, ε') are joined by an edge (a special case of Example 2.1.8 (3)) as follows: there is no edge unless x and y are connected by an edge α in Γ; if this is the case, there is an edge if and only if $\varepsilon' = s(\alpha)\varepsilon$.

In particular, the graph Γ_s is d-regular, since each vertex (x, ε) is connected exactly to one vertex (y, ε') for each $y \sim x$.

To understand the definition more intuitively, it is always useful to make some pictures. There is a surjective graph map f from Γ_s to Γ defined by $f((x, \varepsilon)) = x$, since (x, ε) and (y, ε') can be connected only if x and y are (we do not need to specify f_* more precisely since Γ_s is a simple graph). For each x and y in Γ with an edge α joining them, there are two vertices $(x, 1)$ and $(x, -1)$ mapping to x and two vertices $(y, 1)$ and $(y, -1)$ mapping to y. Then there are two edges mapping to α: they are either "parallel" edges from $(x, 1)$ to $(y, 1)$ and from $(x, -1)$ to $(y, -1)$, when $s(\alpha) = 1$, or they are "crossing" edges from $(x, 1)$ to $(y, -1)$ and from $(x, -1)$ to $(y, 1)$, in the case $s(\alpha) = -1$.



If Γ is bipartite, then composing $f\colon \Gamma_s \to \Gamma$ with a graph map $\Gamma \to P_1$ corresponding to the bipartite decomposition (Proposition 2.1.25), we see that Γ_s is also bipartite. In particular, its spectrum is then symmetric (Corollary 3.2.20 (4)).

There is a further graph map $\iota\colon \Gamma_s \to \Gamma_s$ given on vertices by $(x, \varepsilon) \mapsto (x, -\varepsilon)$, since the definition shows that (x, ε) is connected to (y, ε') if and only if $(x, -\varepsilon)$ is connected to $(y, -\varepsilon')$. The map ι is an involution without fixed points, and satisfies $\pi \circ \iota = \pi$. In particular, x and y are connected if and only if $\iota(x)$ and $\iota(y)$ are connected.

Lemma 4.2.4. *There exists an isometry*

$$L^2(\Gamma) \oplus L^2(\Gamma) \xrightarrow{j} L^2(\Gamma_s),$$

where the direct sum is orthogonal, such that the Markov operator M_s of Γ_s is identified with $M \oplus \widetilde{M}_s$, where M is the Markov operator of Γ and $\widetilde{M}_s$ is the self-adjoint linear *operator on $L^2(\Gamma)$ such that*

$$\widetilde{M}_s\varphi(x) = \frac{1}{d}\sum_{y\sim x} s(x,y)\varphi(y). \tag{4.4}$$

In particular, the spectrum of M_s is the union of the spectrum of M and the spectrum of $\widetilde{M}_s$.

Proof. Define subspaces H_+ and H_- of $L^2(\Gamma_s)$ by

$$H_+ = \{f\colon V_s \to \mathbf{C} \mid f\circ\iota = f\}, \qquad H_- = \{f\colon V_s \to \mathbf{C} \mid f\circ\iota = -f\}$$

(which one can think as "even" and "odd" functions). The decomposition

$$f = \frac{f+f\circ\iota}{2} + \frac{f-f\circ\iota}{2}$$

shows that $H_+\oplus H_- = L^2(\Gamma_s)$, and the sum is orthogonal since

$$\langle f+f\circ\iota, f-f\circ\iota\rangle = \langle f,f\rangle - \langle f\circ\iota, f\circ\iota\rangle + \langle f\circ\iota, f\rangle - \langle f, f\circ\iota\rangle = 0.$$

We define j_+ for $\varphi\in H_+$ by $j_+(\varphi)(x,\varepsilon) = \varphi(x)$, and for $\varphi\in H_-$, we define

$$j_-(\varphi)(x,1) = \varphi(x), \qquad j_-(\varphi)(x,-1) = -\varphi(x).$$

It is then easy to check that j_+ (resp. j_-) is a linear isomorphism from $L^2(\Gamma)$ to H_+ (resp. H_-). Moreover, we have

$$\|j_+(\varphi)\|^2 = \frac{1}{2|V|}\sum_{x\in V}\sum_{\varepsilon=\pm1}|\varphi(x)|^2 = \|\varphi\|^2,$$

and similarly $\|j_-(\varphi)\|^2 = \|\varphi\|^2$. The linear map $j = j_+\oplus j_-$ is therefore an isometric isomorphism from the orthogonal direct sum $L^2(\Gamma)\oplus L^2(\Gamma)$ to $L^2(\Gamma_s)$.

Furthermore, the Markov operator M_s of Γ_s commutes with ι: for $\varphi\in L^2(\Gamma_s)$ and any $x\in V_s$, we have

$$(M_s\varphi)(\iota(x)) = \frac{1}{d}\sum_{y\sim_s\iota(x)}\varphi(y) = \frac{1}{d}\sum_{w\sim_s x}\varphi(\iota(w)) = \frac{1}{d}\sum_{w\sim_s x}\varphi(w) = (M_s\varphi)(x)$$

since $\varphi(\iota(w)) = \varphi(w)$. It follows that both H_+ and H_- are stable under M_s.

Moreover, this argument also shows that the endomorphism of H_+ induced by M_s is the Markov operator M of Γ. If $\varphi\in H_-$, on the other hand, we compute for any $x\in V$ that

$$(M_s\varphi)((x,1)) = \frac{1}{d}\sum_{(y,\varepsilon)\sim(x,1)}\varphi((y,\varepsilon)) = \frac{1}{d}\sum_{\substack{y\sim x\\ s(x,y)=1}}\varphi((y,1)) + \frac{1}{d}\sum_{\substack{y\sim x\\ s(x,y)=-1}}\varphi((y,-1)),$$

By definition of H_-, this is

$$(M_s\varphi)((x,1)) = \frac{1}{d}\sum_{y\sim x}s(x,y)\varphi(y),$$

which means that M_s restricted to H_- is identified with the endomorphism $\widetilde{M}_s$ in the statement of the lemma. □

We denote by $\varrho_\Gamma(s)$ the spectral radius of $\widetilde{M}_s$ acting on H_-. It corresponds, intuitively, to the largest "new" eigenvalue of the graph Γ_s, when comparing its spectrum with that of Γ, which it always contains. (Note that it may be that $\varrho_\Gamma(s) = 1$: this will happen for instance if $s(\alpha) = 1$ for all edges α, since in that case Γ_s is the disjoint union of two copies of Γ, hence is not connected, and $\widetilde{M}_s$ has 1 as eigenvalue in that case).

The main result of this section is the following:

THEOREM 4.2.5 (Marcus, Spielman, Srivastava). *Let $d \geqslant 3$. Let Γ be a finite connected d-regular* bipartite *graph. There exists $s \in \mathcal{S}_\Gamma$ such that $\varrho_\Gamma(s) \leqslant \frac{2\sqrt{d-1}}{d}$.*

To prove this, the idea is to average over $\mathcal{S}_\Gamma$, in the same way that one can prove that a function is non-zero somewhere by showing that its average is non-zero. But the problem is, what should be averaged here? We need to control the largest eigenvalue of $\widetilde{M}_s$ (in fact, the largest absolute value of an eigenvalue). Because we work with bipartite graphs, this is the same as to control the largest (real) zero of the characteristic polynomial of $\widetilde{M}_s$, which is a rather mysterious quantity, and one which is by no means well-behaved for arbitrary polynomials. However, this characteristic polynomial turns out to have specific features that lead to the conclusion.

The next lemma refers to the properties of matchings considered in Section 3.6.

LEMMA 4.2.6. *Let $d \geqslant 3$. Let Γ be a finite connected d-regular graph. Then $\mathbf{E}(\det(X - d\widetilde{M}))$ is the matching polynomial $p(\Gamma)$ of Γ.*

PROOF. For any $s \in \mathcal{S}_\Gamma$, the linear map $\widetilde{M}_s$ is an endomorphism of $L^2(\Gamma)$. We denote the matrix of $d\widetilde{M}_s$ in the basis of characteristic functions of vertices of Γ by $(a_s(x, y))_{x,y\in V}$. By (4.4), we have $a_s(x, y) = 0$ unless x and y are adjacent, in which case $a_s(x, y) = s(\alpha)$, where α is the edge joining x and y. Viewed as random variables on $\mathcal{S}_\Gamma$, the coefficient maps $s \mapsto a_s(x, y)$, for (x, y) in V, satisfy $\mathbf{E}(a_s(x, y)) = 0$ in all cases.

We expand the determinant as a sum over the permutations σ of V, and obtain

$$\det(X - d\widetilde{M}_s) = \sum_\sigma \varepsilon(\sigma) X^{f(\sigma)} (-1)^{|V|-f(\sigma)} g(s, \sigma)$$

where $f(\sigma)$ is the number of fixed points of σ and

$$g(s, \sigma) = \prod_{\sigma(x)\neq x} a_s(x, \sigma(x)).$$

We now claim that $\mathbf{E}(g(s, \sigma)) = 0$ unless the permutation σ is an involution realized in Γ, in which case $\mathbf{E}(g(s, \sigma)) = 1$. If we prove this, then the lemma follows from Remark 3.6.5 (2).

First of all, we have $g(s, \sigma) = 0$ for all s unless x and $\sigma(x)$ are always joined by an edge if $\sigma(x) \neq x$. Next, assume that σ is a permutation realized in Γ. For each $x \in V$, let α_x be the edge joining x to $\sigma(x)$. We then have

$$\mathbf{E}(g(s, \sigma)) = \mathbf{E}\Bigl(\prod_{\sigma(x)\neq x} s(\alpha_x)\Bigr).$$

A given edge α can appear at most twice as α_x, namely we can have $\alpha_x = \alpha_{\sigma^{-1}(x)}$ if $\sigma^{-1}(x) = \sigma(x)$. Since the random variables $(s \mapsto s(\alpha))_{\alpha\in E}$ are independent, if some α_{x_0} appears a single time, we get

$$\mathbf{E}(g(s, \sigma)) = \mathbf{E}(s(\alpha_{x_0}))\,\mathbf{E}\Bigl(\prod_{\substack{\sigma(x)\neq x\\ \alpha_x\neq\alpha_{x_0}}} s(\alpha_x)\Bigr) = 0.$$

Hence $\mathbf{E}(g(s,\sigma)) = 0$ unless each edge α_x appears twice, which means that $\sigma^2(x) = x$ for all x, or in other words if σ is a realizable involution. In that case, since $s(\alpha_x)s(\alpha_{\sigma(x)}) = s(\alpha_x)^2 = 1$, we have $\mathbf{E}(g(s,\sigma)) = 1$, as claimed. □

Now comes the crux of the proof. First, the structure of the graphs Γ_s comes into play again through the following easy lemma:

Lemma 4.2.7. *Let $d \geqslant 3$. Let $\Gamma = (V, E, \mathrm{ep})$ be a finite simple connected d-regular graph with m edges. There exist independent random variables $(\ell_\alpha)_{\alpha\in E}$ on $\mathcal{S}_\Gamma$, with values in the set of positive operators of rank 1 on the Hilbert space $L^2(\Gamma)$, such that*

$$\sum_{\alpha\in E} \ell_\alpha(s) = d + d\widetilde{M}_s.$$

Proof. Just spelling out the definition from Lemma 4.2.4, we have a decomposition of $d+d\widetilde{M}_s$ as the sum of $\ell_\alpha(s)$, where $\ell_\alpha(s)$ is the operator on $L^2(\Gamma)$ such that $\ell_\alpha(s)\varphi(x) = 0$ if x is not an extremity of α, and otherwise, if $\mathrm{ep}(\alpha) = \{x, y\}$, then

$$\ell_\alpha(s)\varphi(x) = \varphi(x) + s(\alpha)\varphi(y).$$

Let α be an edge with $\mathrm{ep}(\alpha) = \{x, y\}$. The image of $\ell_\alpha(s)$ is contained in the space of functions vanishing outside $\{x, y\}$. But note also that

$$\ell_\alpha(s)\varphi(y) = \varphi(y) + s(\alpha)\varphi(x) = s(\alpha)\cdot \ell_\alpha(s)\varphi(x),$$

from which we deduce that the image of $\ell_\alpha(s)$ is in fact the one-dimensional space generated by the function $\psi_{\alpha,s}$ mapping all $z \notin \mathrm{ep}(\alpha)$ to 0, and sending x to 1 and y to $s(\alpha)$. In fact, a few seconds of computations reveal that

$$\ell_\alpha(s)\varphi = d|\Gamma|\langle\varphi, \psi_{\alpha,s}\rangle\psi_{\alpha,s},$$

where $d|\Gamma|$ is the sum of the valencies in Γ. This computation shows that $\ell_\alpha(s)$ is also a positive linear map.

By definition, the random variables $s \mapsto s(\alpha)$ are independent for $\alpha \in E$, and it follows formally that the family $(\ell_\alpha)_{\alpha\in E}$ is also a family of independent random variables. □

With this lemma in hand, the key step is the next proposition (this could be phrased more generally so that it would have nothing to do with graphs anymore, as in [114, Th. 1.8]).

Proposition 4.2.8. *There exists $s \in \mathcal{S}_\Gamma$ such that*

$$\varrho^+\Bigl(\det\Bigl(X - \sum_\alpha \ell_\alpha(s)\Bigr)\Bigr) \leqslant \varrho^+\Bigl(\mathbf{E}\Bigl(\det\Bigl(X - \sum_\alpha \ell_\alpha\Bigr)\Bigr)\Bigr),$$

where $\varrho^+(f)$ denotes the largest real root of a polynomial $f \in \mathbf{R}[X]$.

The proof of this proposition ultimately relies on some properties of "real stable" polynomials, whose proofs can be found in Section C.4 of Appendix C.

Proof. We order the set of edges (arbitrarily) $E = \{\alpha_1, \ldots, \alpha_m\}$, and denote $\ell_i = \ell_{\alpha_i}$ for simplicity. We also write $\lambda_i = \mathbf{E}(\ell_i(s))$ for $1 \leqslant i \leqslant m$. We recall the notation $\boldsymbol{\mu}(\boldsymbol{u})$ for the mixed characteristic polynomial associated to a tuple of endomorphisms of $L^2(\Gamma)$ (see Definition C.5.1).

We will show below, by descending induction on j, where $1 \leqslant j \leqslant m$, that there exist $s \in \mathcal{S}_\Gamma$ such that

$$\varrho^+(\boldsymbol{\mu}(\ell_1(s), \ldots, \ell_j(s), \lambda_{j+1}, \ldots, \lambda_m)) \leqslant \varrho^+(\boldsymbol{\mu}(\ell_1(s), \ldots, \ell_{j-1}(s), \lambda_j, \ldots, \lambda_m)).$$

At the end of the induction, we get some $s \in \mathcal{S}_\Gamma$ such that

$$\varrho^+(\boldsymbol{\mu}(\ell_1(s), \ldots, \ell_m(s))) \leqslant \varrho^+(\boldsymbol{\mu}(\lambda_1, \ldots, \lambda_m)).$$

The mixed characteristic polynomial in the right-hand side of this inequality is equal to

$$\mathbf{E}\Big(\det\Big(X - \sum_{i=1}^{m} \ell_i\Big)\Big)$$

by Corollary C.5.8, and the one on the left-hand side is

$$\det\Big(X - \sum_{i=1}^{m} \ell_i(s)\Big)\Big)$$

by Corollary C.5.7, hence the result.

We now check the induction. The base case and the inductive step are in fact the same here. The endomorphism $\lambda_j = \mathbf{E}(\ell_j)$ is a convex combination of the endomorphisms $\ell_j(s)$ for $s \in \mathcal{S}_\Gamma$, so by Proposition C.5.5 (and the symmetry of the arguments of mixed characteristic polynomials), the mixed characteristic polynomial

$$\boldsymbol{\mu}(\ell_1(s), \ldots, \ell_{j-1}(s), \lambda_j, \ldots, \lambda_m)$$

is a convex combination of the polynomials

$$\boldsymbol{\mu}(\ell_1(s), \ldots, \ell_j(s), \lambda_{j+1}, \ldots, \lambda_m).$$

as s varies in $\mathcal{S}_\Gamma$. We can apply Proposition C.4.4, since Corollary C.5.9 shows that any convex combination of these polynomials is real stable. It follows that the quantity

$$\varrho^+(\boldsymbol{\mu}(\ell_1(s), \ldots, \ell_{j-1}(s), \lambda_j, \ldots, \lambda_m))$$

belongs to the convex hull of the numbers

$$\varrho^+(\boldsymbol{\mu}(\ell_1(s), \ldots, \ell_j(s), \lambda_{j+1}, \ldots, \lambda_m))$$

for $s \in \mathcal{S}_\Gamma$. In particular, for some s, we have

$$\varrho^+(\boldsymbol{\mu}(\ell_1(s), \ldots, \ell_{j-1}(s), \lambda_j, \ldots, \lambda_m)) \leqslant \varrho^+(\boldsymbol{\mu}(\ell_1(s), \ldots, \ell_j(s), \lambda_{j+1}, \ldots, \lambda_m)).$$

□

Proof of Theorem 4.2.5. By induction, it suffices to prove that if Γ is a finite bipartite simple d-regular Ramanujan graph, then there exists $s \in \mathcal{S}_\Gamma$ such that Γ_s is a (bipartite, simple, d-regular) Ramanujan graph. From the above proposition and Lemma 4.2.7, we know that there exists $s \in \mathcal{S}_\Gamma$ such that

$$\varrho^+(\det(X - (d + d\widetilde{M}_s))) \leqslant \varrho^+\Big(\mathbf{E}\Big(\det\Big(X - \sum_{\alpha} \ell_\alpha\Big)\Big)\Big),$$

or in other words

$$d + \varrho^+(\det(X - d\widetilde{M}_s)) \leqslant d + \varrho^+(\mathbf{E}(\det(X - d\widetilde{M}))).$$

The right-hand side is $d + \varrho^+(p(\Gamma))$ by Lemma 4.2.6, hence is $\leqslant d + 2\sqrt{d-1}$ by Godsil's Theorem (Theorem 3.6.8), and we conclude that $\varrho^+(\det(X - d\widetilde{M}_s)) \leqslant 2\sqrt{d-1}$. Since Γ_s is bipartite, its spectrum is symmetric, so that the spectral radius of $\widetilde{M}_s$ is $\leqslant 2\sqrt{d-1}$. Since Γ itself is a Ramanujan graph, Lemma 4.2.4 allows us to conclude that Γ_s is a Ramanujan graph. □

4.3. Cayley graphs of finite linear groups

For many of the applications of expander graphs that we will discuss in Chapter 5, the most important families of graphs are those arising from Cayley graphs of finite linear groups. Considerable progress has been made in recent years in understanding the expansion properties of these graphs.

There are two general, related, constructions of such families. We may consider a family (G_i) of finite groups, with $|G_i| \to +\infty$, given with symmetric generating subsets $S_i \subset G_i$ of fixed cardinality k, and the family $(\mathcal{C}(G_i, S_i))$. (For example, consider $\mathrm{SL}_2(\mathbf{F}_p)$ with generating set

$$\left\{ \begin{pmatrix} 1 & \pm(p-1)/2 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ \pm(p-1)/2 & 1 \end{pmatrix} \right\}$$

for p prime $\geqslant 3$). Alternatively, we may consider an *infinite* finitely generated group G, with a fixed symmetric finite set of generators $S \subset G$, and a family K_i of normal subgroups $K_i \lhd G$ with finite index $[G : K_i] \to +\infty$, and consider the relative Cayley graphs $\mathcal{C}(G/K_i, S)$. (For example, take $G = \mathrm{SL}_2(\mathbf{Z})$ with

$$S = \left\{ \begin{pmatrix} 1 & \pm 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ \pm 1 & 1 \end{pmatrix} \right\}$$

and K_p, for p prime, the kernel of reduction modulo p).

Note that when the quotient maps $G \to G_i = G/K_i$ are injective on S, with image $S_i \subset G_i$, then the action graphs are isomorphic to $\mathcal{C}(G_i, S_i)$, and the second family becomes a special case of the first type. Indeed, in the cases considered here, this will hold except for finitely many i, so it seems that we could restrict in principle without much loss to the first case, *except that in general this first case remains very mysterious.*

The question we wish to address is, quite generally: *under which type of condition is it true that a family of Cayley graphs as above is an expander family?*

Up to now, only two examples of sequences of Cayley graphs have appeared in this book, but these are not representative of the general case: the cycles C_m for $m \geqslant 2$ (which are 2-regular Cayley graphs of $G_m = \mathbf{Z}/m\mathbf{Z}$) or the graphs $G_n = \mathcal{C}(\mathfrak{S}_n, S_n)$ of Example 2.3.2 (4)). In both cases, we have seen that these are *not* expanders (though the second is not too far, being an esperantist family, by Example 3.5.7). But it turns out that, for many interesting sequences of "complicated" non-abelian groups, the answer to the question is positive, or conjectured to be so. For instance, in Section 4.4, we will give a fairly detailed sketch of the proof of the case $m = 3$ of the following theorem that combines results of Kazhdan and Margulis:

Theorem 4.3.1 (Kazhdan, Margulis). *Let $m \geqslant 3$ be an integer. For any finite symmetric generating set S of $\mathrm{SL}_m(\mathbf{Z})$, the family of relative Cayley graphs*

$$(\mathcal{C}(\mathrm{SL}_m(\mathbf{Z})/H, S))_{H \lhd \mathrm{SL}_3(\mathbf{Z})},$$

where H runs over all finite index normal subgroups of $\mathrm{SL}_m(\mathbf{Z})$, is an expander family.

This is an important and useful result, but the method of proof shows that the groups concerned are fairly special. In particular, it does not apply to $\mathrm{SL}_2(\mathbf{Z})$ (and indeed, the analogue statement is false for $\mathrm{SL}_2(\mathbf{Z})$).

On the other hand, in Chapter 6, we will prove the special case $m = 2$ of a theorem of Bourgain and Gamburd [12] and Varjú [116] concerning expansion of quotients of much more general subgroups of $\mathrm{SL}_m(\mathbf{Z})$. The price in this generalization is that we must restrict the family of quotients that are expanding.

Theorem 4.3.2 (Expansion in Zariski-dense subgroups of $\mathrm{SL}_m(\mathbf{Z})$). *Let $m \geqslant 2$ be an integer. Let $S \subset \mathrm{SL}_m(\mathbf{Z})$ be any finite symmetric subset and let G be the subgroup generated by S. Assume that G is Zariski-dense in SL_m. For prime numbers p, let $\Gamma_p = \mathcal{C}(\mathrm{SL}_m(\mathbf{F}_p), S)$ be the relative Cayley graph of the finite quotient group $\mathrm{SL}_m(\mathbf{F}_p)$ with respect to the reduction modulo p of the set S. Then there exists p_0 such that the family $(\Gamma_p)_{p \geqslant p_0}$ is an expander family.*

Remark 4.3.3. (1) The difference with Theorem 4.3.1 is that the previous result holds for *any* collection of finite index subgroups, not only for a specific family such as the kernels of reduction modulo primes, or even modulo any integer.

(2) In the special case of $\mathrm{SL}_2(\mathbf{Z})$, although Theorem 4.3.1 does not hold, there were important special cases of Theorem 4.3.2 that had been proved much earlier, and that were of great importance (both in terms of applications and of history). In particular, when $G = \mathrm{SL}_2(\mathbf{Z})$ itself, Theorem 4.3.2 follows from a crucial result of Selberg concerning the spectral gap of the hyperbolic Laplace operator and the comparison principle of Brooks and Burger (described in Section 5.4). This is related to Lubotzky's Property (τ), and we refer to [78, §4.4] for more discussion. The most general result along these lines is due to Clozel [29].

In the setting of Theorem 4.3.2, the condition that G is Zariski-dense has a very simple equivalent formulation: it means that for all primes p large enough, the reduction modulo p maps G *surjectively* to $\mathrm{SL}_m(\mathbf{F}_p)$. In terms of graphs, it therefore means that there exists p_0 such that $\mathcal{C}(\mathrm{SL}_m(\mathbf{F}_p), \mathbf{F}_p)$ is connected for all primes $p > p_0$, which is clearly a necessary condition for the expansion! It is also an elementary condition to check in many cases. For example, we obtain the following corollary:

Corollary 4.3.4 (Bourgain–Gamburd). *Let $k \geqslant 1$ be an integer, let*

$$S = \left\{ \begin{pmatrix} 1 & \pm k \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ \pm k & 1 \end{pmatrix} \right\} \subset \mathrm{SL}_2(\mathbf{Z}),$$

and for p prime, let S_p denote the image of S modulo p. Then the family of Cayley graphs $\mathcal{C}(\mathrm{SL}_2(\mathbf{F}_p), S_p)$ for $p \nmid k$ is an expander family.

For $k = 1$ or $k = 2$, this result was part of the special cases known classically that we mentioned above. However, for $k \geqslant 3$, this was a notorious open question until the results of Bourgain–Gamburd led to a general proof. The difference between these two cases is that S generates a *finite index* subgroup of $\mathrm{SL}_2(\mathbf{Z})$ for $k = 1$ or $k = 2$, but an *infinite index* subgroup if $k \geqslant 3$ (see Proposition B.1.3 for the proof). These groups are examples of what are now often known as "thin" subgroups of $\mathrm{SL}_2(\mathbf{Z})$ (see the book [15] for many aspects of the fascinating properties of these groups).

As we will also explain, a crucial step in the proof of Theorem 4.3.2 is a very important theorem that was proved by Helfgott [52] for SL_2 and SL_3 (and "almost" for SL_m), and then generalized by Pyber–Szabó [97] and Breuillard–Green–Tao [16] independently. We will state this precisely when needed, but it is helpful to state right now the following result, which turns out to be an immediate corollary.

Theorem 4.3.5 (Esperantism for Cayley graphs of $\mathrm{SL}_m(\mathbf{F}_p)$). *Let $m \geqslant 2$ be an integer. For any prime number p, let $S_p \subset \mathrm{SL}_m(\mathbf{F}_p)$ be a symmetric generating set of $\mathrm{SL}_m(\mathbf{F}_p)$, and assume that*

$$|S_p| \leqslant k$$

for some fixed $k \geqslant 1$. Then the family of Cayley graphs $(\mathcal{C}(\mathrm{SL}_m(\mathbf{F}_p), S_p))$ is an esperantist family, i.e., there exists $c > 0$ and $A \geqslant 0$ such that

$$\lambda_1(\mathcal{C}(\mathrm{SL}_m(\mathbf{F}_p), S_p)) \geqslant \frac{c}{(\log p)^A}.$$

Remark 4.3.6. The following example shows that for certain families of finite groups, there may exist families of generators for which the associated Cayley graphs are expanders, and others for which they are not. The Cayley graphs $G_n = \mathcal{C}(\mathfrak{S}_n, S_n)$ of Example 2.3.2 (4) are *not* expanders, but a remarkable result of Kassabov [63] shows that there exist (effectively computable) generating sets T_n of $\mathfrak{S}_n$, of bounded size as $n \to +\infty$, such that the Cayley graphs $(\mathcal{C}(\mathfrak{S}_n, T_n))$ do form an expander. Hence, for symmetric groups at least, the expansion property is not purely group-theoretical.

The restriction to subgroups of $\mathrm{SL}_m(\mathbf{Z})$ and to reduction modulo primes in Theorem 4.3.2, and to subgroups of $\mathrm{SL}_m(\mathbf{F}_p)$ for Theorem 4.3.5, is only present for the sake of simplicity. Much successful work was done from 2010 to around 2014 to generalize these results to other groups and to reduction modulo other integers, and the current state of knowledge goes much further. To state these extensions requires the use of the language of algebraic groups; the reader who is not familiar with the terminology need only know that the groups SL_m for $m \geqslant 2$ and Sp_{2g} for $g \geqslant 2$ satisfy the conditions of both theorems we will now state.

The general version of Theorem 4.3.5 was proved by Pyber and Szabó [97] and Breuillard–Green–Tao [16] independently. Precisely we have, again in the esperantist form of the statement:

Theorem 4.3.7. *Let $\mathbf{G}$ be a semisimple almost-simple linear algebraic group over $\mathbf{Q}$. For p prime, let $S_p \subset \mathbf{G}(\mathbf{F}_p)$ be a symmetric generating set of $\mathbf{G}(\mathbf{F}_p)$. Assume that there exists an integer $k \geqslant 1$ such that $|S_p| \leqslant k$ for all p. Then the family of Cayley graphs $(\mathcal{C}(\mathbf{G}(\mathbf{F}_p), S_p))$ is an esperantist family.*

For expanders, Salehi-Golsefidy and Varjú [99] proved the following remarkable result, where the last addition corresponding to SL_m is due to Bourgain and Varjú [14].

Theorem 4.3.8. *Let $\mathbf{G}$ be a semisimple almost-simple linear algebraic group over $\mathbf{Q}$. Let Γ be a Zariski-dense finitely generated discrete subgroup of $\mathbf{G}(\mathbf{Z})$. Let S be a finite symmetric generating set of Γ. There exists an integer $N \geqslant 1$ such that the family of relative Cayley graphs $\mathcal{C}(\mathbf{G}(\mathbf{Z}/n\mathbf{Z}), S)$ for n squarefree and coprime to N is an expander family.*

If $\mathbf{G} = \mathrm{SL}_m$, then the same holds for the family $\mathcal{C}(\mathbf{G}(\mathbf{Z}/n\mathbf{Z}), S)$ for all integers $n \geqslant 1$ coprime to N.

The case of $\mathbf{G} = \mathrm{SL}_2$ and squarefree n is due to Bourgain, Gamburd and Sarnak [13]. The remaining open problem in this area is to extend to all groups $\mathbf{G}$ the final statement of Bourgain and Varjú.

4.4. Property (T)

In the 1960's, Kazhdan [64] introduced an important property of locally compact groups, related to their unitary representations. A few years later, it was realized by Margulis that this led to examples of expanders from Cayley graphs of finite quotients of discrete groups satisfying Kazhdan's property.

We first explain this result of Margulis, taking a practical point of view where we specialize the definitions from the outset to discrete groups.

Definition 4.4.1 (Kazhdan's Property (T)). Let G be a discrete group. One says that G has Property (T) if there exists a finite subset S of G and a positive real number $\delta > 0$ such that for any unitary representation

$$\varrho\colon G \to \mathrm{U}(E),$$

where E is a Hilbert space, either there exists a non-zero vector $v \in E$ fixed by ϱ (i.e., $\varrho(g)v = v$ for all $g \in G$) or for all $v \neq 0$, we have

$$\max_{s\in S} \|\varrho(g)v - v\| \geqslant \delta\|v\|.$$

One then says that (S, δ) is a *Kazhdan pair* for G. If S is fixed, δ is said to be a *Kazhdan constant.*

The shorthand for this definition is: G has Property (T) if, whenever G acts linearly by unitary transformations on a Hilbert space, either it has a (non-zero) invariant vector, or it doesn't even have "almost" invariant vectors: any vector is moved by a non-trivial amount by some element of S.

Theorem 4.4.2 (Margulis). *Let G be a discrete group with Property* (T)*. Let (S, δ) be a Kazhdan pair for G such that S generates G. Let X be the family of all finite index normal subgroups of G. For all $H \in X$, we have*

$$h(\mathcal{C}(G/H, S)) \geqslant \delta^2.$$

In particular, if X contains elements of arbitrarily large index in G, the family of Cayley graphs of G/H, with respect to the image of S, is an expander family.

Proof. Let $H \in X$ and denote $\Gamma = \mathcal{C}(G/H, S)$. We consider the (finite-dimensional) Hilbert space $E = L^2(G/H)$ (i.e., the L^2-space for the Cayley graph Γ, where the inner product is defined by

$$\langle f_1, f_2, \rangle = \frac{1}{|G/H|} \sum_{x\in G/H} f_1(x)\overline{f_2(x)}$$

for $f_1,\ f_2\colon G/H \to \mathbf{C}$) and the homomorphism $G \to \mathrm{U}(E)$ defined by

$$\varrho(g)f(x) = f(xg)$$

(where we write $xg = x\pi(g)$ in terms of the projection $\pi\colon G \to G/H$). It is indeed elementary to check that ϱ is a homomorphism, and that $\varrho(g)$ is unitary.

Let E_0 be the orthogonal complement of the constant functions in E. Since the constant functions are invariant, under ϱ, and the representation is unitary, the subspace E_0 is also invariant. Thus ϱ induces a unitary representation $\varrho_0\colon G \to \mathrm{U}(E_0)$. Since S generates G, there is no function in E_0 invariant under the action of G.

Let now $W \subset G/H$ be a set of vertices of Γ with $|W| \leqslant \frac{1}{2}|G/H|$, and let

$$f = \mathbf{1}_W - \frac{|W|}{|G/H|}$$

be its normalized characteristic function. Then f belongs to E_0, and Kazhdan's Property (T) therefore implies that there exists $s \in S$ such that $\|\varrho(s)f - f\|^2 \geqslant \delta^2\|f\|^2$. However we have

$$\|\mathbf{1}_W\|^2 = \frac{|W|}{|G/H|}$$

and

$$\begin{aligned}\|\varrho(s)f-f\|^2=\|\varrho(s)\mathbf{1}_W-\mathbf{1}_W\|^2&=\frac{1}{|G/H|}\sum_{x\in G/H}|\mathbf{1}_W(xs)-\mathbf{1}_W(x)|^2\\&=\frac{1}{|G/H|}\Bigl(\sum_{\substack{x\in W\\xs\notin W}}1+\sum_{\substack{x\notin W\\xs\in W}}1\Bigr)\leqslant\frac{|\mathcal{E}(W)|}{|G/H|}.\end{aligned}$$

It follows that

$$|\mathcal{E}(W)|\geqslant\delta^2|W|.$$

Taking the minimum over W, we see that the Cheeger constant of Γ is $\geqslant\delta^2$. □

Remark 4.4.3. One can show (see, e.g., [7, Prop. 1.3.2]) that in fact, for a discrete group G with Property (T), any Kazhdan pair (S,δ) has the property that S generates G. Note that this implies that G is finitely generated; this fact was one of the motivating applications of Property (T), since Kazhdan was able to prove Property (T) for certain groups that were not previously known to be finitely generated. Conversely, for any finite generating set S of G, one can show that there exists $\delta>0$ (a Kazhdan constant for S) such that (S,δ) is a Kazhdan pair.

We will now give most steps of the proof of one of the first, and most important, results of Kazhdan, following the method of Shalom [104]; this implies Theorem 4.3.1 in the case $m=3$.

Theorem 4.4.4. *The group* $\mathrm{SL}_3(\mathbf{Z})$ *has Property* (T). *In particular, for any finite symmetric generating set* S *of* $\mathrm{SL}_3(\mathbf{Z})$, *the family of relative Cayley graphs*

$$(\mathcal{C}(\mathrm{SL}_3(\mathbf{Z})/H,S))_{H\lhd\mathrm{SL}_3(\mathbf{Z})},$$

where H *runs over all finite index normal subgroups of* $\mathrm{SL}_3(\mathbf{Z})$, *is an expander family.*

We will only discuss the proof of this result when looking at the defining property (Definition 4.4.1) restricted to *finite-dimensional* unitary representations of G; the proof of Theorem 4.4.2 shows that this is sufficient to obtain expander graphs. We will also work with the symmetric generating set

$$S=\{\mathrm{Id}\pm E_{i,j}\mid 1\leqslant i\neq j\leqslant 3\}$$

of $\mathrm{SL}_3(\mathbf{Z})$, where the matrices $E_{i,j}$ are integral matrices with all coefficients equal to 0 except for the (i,j)-th coefficient, that is equal to 1. We denote the generators

$$s_{i,j}=\mathrm{Id}+E_{i,j}$$

(that these matrices generate $\mathrm{SL}_3(\mathbf{Z})$ follows from the next theorem, but is of course a classical statement; see Lemma B.2.5). These are elements of infinite order; any element of $\mathrm{SL}_3(\mathbf{Z})$ of the form $s_{i,j}^m$ for some integer $m\in\mathbf{Z}$ is called an *elementary matrix* in $\mathrm{SL}_3(\mathbf{Z})$.

The proof of Theorem 4.4.4 that we give, for finite-dimensional representations, will be complete, except that we will rely on a non-trivial property of $\mathrm{SL}_3(\mathbf{Z})$.

Theorem 4.4.5 (Bounded elementary generation). *Any matrix* $g\in\mathrm{SL}_3(\mathbf{Z})$ *can be expressed as the product of at most* 48 *elementary matrices: there exists an integer* $k\leqslant 48$, *integers* $m_i\in\mathbf{Z}$ *and elementary matrices* $s_i\in S$ *for* $1\leqslant i\leqslant k$, *such that*

$$g=s_1^{m_1}\cdots s_k^{m_k}.$$

(Expressed differently: the diameter of the Cayley graph of $\mathrm{SL}_3(\mathbf{Z})$ with respect to the infinite generating set of *all* elementary matrices is at most 48).

This is a special case of more general results of Carter and Keller [24]; see also [7, §4.1, Th. 4.1.3] for the proof, which is to a large extent elementary (if somewhat complicated), although it does rely at one step on Dirichlet's Theorem on primes in arithmetic progressions ([7, Lemma 4.1.6]). A similar result (with a larger number of elementary matrices) is also true for $\mathrm{SL}_m(\mathbf{Z})$ for all $m \geqslant 3$, as proved in these references, but the corresponding property is not true for $\mathrm{SL}_2(\mathbf{Z})$.

On the other hand, a key step in Shalom's proof is a property of $\mathrm{SL}_2(\mathbf{Z})$, which is known as relative Property (T) (and goes back also to Kazhdan).

Theorem 4.4.6. *Let G be the semi-direct product $\mathbf{Z}^2 \rtimes \mathrm{SL}_2(\mathbf{Z})$, with $\mathrm{SL}_2(\mathbf{Z})$ acting on $\mathbf{Z}^2$ by $g \cdot x = gx$ for a matrix $x \in \mathrm{SL}_2(\mathbf{Z})$ and $x \in \mathbf{Z}^2$. Let*

$$F = \Big\{\Big(0, \begin{pmatrix} 1 & \pm 1 \\ 0 & 1\end{pmatrix}\Big), \Big(0, \begin{pmatrix} 1 & 0 \\ \pm 1 & 1\end{pmatrix}\Big), \Big(\begin{pmatrix} \pm 1 \\ 0\end{pmatrix}, 1\Big), \Big(\begin{pmatrix} 0 \\ \pm 1\end{pmatrix}, 1\Big), \Big\} \subset G.$$

Let $\varrho\colon G \to \mathrm{U}(E)$ be a finite-dimensional unitary representation of G such that E contains no non-zero vector invariant under $\mathbf{Z}^2$ viewed as a subgroup of G. Then for any $v \neq 0$ in E, we have

$$\max_{s \in F} \|\varrho(s)v - v\| \geqslant \frac{1}{1000}\|v\|.$$

Remark 4.4.7. The numerical value could easily be improved (for instance, in [7, Th. 4.2.2], it is 1/10).

We recall that the group G is, as a set, equal to $\mathbf{Z}^2 \times \mathrm{SL}_2(\mathbf{Z})$, with the group law

$$(x, g) \cdot (y, h) = (x + gy, gh).$$

It follows that

$$(x, g)^{-1} = (-g^{-1}x, g^{-1}),$$

and that

$$(x, g)(y, 1)(x, g)^{-1} = (g \cdot y, 1).$$

We begin by explaining the intuitive argument behind this theorem, which may seem mysterious at first, but has actually a very beautiful geometric interpretation. This arises from the fact that we may understand "geometrically" the representation ϱ using the decomposition of E as a unitary representation of the abelian subgroup $\mathbf{Z}^2$. By the results recalled in Section C.2, there exists a finite subset $T \subset \mathbf{T} = \widehat{\mathbf{Z}}^2 = (\mathbf{R}/\mathbf{Z})^2$ such that

$$E_\xi = \{v \in E \mid \varrho(m)v = e(\langle m, \xi\rangle)v \text{ for } m \in \mathbf{Z}^2\}$$

is non-zero if and only if $\xi \in T$, and such that we have an orthogonal decomposition

$$E = \bigoplus_{\xi \in T} E_\xi. \tag{4.5}$$

The assumption that E has no vector that is invariant under the action of $\mathbf{Z}^2$ is equivalent to the condition that $0 \notin T$. Moreover, the fact that ϱ is a representation of the *whole* group G, and not only of the normal subgroup $\mathbf{Z}^2$, is encoded in the fact that the subgroup $\mathrm{SL}_2(\mathbf{Z})$ of G permutes the spaces E_ξ, in the following sense. The group $\mathrm{SL}_2(\mathbf{Z})$ acts on $\mathbf{T}$ on the left by

$$\begin{pmatrix} a & b \\ c & d\end{pmatrix} \cdot (\xi_1, \xi_2) = (\xi_1, \xi_2) \cdot {}^t\begin{pmatrix} a & b \\ c & d\end{pmatrix}^{-1} = (\xi_1, \xi_2) \cdot \begin{pmatrix} d & -b \\ -c & a\end{pmatrix} = (d\xi_1 - b\xi_2, -c\xi_1 + a\xi_2),$$

and for any $g \in \mathrm{SL}_2(\mathbf{Z})$ and $\xi \in T$, the linear map $\varrho(g)$ is an isometry

$$E_\xi \to E_{g\cdot\xi}.$$

Indeed, it suffices to check that if $w \in E_\xi$, then $\varrho(g)w \in E_{g\cdot\xi}$. This is the case since for $m \in \mathbf{Z}^2$, we have

$$\varrho(m)\varrho(g)w = \varrho(g)\varrho(g^{-1}mg)w = \varrho(g)\varrho(g^{-1}\cdot m)w = e(\langle g^{-1}\cdot m, \xi\rangle)\varrho(g)w,$$

and

$$\langle g^{-1}\cdot m, \xi\rangle = \langle m, g\cdot\xi\rangle$$

with the definition of the action of G on $\mathbf{T}$ as above.

Now comes the crucial point. Let $v \neq 0$ be some vector in E, say with $\|v\| = 1$. We first consider

$$\varepsilon = \max_{s\in F_1} \|\varrho(s)v - v\|.$$

Since there is no non-zero vector in E invariant under $\mathbf{Z}^2$ by assumption, we have $\varepsilon > 0$. If $\varepsilon \geqslant 1/1000$, then of course we obtain the statement of the theorem for v. But, for any subset $Y \subset X$ and any $g \in \mathrm{SL}_2(\mathbf{Z}) \subset G$, we can “read off” a lower bound for $\|\varrho(g)v - v\|$ by computing the difference of the “mass” of v that is located in Y and the mass in the subset $g\cdot Y$.

Precisely, let

$$v = \sum_{\xi\in T} v_\xi$$

be the decomposition of v with $v_\xi \in E_\xi$. One defines the *spectral measure* μ associated to v as the measure on $\mathbf{T}$ such that

$$\mu(Y) = \sum_{\xi\in T\cap Y} \|v_\xi\|^2 \tag{4.6}$$

(this is a sum of Dirac masses, and by no means a complicated measure, because of the restriction to finite-dimensional representations; the general case involves a form of the spectral theorem for commuting unitary operators on an infinite-dimensional Hilbert space). Note that $\mu(\mathbf{T}) = 1$ since $\|v\| = 1$, so μ is a probability measure.

We claim that if $g \in \mathrm{SL}_2(\mathbf{Z})$, then we have

$$\|\varrho(g)v - v\| \geqslant \frac{1}{2}(\mu(Y) - \mu(g^{-1}\cdot Y)). \tag{4.7}$$

To see this, let P_Y denote the orthogonal projection of E with image the direct sums of E_ξ for $\xi \in Y$, so that

$$P_Y(v) = \sum_{\xi\in T\cap Y} v_\xi,$$

and hence $\|P_Y(v)\|^2 = \langle P_Y v, v\rangle = \mu(Y)$. The fact that $\varrho(g)$ is an isometry from E_ξ to $E_{g\cdot\xi}$ means that

$$(\varrho(g)v)_{g\cdot\xi} = \varrho(g)v_\xi,$$

or equivalently that

$$P_{g\cdot\xi}\varrho(g) = \varrho(g)P_\xi,$$

from which we deduce (by summing over $\xi \in Y \cap T$) that $P_{g\cdot Y} = \varrho(g)P_Y\varrho(g)^{-1}$. Hence

$$\mu(Y) - \mu(g^{-1}\cdot Y) = \langle P_Y v, v\rangle - \langle \varrho(g^{-1})P_Y\varrho(g)v, v\rangle.$$

Since ϱ is a unitary representation, we have $\langle P_Y v, \varrho(g)v\rangle = \langle \varrho(g)^{-1}P_Y v, v\rangle$, and we can rearrange this expression as

$$\begin{aligned}&\langle P_Y v, v - \varrho(g)v\rangle - \langle \varrho(g)^{-1}P_Y(\varrho(g)v - v), v\rangle \\ &\qquad\qquad \leqslant \|P_Y\|\|v - \varrho(g)v\| + \|\varrho(g)^{-1}P_Y\|\|\varrho(g)v - v\| \leqslant 2\|\varrho(g)v - v\|.\end{aligned}$$

The formula (4.7) reveals that in order to show that some $s \in F \cap \mathrm{SL}_2(\mathbf{Z})$ displaces v significantly, we may look for a subset Y of $\mathbf{T}$ for which $s\cdot Y$ is contained in, but "smaller" than Y, in the sense that the measure $\mu(Y - s\cdot Y)$ is quite large. Such behavior for a measure would be somewhat paradoxical if μ was invariant under the action of $\mathrm{SL}_2(\mathbf{Z})$, but the point is precisely that the measure is far from being invariant under this type of action. This is essentially a result of Burger [20]. The geometric intuition is quite easy to explain if we consider instead the action of $\mathrm{SL}_2(\mathbf{Z})$ on $\mathbf{R}^2$: take for instance

$$s = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

and consider the subset

$$\mathcal{Y} = \{(x_1, x_2) \in \mathbf{R}^2 \mid x_1x_2 \geqslant 0\}$$

(the union of the north-east and the south-west quadrants). Then since

$$s\cdot(x, y) = (x + y, y),$$

we get

$$s\cdot\mathcal{Y} = \{(x_1, x_2) \in \mathbf{R}^2 \mid x_1 \geqslant x_2 \geqslant 0 \text{ or } x_1 \leqslant x_2 \leqslant 0\} \subset Y,$$

which is intuitively "half" of $\mathcal{Y}$. So if the measure of "the other half" of $\mathcal{Y}$ is large, we will be precisely in the situation we want.

We now give the precise details, where we must essentially be careful because we have an action of $\mathrm{SL}_2(\mathbf{Z})$ on $\mathbf{T}$, and not on $\mathbf{R}^2$, so that "wrapping up" falsifies the relation $s\cdot Y \subset Y$ when Y is the analogue in $\mathbf{T}$ of $\mathcal{Y}$. To deal with this will require a careful game between the subgroups $\mathbf{Z}^2$ and $\mathrm{SL}_2(\mathbf{Z})$ of G.

Proof of Theorem 4.4.6. Let $v \in E$ be a given vector with $\|v\| = 1$. Since v is not fixed under $\mathbf{Z}^2$, which is generated by F_1, the number

$$\varepsilon = \max_{s\in F_1}\|\varrho(s)v - v\|$$

is strictly positive. The idea is to show, using (4.7), that if ε is too small, then some element $s \in F - F_1$ must move v by a non-trivial amount, so that the maximum of $\|\varrho(s)v - v\|$ over $s \in F$ will be large. One can optimize the argument but for simplicity we assume that $\varepsilon < 1/130$, and will then show that $\varepsilon \geqslant 1/1000$.

We decompose the vector v as

$$v = \sum_{\xi\in T} v_\xi$$

with $v_\xi \in E_\xi$. We define the associated spectral measure μ as in (4.6).

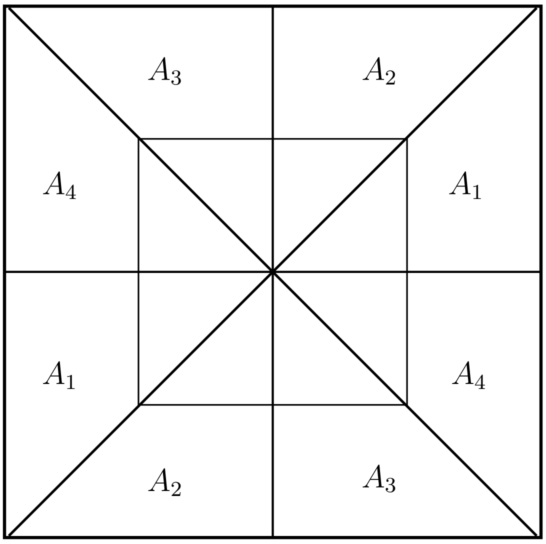
Since the decomposition (4.5) is orthogonal, and $\varrho(m)v_\xi = e(\langle m, \xi\rangle)v_\xi$, we have

$$1 = \|v\|^2 = \sum_{\xi\in T}\|v_\xi\|^2, \qquad \|\varrho(m)v - v\|^2 = \sum_{\xi\in T}|e(\langle m, \xi\rangle) - 1|^2\|v_\xi\|^2$$

for $m \in \mathbf{Z}^2$. Note that

$$|e(\langle m, \xi\rangle) - 1|^2 = 4\sin^2(\pi\langle m, \xi\rangle),$$

Figure 4.1. The four regions



and in particular for the elements

$$m_1 = \left(\begin{pmatrix}\pm 1\\ 0\end{pmatrix}, 1\right), \quad m_2 = \left(\begin{pmatrix}0\\ \pm 1\end{pmatrix}, 1\right),$$

of F_1, we have

$$|e(\langle m_1, \xi\rangle) - 1|^2 = 4\sin^2(\pi\xi_1), \qquad |e(\langle m_2, \xi\rangle) - 1|^2 = 4\sin^2(\pi\xi_2)$$

where we write each element of T in the form $\xi = (\xi_1, \xi_2) \in \mathbf{T}$. Hence, by definition of ε, we have

$$4\sum_{\xi\in T} \sin^2(\pi\xi_j)\|v_\xi\|^2 \leqslant \varepsilon^2$$

for $j = 1$, 2. If we represent each coordinate $\xi_j \in \mathbf{R}/\mathbf{Z}$ by a real number in the interval $]-\frac{1}{2}, \frac{1}{2}]$, one of which at least is non-zero (since $0 \notin T$), the fact that $\sin^2(\pi t)$ is not "small" for $1/4 \leqslant |t| \leqslant 1/2$ means that the corresponding vectors v_ξ must have rather small norms (here the choice of the constant 1/4 is convenient but not essential). This property is a precise analytic translation of the fact that v is not "almost" invariant under $\mathbf{Z}^2$.

We identify the dual group $(\mathbf{R}/\mathbf{Z})^2$ of $\mathbf{Z}^2$ with the square $X =]-\frac{1}{2}, \frac{1}{2}]$. In Figure 4.1, the small inner square corresponds to the set Y of those ξ with $|\xi_j| \leqslant 1/4$. Since

$$|\sin^2(\pi t)| \leqslant \frac{1}{4}$$

for $|t| \leqslant 1/4$, the vector

$$w = \sum_{\xi\in T\cap Y} v_\xi = P_Y v$$

satisfies

$$\|w\|^2 \geqslant (1-\varepsilon^2)\|v\|^2 = 1 - \varepsilon^2.$$

We denote by ν the spectral measure associated to w. Since $0 \notin T$, one of the following four sets A_i, whose union is $X - \{0\}$, must satisfy $\nu(A_i) \geqslant \|w\|^2/4$:

$$\begin{aligned}
A_1 &= \{(\xi_1, \xi_2) \in X \mid 0 \leqslant \xi_2 < \xi_1 \text{ or } \xi_2 < \xi_1 \leqslant 0\},\\
A_2 &= \{(\xi_1, \xi_2) \in X \mid 0 \leqslant \xi_1 < \xi_2 \text{ or } \xi_1 < \xi_2 \leqslant 0\},\\
A_3 &= \{(\xi_1, \xi_2) \in X \mid 0 \leqslant -\xi_1 < \xi_2 \text{ or } -\xi_1 < \xi_2 \leqslant 0\},\\
A_4 &= \{(\xi_1, \xi_2) \in X \mid 0 \leqslant -\xi_2 < \xi_1 \text{ or } -\xi_2 < \xi_1 \leqslant 0\}.
\end{aligned}$$

For each i, there exists $s \in F \cap \mathrm{SL}_2(\mathbf{Z})$ and $j \neq i$ such that

$$s((A_i \cup A_j) \cap Y) = A_i,$$

namely

$$\begin{aligned}
s &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \text{ and } A_j = A_2 \text{ if } i = 1,\\
s &= \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \text{ and } A_j = A_1 \text{ if } i = 2,\\
s &= \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \text{ and } A_j = A_4 \text{ if } i = 3,\\
s &= \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \text{ and } A_j = A_3 \text{ if } i = 4,
\end{aligned}$$

all these assertions being elementary.

We then argue as sketched before the proof: we have

$$\begin{aligned}
\|\varrho(s)w - w\| &\geqslant \frac{1}{2}\Big(\nu(A_i \cup A_j) - \nu(s \cdot (A_i \cup A_j))\Big)\\
&\geqslant \frac{1}{2}\nu(A_i) \geqslant \frac{1}{8}\|w\|^2 \geqslant \frac{1 - \varepsilon^2}{8}.
\end{aligned}$$

We conclude by observing that since $\|w - v\|^2 \leqslant \varepsilon^2$, we have

$$\|\varrho(s)w - \varrho(s)v\| = \|w - v\| \leqslant \varepsilon$$

and applying the triangle inequality gives

$$\|\varrho(s)v - v\| \geqslant \frac{1 - \varepsilon^2}{8} - 2\varepsilon,$$

which is $\geqslant 1/1000$ if $\varepsilon < 1/130$, for instance. □

Corollary 4.4.8. *Let $G = \mathbf{Z}^2 \rtimes \mathrm{SL}_2(\mathbf{Z})$. Let $F \subset G$ be as in Theorem 4.4.6. For any finite-dimensional unitary representation $\varrho \colon G \to \mathrm{U}(E)$ of G and for any $v \in E$, we have*

$$\max_{x \in \mathbf{Z}^2} \|\varrho(x)v - v\| \leqslant 2000 \max_{s \in F} \|\varrho(s)v - v\|.$$

Proof. We have an orthogonal direct sum $E = E_0 \oplus E_1$, where E_0 is the space of vectors which are $\mathbf{Z}^2$-invariant and E_1 its orthogonal complement. Because $\mathbf{Z}^2$ is normal in G, the space E_0 is in fact G-invariant (namely if $g \in G$ and $m \in \mathbf{Z}^2$, then for any $v \in E_0$, we have $\varrho(m)(\varrho(g)v) = \varrho(g)(\varrho(g^{-1}mg)v) = \varrho(g)v$ since $g^{-1}mg \in \mathbf{Z}^2$). Since ϱ is unitary, the space E_1 is also G-invariant.

Let v be a vector in E of norm 1 and write $v = v_0 + v_1$ with $v_i \in E_i$. Since E_1 has no $\mathbf{Z}^2$-invariant vectors, there exists $s \in F$ such that

$$\|\varrho(s)v - v\|^2 \geqslant \|\varrho(s)v_1 - v_1\|^2 \geqslant \delta^2\|v_1\|^2,$$

with $\delta = 1/1000$. For any $m \in \mathbf{Z}^2$, we have

$$\|\varrho(m)v - v\|^2 = \|\varrho(m)v_0 - v_0\|^2 + \|\varrho(m)v_1 - v_1\|^2 = \|\varrho(m)v_1 - v_1\|^2$$
$$\leqslant 4\|v_1\|^2 \leqslant \frac{4}{\delta^2} \max_{s \in F} \|\varrho(s)v - v\|^2.$$

□

The final lemma is quite general.

Lemma 4.4.9. *Let G be a topological group and ϱ a unitary representation of G on a Hilbert space E. If there exists a vector v of norm 1 in E such that*

$$\sup_{g \in G} \|\varrho(g)v - v\| \leqslant 1$$

then E has a non-zero invariant vector.

Proof. Let $C \subset E$ be the closure of the convex hull of the orbit $\varrho(G)v$. This is a convex subset of the Hilbert space E, so it contains a unique element w of minimal norm (this is a standard result, which is relatively clear intuitively for finite-dimensional spaces; for the general case, see, e.g., [11, V, p. 10, th. 1]). Since C is G-invariant, by uniqueness, we have $\varrho(g)w = w$. It only remains to prove that $w \neq 0$.

Let $g \in G$. We have

$$\mathrm{Re}(\langle \varrho(g)v, v\rangle) = 1 - \frac{1}{2}\|\varrho(g)v - v\|^2 \geqslant \frac{1}{2}$$

by assumption, hence taking convex combinations, we obtain

$$\mathrm{Re}(\langle x, v\rangle) \geqslant \frac{1}{2}$$

for all x which are convex combinations of elements of $\varrho(G)v$, and hence by continuity for all $x \in C$. Taking $x = w$, this shows that $w \neq 0$. □

We can now conclude.

Proof of Theorem 4.4.4. Let $\varrho\colon \mathrm{SL}_3(\mathbf{Z}) \to \mathrm{U}(E)$ be a finite-dimensional unitary representation of $\mathrm{SL}_3(\mathbf{Z})$ without invariant vectors. Let $v \in E$ be a vector of norm 1. Let $\varepsilon = \max_{s \in S} \|\varrho(s)v - v\|$.

Let i, j be distinct integers between one and three. It is not difficult to find a subgroup $G_{i,j}$ of $\mathrm{SL}_3(\mathbf{Z})$ isomorphic to $\mathbf{Z}^2 \rtimes \mathrm{SL}_2(\mathbf{Z})$ such that $s_{i,j}$ is an element of the corresponding subset $F \cap \mathbf{Z}^2$, so that all elementary matrices $s_{i,j}^k$ with $k \in \mathbf{Z}$ belong to the subgroup $\mathbf{Z}^2$ of $G_{i,j}$. For instance, for $i = 1$ and $j = 2$, one can take

$$G_{1,2} = \left\{ \begin{pmatrix} 1 & x & y \\ 0 & a & b \\ 0 & c & d \end{pmatrix} \mid ad - bc = 1, x,\ y \in \mathbf{Z} \right\}$$

(the reader should check that the matrix product does correspond to the semi-direct product). In addition, not only is $s_{i,j} \in F$, but more precisely the set of basic elementary matrices $s_{k,\ell}$ belonging to $G_{i,j}$ coincides precisely with the set F. By Corollary 4.4.8 applied to the restriction of ϱ to $G_{i,j}$, with i and j varying, we deduce that for any elementary matrix $s = s_{i,j}^k$ with $k \in \mathbf{Z}$, we have

$$\|\varrho(s)v - v\| \leqslant 2000 \max_{x \in S} \|\varrho(x)v - v\| = 2000\varepsilon.$$

Let now $g \in \mathrm{SL}_3(\mathbf{Z})$. Write

$$g = s_1 \cdots s_m$$

with $m \leqslant 48$ and each s_i an elementary matrix (by Theorem 4.4.5). Then

$$\begin{aligned}\|\varrho(g)v - v\| &\leqslant \sum_{i=0}^{k-1} \|\varrho(s_1 \dots s_{m-i})v - \varrho(s_1 \dots s_{m-i-1})v\| \\ &\leqslant 48 \max_{s \text{ elementary}} \|\varrho(s)v - v\| \leqslant 96000\varepsilon.\end{aligned}$$

If $\varepsilon \leqslant 1/96000$, this means that $\|\varrho(g)v - v\| \leqslant 1$ for all $g \in \mathrm{SL}_3(\mathbf{Z})$. But by Lemma 4.4.9, this contradicts the assumption that E has no $\mathrm{SL}_3(\mathbf{Z})$-invariant vectors. □

Remark 4.4.10. (1) One virtue of this approach to expander graphs (first visible in the work of Burger [20]) is that it leads to fully explicit expansion parameters, and that these are not as ridiculously small as those known from the Bourgain–Gamburd approach (compare with [70]). In fact, although we obtained here rather weak bounds, one can improve them significantly; Kassabov [62] refined considerably Shalom's method to get

$$\max_{s \in S} \|\varrho(g)v - v\| \geqslant \frac{1}{42\sqrt{3} + 860}$$

in Theorem 4.4.4 for unit vectors v in a unitary representation without invariant vectors.

This is not the best known bound, and a recent computation of Netzer and Thom [89], exploiting a striking result of Ozawa [92] that gives a "numerical" criterion for Property (T), shows that the spectral gap for the random walk on any finite quotient of $\mathrm{SL}_3(\mathbf{Z})$ is at least $1/72$. Even more recently, Kaluba, Nowak and Ozawa [59] (for $n = 5$) and Kaluba, Kielak and Nowak [60] (for $n \geqslant 6$) have used Ozawa's criterion to prove that the automorphism group of the free group on n generators has Property (T), which was not known before.

(2) We should mention another fascinating approach towards explicit bounds, that in fact motivated the work of Ozawa, and that is due to Żuk [122]. Given a finite symmetric generating set S of a discrete group G not containing 1, Żuk constructs an auxiliary simple graph Γ_S with vertices S and edges joining s_1 and s_2 if $s_1^{-1}s_2 \in S$. Assuming that Γ_S is connected (which may be achieved, if need be, by replacing S with $(S \cup S \cdot S) - \{1\}$), Żuk [122, Th. 1] proves that if *its* normalized spectral gap satisfies $\boldsymbol{\lambda}_1(\Gamma_S) > \frac{1}{2}$, then the group G has Property (T), and a Kazhdan constant (with respect to S) is

$$\frac{2}{\sqrt{3}}\Big(2 - \frac{1}{\boldsymbol{\lambda}_1(\Gamma_S)}\Big).$$

Żuk shows that this criterion is particularly useful in the study of certain models of random groups. He also gives the example of the exercise below to show that the value $\frac{1}{2}$ is best possible in the assumption.

Exercise 4.4.11. Let $G = \mathbf{Z}^2$ and $S = \{(1,0), (-1,0), (0,1), (0,-1), (1,1), (-1,-1)\}$. Construct the graph Γ_S in that case, and show that it is connected and satisfies $\boldsymbol{\lambda}_1(\Gamma_S) = 1/2$. Explain why G does not have Property (T).

Exercise 4.4.12. Let $n \geqslant 2$. Let $S_n = \{\mathrm{Id} + E_{i,j} \mid 1 \leqslant i \neq j \leqslant n\}$ be the generating set of elementary matrix of $\mathrm{SL}_n(\mathbf{Z})$. Consider the unitary representation of $\mathrm{SL}_n(\mathbf{Z})$ on $L^2(\mathbf{Z}^n - \{0\})$ by $\varrho(g)\varphi(m) = \varphi(g^{-1}m)$ for $g \in \mathrm{SL}_n(\mathbf{Z})$, $\varphi \in L^2(\mathbf{Z}^n - \{0\})$ and $m \in \mathbf{Z}^n$. Let φ be the characteristic function of the n canonical basis vectors in $\mathbf{Z}^n - \{0\}$. Show that

$$\max_{s \in S} \|\varrho(s)\varphi - \varphi\| \geqslant \sqrt{\frac{2}{n}}\,\|\varphi\|.$$

(This result is also due to Żuk, and is reported in [104, p. 149]; it shows that the best possible Kazhdan constant for the generating set of elementary matrices must depend on n, and tends to 0 with n).

CHAPTER 5

Applications of expander graphs

This chapter will present (sometimes merely in a survey style) some of the applications of expander graphs. This is very far from exhaustive – the reader will find many more applications, especially to combinatorics and theoretical computer science, in [54], and to "pure" mathematics in the books and lectures of Lubotzky [78], [79] and of Sarnak [101]. Our selection is guided in great part by personal interest (such as a sense of wonder at the Gromov–Guth distorsion theorem) and by a preference for topics which have not yet appeared in textbooks.

5.1. The Barzdin-Kolmogorov graph-embedding theorem

We now explain the first application of expander graphs, or precisely of the variant developed by Barzdin and Kolmogorov [5]. We refer to the paper of Gromov–Guth [48] and to Bergeron's survey [8] for more precise results in this direction.

Given a finite graph Γ, which we assume to be without loops or multiple edges, we define a *thick embedding* of Γ in $\mathbf{R}^3$ to be a pair (ι, j) such that ι is a map

$$\iota : V \longrightarrow \mathbf{R}^3$$

and j is a map $E \to C^1([0,1],\mathbf{R}^3)$, where $C^1([0,1],\mathbf{R}^3)$ is the space of C^1-functions from $[0,1]$ to $\mathbf{R}^3$, with the following properties:

- the map ι is injective;
- for any edge $\alpha \in E$ with extremities x_1 and x_2, the path $j(\alpha)$ goes from $\iota(x_1)$ to $\iota(x_2)$, or the opposite;
- the balls of center $\iota(x)$ and radius 1 are disjoint;
- for any distinct edges α and α', if there exists t and t' in $[0,1]$ such that the functions $\varphi = j(\alpha)$ and $\varphi' = j(\alpha')$ satisfy $|\varphi(t) - \varphi(t')| \leqslant 1/2$, then α and α' have a common extremity x, and $\varphi(t)$ and $\varphi(t')$ are at distance $\leqslant 1/2$ of $\iota(x)$.

Intuitively, we view this data as a way of embedding the graph in $\mathbf{R}^3$: each vertex x maps to $\iota(x)$, and each edge maps to a smooth curve joining the images of its extremities. The last two conditions above mean that the images of these segments are never close to each other, except in the neighborhood of a common extremity. We interpret this as giving a precise manner in which we can "draw" the graph faithfully in space, with enough space between vertices and edges.

We first can check that such an embedding always exists:

Proposition 5.1.1. *Let Γ be a finite graph with maximal degree at most 6 and no loops or multiple edge. Then Γ admits a thick embedding.*

Proof. It is fairly easy to convince oneself that this is correct, and we only sketch a proof. Here is a way to do it: map the vertices arbitrarily to points in the horizontal plane $z = 0$ which have coordinates multiples of 4 (say), so that they are well-separated. Then add edges one by one. Since the degree is at most 6, we can "start" each path in a different direction in $\mathbf{R}^3$. For each new edge α, we must ensure that the "main part" of the path $\varphi = j(\alpha)$ is contained in $\mathbf{R}^3$ minus the finite union of the points at distance $\leqslant 1$ of the previous paths. This is certainly possible since this space is pathwise connected. □

Given a thick embedding ι, we define its *radius* $r(\iota)$ to be the infimum of those real numbers $r \geqslant 0$ such that, for some $x \in \mathbf{R}^3$, the image of ι is contained in the ball of radius r centered at x. In particular, the smallest volume of a ball in $\mathbf{R}^3$ in which Γ can be "drawn" using a thick embedding is $\frac{4}{3}\pi r(\iota)^3$. The theorem of Barzdin and Kolmogorov is concerned with how large $r(\iota)$ should be.

THEOREM 5.1.2 (Barzdin–Kolmogorov). *Let Γ be a finite graph with valency at most 6 at each vertex.*

(1) *Let ι be a thick embedding of Γ. There exists a constant $c > 0$, independent of Γ and ι, such that the radius of ι is at least $c\sqrt{h(\Gamma)|V|}$.*

(2) *There exists a constant $c' > 0$, independent of Γ, such that Γ admits a thick embedding with radius $\leqslant c'\sqrt{|V|}$.*

In particular, if $(\Gamma_n)_{n\geqslant 1}$ is a family of expander graphs with degree at most 6, the optimal order of magnitude of the radius of a thick embedding of Γ_n as n tends to infinity is $\sqrt{|V_n|}$. For this result to be of interest, it is of course essential to know that expander graphs do exist, but an explicit construction is not needed.

PROOF. We prove only (1), and refer to [8] and [5] for (2). We may certainly assume that $|V| \geqslant 3$. We assume first that ι is such that the z-coordinate of all points $\iota(x)$, for $x \in V$, are distinct. Let ι be a thick embedding of Γ and let $x_0 \in \mathbf{R}^3$ and $r > 0$ be such that the image of V is contained in the ball of radius r around x_0. Let z_0 be a real number such that at least half of the vertices x of Γ are such that the third coordinate of $\iota(x)$ is $\leqslant z_0$, and at least half are such that the third coordinate is $\geqslant z_0$ (z_0 is a median of the third coordinate function on V).

Let $V_1 \subset V$ be the set of those vertices with third coordinate $\leqslant z_0$. We have $\frac{1}{2}|V|-1 \leqslant |V_1| \leqslant \frac{1}{2}|V|$. By definition of the expansion constant, there are at least $h(\Gamma)|V_1| \geqslant \frac{1}{4}h(\Gamma)|V|$ edges in Γ with one extremity in V_1 and one extremity in V_2. For each such edge α, with extremities x_1 and x_2, the continuous path $j(\alpha)$ joining $\iota(x_1)$ and $\iota(x_2)$ intersects the horizontal plane with equation $z = z_0$ in at least one point. Let $f(\alpha)$ be one arbitrarily chosen such intersection point. All the points of the form $f(\alpha)$ belong to a disc in the plane $z = z_0$ with radius $\leqslant r$. On the other hand, by the last condition in the definition of a thick embedding, if α and α' are distinct edges, then the balls with radius 1/2 and centers $f(\alpha)$ and $f(\alpha')$ are disjoint, unless α and α' have a common extremity. Since the maximal degree is $\leqslant 6$, this means that there are at least $\frac{1}{24}h(V)|V|$ disjoint discs of radius 1/2 contained in a disc of radius r in the plane $z = z_0$. Hence

$$\frac{1}{96}h(\Gamma)|V| \leqslant \pi r^2.$$

Consider now the general case where no assumption on the z-coordinates of the points $\iota(x)$ is made. Then rotate the thick embedding slightly; this doesn't change the radius of the embedding, but one can see that some rotated embedding will have the required property. (Dually, replace the linear form $(x, y, z) \mapsto z$ by another one that is injective on the vertices). □

5.2. Error reduction in probabilistic algorithms

In this section, which is based on [54, Section 3.3], we present one application of expander graphs in theoretical computer science. Our exposition will not be completely formal, since we will not give a rigorous definition of "algorithm" or "computer", but the basic ideas should be clear. Moreover, this gives further motivation for particular problems concerning expanders, and the main technical tool that is used is clearly relevant in other contexts.

Informally, an *algorithm* A with inputs I, outputs O and running time

$$r : I \longrightarrow [0, +\infty[$$

is a (deterministic) computer program which takes as input an element $i \in I$, and (always) ends its run by printing an (output) element of O, which we denote $\mathsf{A}[i]$, and takes time $r(i)$ to do so (for instance, A can be defined as a Lisp function [85], with I and O defined as sets of arbitrarily long finite binary strings, and running time the number of elementary operations used in the computation.)

Below, each element of I will have a well-defined "length" $\ell(i)$, corresponding intuitively to the number of binary digits needed to encode i, and $r(i)$ will be a function of $\ell(i)$ only. For instance, if $I = \mathbf{Z}$, then we take $\ell(i)$ to be the number of binary digits used in expressing i. We will then be interested in *polynomial-time algorithms*, which are those for which

$$r(i) \leqslant c_1 \ell(i)^A$$

for some constants $c_1 \geqslant 0$ and $A \geqslant 0$ which are independent of $i \in I$.

Remark 5.2.1. In principle, the running time should include the time needed to "output" the value $\mathsf{A}[i]$. However, we will consider algorithms with $O = \{0, 1\}$ for which such a distinction is irrelevant.

Example 5.2.2. Given a subset $M \subset \mathbf{Z}$, one can ask for a fast algorithm A_M which "recognizes" M, i.e., which has input set $I = \mathbf{Z}$, output $O = \{0, 1\}$, runs in polynomial time (relative to the number of digits of i) and is such that

$$\mathsf{A}_M[i] = \begin{cases} 1 & \text{if } i \in M, \\ 0 & \text{if } i \notin M. \end{cases}$$

A natural number-theoretic example is the set M of prime numbers. In that case, the naive "trial-division" algorithm certain has the right output, but is *not* fast: its running time satisfies $r(i) \leqslant \sqrt{i} \asymp 2^{\ell(i)/2}$.

Trying to find a polynomial-time algorithm to recognize the set of primes should convince the reader that this is not a trivial problem, and the discovery of such an algorithm by Agrawal, Kayal and Saxena [3] is rather recent. However, if one allows a bit of luck to come into the game, and allows some possibility of error, one can work somewhat quicker. These relaxations of the rules lead to the notion of probabilistic (or randomized) algorithms.

We consider these only for algorithms which are supposed to compute a function

$$f : I \longrightarrow \{0, 1\}$$

where I is given with a size function ℓ as above, taking non-negative integer values. We write I_m for the set of $i \in I$ of size m.

A *randomized algorithm* for the computation of f is an algorithm $\tilde{\mathsf{A}}$ with input set

$$\tilde{I} = \bigcup_{i \in I} (\{i\} \times \Omega_i)$$

such that

(1) The auxiliary non-empty finite sets Ω_i (the sets of "random bits") are also given with a size function $\ell(\omega)$ such that $\ell(\omega) \leqslant c_2 \ell(i)^B$, where $c_2 \geqslant 0$ and $B \geqslant 0$ are fixed;

(2) The algorithm $\tilde{\mathsf{A}}$ runs in polynomial time relative to the size

$$\ell(i, \omega) = \ell(i) + \ell(\omega), \qquad \omega \in \Omega_i,$$

and hence, for any $i \in I$ and any choice of $\omega \in \Omega_i$, $\tilde{\mathsf{A}}[i, \omega]$ runs in polynomial time in terms of $\ell(i)$;

(3) For all $i \in I$ such that $f(i) = 1$, we have

$$\tilde{\mathsf{A}}[i, \omega] = 1$$

for arbitrary $\omega \in \Omega_i$;

(4) For all $i \in I$ such that $f(i) = 0$, the algorithm *may* return the wrong answer 1 for certain choices of "random bits" $\omega \in \Omega_i$, *but* at most with a fixed probability $p < 1$, i.e., if $f(i) = 0$, we have

$$\frac{1}{|\Omega_i|} |\{\omega \in \Omega_i \mid \tilde{\mathsf{A}}[i, \omega] = 1\}| \leqslant p. \tag{5.1}$$

The probability p is called the *error rate*, or *failure rate* of the probabilistic algorithm. Intuitively, the idea is to attempt to compute $f(i)$ by selecting $\omega \in \Omega_i$ uniformly at random[1] and running $\tilde{\mathsf{A}}[i, \omega]$. If the answer is 0, it follows that $f(i) = 0$, by Property (3), but if the answer is 1, we can only take this as a hint that $f(i)$ *could* be equal to 1. By (5.1), this hint will be the right answer with probability at least $1 - p$.

Example 5.2.3 (The Solovay-Strassen primality test). A good practical example should clarify a lot what is happening. Consider once more the problem of finding a good primality test. It turns out that a probabilistic polynomial-time algorithm for this question, with error rate $\leqslant 1/2$, can be devised quite easily: this is the Solovay-Strassen test [106], which goes back to 1977.

This test starts with the definition of the *Legendre symbol* modulo an odd prime p, which is the function

$$\begin{cases} \mathbf{Z}/p\mathbf{Z} & \longrightarrow & \{-1, 0, 1\} \\ a & \mapsto & \left(\frac{a}{p}\right) \end{cases}$$

where

$$\left(\frac{a}{p}\right) = \begin{cases} 0 & \text{if } a = 0 \\ 1 & \text{if there exists } y \in \mathbf{F}_p \text{ such that } x = y^2 \\ -1 & \text{otherwise.} \end{cases}$$

The first crucial ingredient of the test is the fact, due to Euler, that the Legendre symbol, for a fixed odd prime p, can be computed by means of the congruence

$$\left(\frac{a}{p}\right) \equiv a^{(p-1)/2} \,(\mathrm{mod}\, p). \tag{5.2}$$

[1] This assumes that this choice can be done efficiently, which may in practice well be another non-trivial problem...

Note that the right-hand side can be computed – using repeated squarings – in polynomial time in terms of the size of p (uniformly for all $a \in \mathbf{Z}/p\mathbf{Z}$).

As a next step, the Legendre symbol is extended to the Jacobi symbol modulo an *odd* integer $n \geqslant 1$, which is defined by

$$\Big(\frac{m}{n}\Big) = \prod_{i=1}^{k} \Big(\frac{m}{p_i}\Big)^{v_i}$$

if $n = p_1^{v_1} \cdots p_k^{v_k}$ is the factorization of n into distinct prime powers (with $p_i \geqslant 3$) and m is any integer. The values of the Jacobi symbol are still among $\{-1, 0, 1\}$. The following is a deep fact, in some sense the founding statement of algebraic number theory:

Theorem 5.2.4 (Quadratic reciprocity law). (1) *For any odd positive integers n and m which are coprime, we have*

$$\Big(\frac{n}{m}\Big)\Big(\frac{m}{n}\Big) = (-1)^{(n-1)(m-1)/4}.$$

(2) *For any odd integer n, we have*

$$\Big(\frac{-1}{n}\Big) = (-1)^{(n-1)/2},$$

and

$$\Big(\frac{2}{n}\Big) = (-1)^{(n^2-1)/8}.$$

For a proof, see, e.g., [57, Prop. 5.2.2] or [102, I.3.3 or I.3, Appendix], or [58, Th. 3.5]...

Note the following corollary, which is of great importance for this example:

Corollary 5.2.5 (Computation of Jacobi symbols). *There is a deterministic*[2] *polynomial time algorithm* J *with inputs* $I = \mathbf{Z} \times \{$*odd integers*$\}$ *and outputs* $O = \{-1, 0, 1\}$ *such that* $\mathsf{J}(m, n)$ *is the Jacobi symbol of* m *modulo* n.

Sketch of the proof. The idea is to use the fact that the Jacobi symbol depends only on m modulo n, select a representative n with $|n| \leqslant m/2$, and use quadratic reciprocity to "switch" n and m, then reduce m modulo n and repeat. The "supplementary laws" computing the Jacobi symbols $\big(\frac{-1}{n}\big)$ and $\big(\frac{2}{n}\big)$ are used to get rid of the sign of m and its 2-power component before switching... (Coprimality of n and m can be tested also in polynomial time using the Euclidean algorithm for computing the greatest common divisor of two integers.) □

Exercise 5.2.6. Finish the proof of this corollary.

We can now define the Solovay-Strassen primality test S: we take as input set

$$I = \bigcup_{i \text{ odd integer}} (\{i\} \times (\mathbf{Z}/i\mathbf{Z})^{\times}),$$

and define $\mathsf{S}[i, a]$, for $i \geqslant 1$ an odd integer and $a \in (\mathbf{Z}/i\mathbf{Z})^{\times}$, to be 1 if

$$\Big(\frac{a}{i}\Big) \equiv a^{(i-1)/2} \pmod{i}.$$

and 0 otherwise. According to Corollary 5.2.5, this algorithm can be run in polynomial time with respect to the size (number of binary digits) of i. We now check that it satisfies the properties for a randomized primality-testing algorithm.

[2] This just means that it is not probabilistic...

First of all, if i is an odd prime number, then by Euler's formula (5.2), we have $\mathsf{S}[i,a]=1$ for all $a\in(\mathbf{Z}/i\mathbf{Z})^\times$, so that the algorithm never returns a wrong answer for prime inputs. It remains to estimate the error rate, i.e., to bound from above the ratio

$$\frac{1}{|(\mathbf{Z}/i\mathbf{Z})^\times|}|\{a\in(\mathbf{Z}/i\mathbf{Z})^\times \mid \Bigl(\frac{a}{i}\Bigr)\equiv a^{(i-1)/2}\,(\mathrm{mod}\, i)\}|.$$

To do this, we claim that if i is not prime, the set

$$B=\{a\in(\mathbf{Z}/i\mathbf{Z})^\times \mid \Bigl(\frac{a}{i}\Bigr)\equiv a^{(i-1)/2}\,(\mathrm{mod}\, i)\}$$

is a proper *subgroup* of $(\mathbf{Z}/i\mathbf{Z})^\times$. If that is the case, then

$$\frac{|B|}{|(\mathbf{Z}/i\mathbf{Z})^\times|}=\frac{1}{[(\mathbf{Z}/i\mathbf{Z})^\times:B]}\leqslant\frac{1}{2},$$

so that the Solovay-Strassen test gives the wrong answer, if i is not prime, at most with probability $1/2$.

As for the claim, the fact that B is a subgroup is easy (and is valid even if i is prime), because both sides of the congruence are multiplicative functions of i. What needs some care is the proof that $B\neq(\mathbf{Z}/i\mathbf{Z})^\times$ if i is *not* prime (which is really the point, since Euler's formula precisely means that $B=(\mathbf{Z}/i\mathbf{Z})^\times$ when i is prime).

Because of the Chinese Remainder Theorem, we may assume that $i=p^v$ is a power of a prime, with $v\geqslant 2$. We recall that, since p is odd, the group $(\mathbf{Z}/p^v\mathbf{Z})^\times$ is cyclic of order p^v-p^{v-1} (see, e.g., [57, Ch. 4, th. 2]). Thus, if we take for a a generator of $(\mathbf{Z}/p^v\mathbf{Z})^\times$, we have $B=(\mathbf{Z}/p^v\mathbf{Z})^\times$ if and only if $a\in B$. First, if v is even, the Jacobi symbol $(\frac{a}{p^v})$ is equal to 1, but $(p^v-1)/2$ is clearly not a multiple of the order p^v-p^{v-1} of a, and therefore $a^{(p^v-1)/2}\neq 1$. On the other hand, if v is odd (not equal to 1), the Jacobi symbol is -1 (because a is not a square modulo p), and since $p^v-1\equiv p^{v-1}-1\,(\mathrm{mod}\,(p^v-p^{v-1}))$, which is non-zero, we have

$$a^{(p^v-1)}\neq 1$$

in $(\mathbf{Z}/p^v\mathbf{Z})^\times$, and a fortiori $a\notin B$.

The way expander graphs intervene in this discussion is to provide a tool to address the following question: given a probabilistic algorithm $\tilde{\mathsf{A}}$ to compute $f : I\rightarrow\{0,1\}$, with error rate $p<1$, can one efficiently diminish the probability of error, ideally to an arbitrarily small error rate $\varepsilon>0$?

An immediate answer comes to mind: if one simply runs $\tilde{\mathsf{A}}$ multiple times with a fixed $i\in I$, say t times, with independent (uniformly distributed) random bits, and output 1 if and only if all runs return 1, then the error rate will drop to p^t. Thus, if the permissible error ε, with $0<\varepsilon\leqslant p$, is fixed beforehand, taking $t=\lceil\frac{\log\varepsilon}{\log p}\rceil$ will give the desired reduction.

Formally, this means that we consider the algorithm $\tilde{\mathsf{A}}_t$ with inputs

$$I_t=\bigcup_{i\in I}(\{i\}\times\Omega_i^t)$$

and

$$\tilde{\mathsf{A}}_t[i,\omega_1,\ldots\omega_t]=\begin{cases}1 & \text{if } \tilde{\mathsf{A}}[i,\omega_1]=\cdots=\tilde{\mathsf{A}}[i,\omega_t]=1\\ 0 & \text{otherwise,}\end{cases}$$

which still uses a polynomially bounded amount of extra randomness. We still have $\tilde{\mathsf{A}}_t[i,\boldsymbol{\omega}]=1$ for any $i\in I$ such that $f(i)=1$ (by definition of $\tilde{\mathsf{A}}$), and the error rate for i

with $f(i)=0$ is given by

$$\frac{1}{|\Omega_i^t|}|\{(\omega_1,\ldots,\omega_t)\in\Omega_i^t\ \mid\ \tilde{\mathsf{A}}[i,\omega_j]=1\text{ for all }1\leqslant j\leqslant t\}|=$$
$$\Bigl(\frac{1}{|\Omega_i|}|\{(\omega_1,\ldots,\omega)\in\Omega_i\ \mid\ \tilde{\mathsf{A}}[i,\omega]=1\}|\Bigr)^t\leqslant p^t.$$

This looks like a good solution, and it certainly is in an abstract realm where randomness is cheap and perfect independent samplings of a uniform distribution on Ω_i is easy. However, neither of these is clear or in fact true in practice. We consider the first problem only: where one run of $\tilde{\mathsf{A}}$ requires intuitively $\ell(\omega)\leqslant c_2\ell(i)^B$ random bits, we need t times as many for $\tilde{\mathsf{A}}_t$. The question is whether one can do better, i.e., bring the error rate below any given value $\varepsilon>0$ using *fewer* extra random bits.

Here is a solution involving expander graphs. Given $\tilde{\mathsf{A}}$ as above, we define a new probabilistic algorithm $\mathsf{E}\tilde{\mathsf{A}}_t$ using the following procedure, which we first describe intuitively before proceeding to formalize it to some extent.

We start with an input $i\in I$.

Step 1. Construct a graph Γ_i with vertex set Ω_i, which is connected, d-regular for some d, non-bipartite, and has equidistribution radius $\varrho_i=\varrho_{\Gamma_i}$;

Step 2. Pick uniformly at random an initial vertex $\omega_0\in\Omega_i$;

Step 3. Start a random walk on Γ_i with initial vertex ω_0, say $(X_n^{(\omega_0)})_{n\geqslant 0}$;

Step 4. Return 1 if

$$\tilde{\mathsf{A}}[i,X_0^{(\omega_0)}]=\cdots=\tilde{\mathsf{A}}[i,X_t^{(\omega_0)}]=1,$$

and 0 otherwise.

In other words, intuitively, we replace the t independent choices of uniformly distributed ω_i's of the previous discussion by $t+1$ choices where one is picked completely randomly, but the others are obtained by the first steps of a random walk on a graph with vertex set Ω_i. The amount of randomness that is necessary to run this algorithm is only the amount needed for the choice of the first vertex ω_0, and for throwing t times a d-sided dice to perform the random walk. (Formally, we take new auxiliary "random bits" $\Omega_i\times\{1,\ldots,d\}^t$, assuming we also fix with Γ_i some explicit bijections from $\{1,\ldots,d\}$ to the set of neighbors of any vertex $\omega\in\Omega_i$.) If d is small, this is significantly less randomness than required for independent trials of $\tilde{\mathsf{A}}$.

It is clear that this algorithm, started with input $i\in I$, will again return 1 whenever $f(i)=1$. Also, since it reduces to $\tilde{\mathsf{A}}$ when $t=0$, it has error rate at most p. Is it significantly smaller? To answer this, note that if $f(i)=0$ and

$$B=\{\omega\in\Omega_i\ \mid\ \tilde{\mathsf{A}}[i,\omega]=1\}\subset\Omega_i \tag{5.3}$$

is the set of random choices of ω for which the original probabilistic algorithm gives the wrong answer, the failure rate for $\mathsf{E}\tilde{\mathsf{A}}$ is given by

$$\mathbf{P}(X_0\in B,\ X_1\in B,\ldots,\ X_t\in B)$$

where (X_n) is now a random walk on Γ_i with uniformly distributed initial step X_0.

To bound this probability, we have the following general result of Ajtai-Komlós-Szemerédi and Alon-Feige-Wigderson-Zuckerman, which has independent interest:

Proposition 5.2.7 (Decay of "confinement" probabilities). *Let $\Gamma=(V,E,\mathrm{ep})$ be a finite connected d-regular, non-bipartite graph, and let $B\subset V$ be a subset of vertices. If (X_n) is the random walk on Γ with uniformly distributed initial step X_0, we have*

$$\mathbf{P}(X_j\in B \textit{ for all } j,\ 0\leqslant j\leqslant t)\leqslant(\mu_\Gamma(B)+\varrho_\Gamma)^t$$

for all $t\geqslant 0$, where $\mu_\Gamma(B)=|B|/|V|$ and ϱ_Γ the equidistribution radius of Γ.

Before proving this, let's see how it applies to the study of the algorithm $\mathsf{E\tilde{A}}_t$. In that case, we apply the proposition to the graph Γ_i of Step 1, with B given by (5.3). By construction, we have $\mu_{\Gamma_i}(B)=|B|/|\Omega_i|\leqslant p$, so we obtain the upper bound

$$(p+\varrho_i)^t$$

for the error rate. This is of course trivial unless $p+\varrho_i<1$. More precisely, this provides exponential decay of the error rate as a function of t, comparable with the independent model $\tilde{\mathsf{A}}_t$, *provided* we have $p+\varrho_i<\varrho_0<1$ where ϱ_0 is independent of i. This certainly requires that the family (Γ_i) be an expander family, but this is not enough if p is relatively large (say if $p=1/2$, as in the Solovay-Strassen primality test).

There is a work-around for this, if one knows an explicit bound ϱ on the equidistribution radius for the family (Γ_i). In that case, one can first replace $\tilde{\mathsf{A}}$ with $\mathsf{B}=\tilde{\mathsf{A}}_s$ for some s such that $p^s+\varrho=\varrho_0<1$, and then construct the corresponding probabilistic algorithm EB_t. Since s will be fixed, this provides then a procedure for reducing arbitrarily the error rate of $\tilde{\mathsf{A}}$ while using much fewer random bits as the corresponding use of independent samples.

There is still one point we haven't addressed, however: the algorithm $\mathsf{E\tilde{A}}_t$ is only effective (i.e., runs in polynomial time) if we also have a deterministic polynomial-time algorithm to construct the graphs Γ_i. Here a probabilistic construction, or a construction of a family where the expansion parameters are not explicitly known, will not be sufficient! A good choice is given by the Zig-zag construction (see [54, §9]), which has the required properties.

Proof of Proposition 5.2.7. Let ϖ be the probability to compute. By direct expansion, we have

$$\varpi=\frac{1}{|V|}\sum_{x_0\in B}\sum\sum_{x_1,\ldots,x_t\in B}\mathbf{P}(X_1=x_1\mid X_0=x_0)\,\mathbf{P}(X_2=x_2\mid X_0=x_0,\ X_1=x_1)$$
$$\cdots\mathbf{P}(X_t=x_t\mid(X_0,\ldots,X_{t-1})=(x_0,\ldots,x_{t-1})),$$

and the Markov property of the random walk (3.8) reduces this to

$$\varpi=\frac{1}{|V|}\sum_{x_0\in B}\sum\sum_{x_1,\ldots,x_t\in B}P(x_0,x_1)\cdots P(x_{t-1},x_t)$$

where the transition probability $P(x,y)$ is defined in (3.9). Now we claim that this can also be expressed, in terms of the Markov operator M, as the inner product

$$\varpi=\langle\mathbf{1}_B,(MP_B)^t\mathbf{1}\rangle,$$

where $P_B:L^2(\Gamma)\to L^2(\Gamma)$ is the orthogonal projection on the space of functions vanishing outside B, which is given by $P_B\varphi=\mathbf{1}_B\varphi$, and $\mathbf{1}$ is here just another notation for the constant function 1. To see this from the previous formula, it is enough to prove that for any $x\in V$, we have

$$(MP)^t\varphi(x)=\sum\sum_{x_1,\ldots,x_t\in V}P(x,x_1)\cdots P(x_{t-1},x_t)\mathbf{1}_B(x_1)\cdots\mathbf{1}_B(x_t)\varphi(x_t),$$

which follows by induction from (3.16).

Now, since M and P_B are self-adjoint and $\mathbf{1}_B = P_B\mathbf{1}$, we get

$$\varpi = \langle (P_B M)^t P_B \mathbf{1}, \mathbf{1}\rangle.$$

But since P_B is a projection, we have $P_B^2 = P_B$, and hence

$$\begin{aligned}(P_B M)^t P_B &= (P_B M)(P_B M)\cdots(P_B M)P\\ &= (P_B M P_B)(P_B M P_B)\cdots(P_B M P_B) = (P_B M P_B)^t,\end{aligned}$$

which gives the upper-bound

$$\varpi \leqslant \|P_B M P_B\|^t.$$

This means that, in order to prove the proposition, it only remains to show that the norm of $P_B M P_B$, as a linear operator from $L^2(\Gamma)$ to itself, is bounded by $\varrho_\Gamma + \mu_\Gamma(B)$.

To do this, take $\varphi \in L^2(V)$. As already done on a few occasions, we write

$$P_B\varphi = m + \psi,$$

where $m = \langle P_B\varphi, 1\rangle$ is the average of $P_B\varphi$ and $\psi \in L_0^2(\Gamma)$ (recall that Γ is not bipartite by assumption, so -1 is not an eigenvalue of M). By orthogonality, we have $\|P_B\varphi\|^2 = |m|^2 + \|\psi\|^2$. By linearity, we obtain

$$P_B M P_B \varphi = m\mathbf{1}_B + P_B M\psi,$$

but since $\|P_B\| \leqslant 1$, the definition of ϱ_Γ gives

$$\|P_B M\psi\| \leqslant \|M\psi\| \leqslant \varrho_\Gamma \|\psi\|,$$

while

$$\|m\mathbf{1}_B\| = |m|\mu_\Gamma(B) \leqslant \|P_B\varphi\|\mu_\Gamma(B) \leqslant \|\varphi\|\mu_\Gamma(B),$$

and hence we get the inequality

$$\|P_B M P_B\varphi\| \leqslant (\mu_\Gamma(B) + \varrho_\Gamma)\|\varphi\|$$

for any $\varphi \in L^2(\Gamma)$, which establishes that the norm of $P_B M P_B$ is at most $\mu_\Gamma(B) + \varrho_\Gamma$, as claimed. □

5.3. Sieve methods

Expansion properties of finite linear groups have been used in recent years in a number of remarkable arithmetic applications involving generalizations to discrete groups with exponential growth of many problems and methods of *sieve theory*, classically used to study problems such as the twin prime conjecture, or the representation of prime values by integral polynomials, or the computation of the Galois group of the splitting field of a “random” polynomial.

We will not discuss the general framework in detail, referring to the surveys [68] and [71] for a general discussion. Instead, we present one particular type of application, corresponding to the so-called “large sieve”, which leads to very nice statements that sometimes seem to be unrelated to graphs or expansion. In particular, we will discuss a result of Lubotzky and Meiri [81] that, to some extent, is a purely algebraic result concerning finitely generated linear groups, and yet depends essentially on an application of the large sieve, and on very recent results on expansion in linear groups. More applications are given in [69, Ch. 7].

The underlying sieve result is the following:

Theorem 5.3.1 (Kowalski; Lubotzky–Meiri). *Let $m \geqslant 2$ be an integer. Let $\Gamma \subset \mathrm{SL}_m(\mathbf{Z})$ be a finitely generated group, and $S = S^{-1}$ a finite symmetric generating set of Γ containing 1. Let $(X_n)_{n\geqslant 1}$ be the corresponding random walk on $\mathcal{C}(\Gamma, S)$. For p prime, let $\Gamma_p \subset \mathrm{SL}_m(\mathbf{Z}/p\mathbf{Z})$ be the image of Γ under the map π_p of reduction modulo p.*

Let $\mathcal{P}$ be an infinite set of primes such that the families of relative Cayley graphs

$$(\mathcal{C}(\Gamma_p, S))_{p\in\mathcal{P}} \tag{5.4}$$

$$(\mathcal{C}(\Gamma_{p_1} \times \Gamma_{p_2}, S))_{\substack{p_1,p_2\in\mathcal{P}\\ p_1\neq p_2}} \tag{5.5}$$

are expander families.

Then there exists $A > 1$ and $\alpha > 0$ such that for any $\delta > 0$ and any subsets $\Omega_p \subset \Gamma_p$ with

$$\frac{|\Omega_p|}{|\Gamma_p|} \geqslant \delta > 0$$

for all $p \in \mathcal{P}$, we have

$$\mathbf{P}(\pi_p(X_n) \notin \Omega_p \text{ for all } p \leqslant A^n \text{ in } \mathcal{P}) \ll H^{-1}$$

for $n \geqslant 1$, where

$$H = \sum_{\substack{p\leqslant A^n\\ p\in\mathcal{P}}} 1.$$

The constants A and α depend only on (m, δ) and on the expansion parameters for the families of Cayley graphs (5.4) and (5.5).

Remark 5.3.2. The second family of graphs is properly speaking the family of action graphs for the natural multiplication action of Γ on $\Gamma_{p_1} \times \Gamma_{p_2}$.

The intuitive meaning of this result can be appreciated in the following manner. Assume that $\mathcal{P}$ is the set of all primes. For a fixed prime p, the random variable $\pi_p(X_n)$ is the n-th step of a random walk on the finite (relative) Cayley graph $\mathcal{C}(\Gamma_p, S)$, and hence it becomes equidistributed in Γ_p when n goes to ∞. So the probability that $\pi_p(X_n)$ is in Ω_p converges to $1 - |\Omega_p|/|\Gamma_p| \leqslant 1 - \delta$. For distinct primes, we expect the reduction to behave independently; this suggests that if we reduce modulo all the primes p in a finite set T, the probability that $\pi_p(X_n)$ is not in Ω_p for $p \in T$ should be

$$\approx \prod_{p\in T}\Bigl(1 - \frac{|\Omega_p|}{|\Gamma_p|}\Bigr) \leqslant (1-\delta)^{|T|}.$$

If we can let $|T|$ grow with n, this certainly suggests that the probability in Theorem 5.3.1 should decay very fast to zero, and this is the conclusion of the theorem, since H is (under our assumption) the number of primes $\leqslant A^n$, which is $\gg A^n/n$ by the Prime Number Theorem (1.1).

The main issues in transforming this intuition into a proof are the uniformity of the convergence to equidistribution for each finite quotient Γ_p, and the approximate independence of reduction modulo different primes. The proof will show that the first problem is (unsurprisingly) handled by the expansion of the Cayley graphs of Γ_p; the second is more subtly dealt with by the expansion of the second family of Cayley graphs, which will be used to control *correlations* between distinct primes after a finite number of steps of the random walk.

We now begin the proof, but the reader might be interested in first looking at the statement of the application that we will prove later, which is due to Lubotzky and Meiri.

The proof itself is due to R. Peled (some ideas are reminiscent of the older large sieve inequalities of Rényi and Turán).

Proof of Theorem 5.3.1. For all primes p, we denote by B_p the random variable

$$B_p = \begin{cases} 1 & \text{if } p \in \mathcal{P} \text{ and } \pi_p(X_n) \in \Omega_p \\ 0 & \text{otherwise,} \end{cases}$$

and for $Q \geqslant 2$, we put

$$N = \sum_{p \leqslant Q} B_p.$$

Our goal is therefore exactly to find an upper bound for $\mathbf{P}(N = 0)$. We use Chebychev's inequality for this purpose: we have

$$\mathbf{P}(N = 0) \leqslant \frac{\mathbf{V}(N)}{\mathbf{E}(N)^2},$$

where $\mathbf{V}(N) = \mathbf{E}(N^2) - \mathbf{E}(N)^2$ is the variance of N. A lower-bound for the expectation of N is easy to obtain: we have

$$\mathbf{E}(N) = \sum_{p \leqslant Q} \mathbf{E}(B_p) = \sum_{\substack{p \leqslant Q \\ p \in \mathcal{P}}} \mathbf{P}(\pi_p(X_n) \in \Omega_p).$$

Since the family (5.4) is an expander family, equidistribution (Corollary 3.2.28) shows that there exists $\varrho < 1$ such that

$$\mathbf{P}(\pi_p(X_n) \in \Omega_p) = \frac{|\Omega_p|}{|\Gamma_p|} + O(|\Gamma_p| \varrho^n) = \frac{|\Omega_p|}{|\Gamma_p|} + O(p^{m^2} \varrho^n) \tag{5.6}$$

for any prime $p \in \mathcal{P}$ (we used a very naive bound for $|\Gamma_p|$). We pick $A > 1$ such that $A^{m^2} \varrho < 1$ and take $Q = A^n$. For n large enough, depending only on δ, ϱ and this choice of A, we then have by assumption

$$\mathbf{P}(\pi_p(X_n) \in \Omega_p) \geqslant \delta + O((A^{m^2} \varrho)^n) \geqslant \frac{\delta}{2}.$$

for $p \leqslant Q$ in $\mathcal{P}$. For any such n and this value of Q, we have

$$\mathbf{E}(N) \gg H$$

where the implied constant depends on δ.

We now estimate from above the variance $\mathbf{V}(N) = \mathbf{E}(N^2) - \mathbf{E}(N)^2$, for $Q = A^n$. Expanding the square leads to

$$\mathbf{V}(N) = \sum_{\substack{p_1, p_2 \leqslant Q \\ p_1, p_2 \in \mathcal{P}}} W(p_1, p_2)$$

where

$$W(p_1, p_2) = \Big(\mathbf{P}(\pi_{p_1}(X_n) \in \Omega_{p_1} \text{ and } \pi_{p_2}(X_n) \in \Omega_{p_2}) - \\ \mathbf{P}(\pi_{p_1}(X_n) \in \Omega_{p_1}) \mathbf{P}(\pi_{p_1}(X_n) \in \Omega_{p_1}) \Big).$$

We distinguish the cases $p_1 = p_2$ and $p_1 \neq p_2$ in this sum. In the first (diagonal) case, we just write

$$W(p_1, p_1) = \mathbf{P}(\pi_{p_1}(X_n) \in \Omega_{p_1}) - \mathbf{P}(\pi_{p_1}(X_n) \in \Omega_{p_1})^2 \leqslant 1,$$

so that the contribution of these terms to the variance is at most

$$\sum_{\substack{p\leqslant Q\\ p\in\mathcal{P}}} 1 = H.$$

For $p_1 \neq p_2$, we have

$$\mathbf{P}(\pi_{p_1}(X_n)\in\Omega_{p_1} \text{ and } \pi_{p_2}(X_n)\in\Omega_{p_2}) = \mathbf{P}((\pi_{p_1}\times\pi_{p_2})(X_n)\in\Omega_{p_1}\times\Omega_{p_2}).$$

Since the family of Cayley graphs (5.5) is an expander, equidistribution in these graphs shows that there exists $\tau<1$ such that

$$\mathbf{P}(\pi_{p_1}(X_n)\in\Omega_{p_1} \text{ and } \pi_{p_2}(X_n)\in\Omega_{p_2}) = \frac{|\Omega_{p_1}||\Omega_{p_2}|}{|\Gamma_{p_1}||\Gamma_{p_2}|} + O((p_1p_2)^{m^2}\tau^n)$$

for any primes $p_1\neq p_2$ in $\mathcal{P}$. If we combine this with (5.6), we obtain

$$W(p_1,p_2) \ll Q^{m^2}\varrho^n + Q^{2m^2}\tau^n$$

for any distinct primes p_1, $p_2 \leqslant Q$, the "main terms" having canceled. Hence

$$\mathbf{V}(N) \ll H + Q^{2+m^2}\varrho^n + Q^{2m^2+2}\tau^n.$$

If A is possibly replaced by a smaller real number (still >1) this implies $\mathbf{V}(N)\ll H$, where the implied constant depends on ϱ and τ, and the choice of A. Therefore, fixing such a value of A, we obtain

$$\mathbf{P}(N=0)\ll H^{-1},$$

as claimed. □

We will show the versatility of this result by proving a special case (for $\Gamma=\mathrm{SL}_3(\mathbf{Z})$) of the following general theorem:

Theorem 5.3.3 (Lubotzky–Meiri). *Let $n\geqslant 2$ and let $\Gamma\subset \mathrm{GL}_n(\mathbf{C})$ be a finitely generated subgroup. Then either there exists a finite index subgroup of Γ that is solvable or, for any finite symmetric generating set S of Γ with $1\in S$, there exists $\alpha>0$ such that the random walk (X_n) on $\mathcal{C}(\Gamma,S)$ satisfies*

$$\mathbf{P}(\textit{there exists } m\geqslant 2 \textit{ and } g\in\Gamma \textit{ with } X_n=g^m)\ll e^{-\alpha n},$$

for $n\geqslant 1$.

In other words, Lubotzky and Meiri prove that if a finitely generated linear group Γ is not "virtually solvable" (i.e., does not have a solvable finite index subgroup), the chance that an element $g\in\Gamma$ is a "proper power" (i.e., an element of the form h^m for some $m\geqslant 2$ and some $h\in\Gamma$) is "exponentially small" in some sense. This is a striking result, especially in view of how general it is!

We will prove Theorem 5.3.3 for $\Gamma=\mathrm{SL}_3(\mathbf{Z})$, in which case Theorem 4.3.1 (obtained through Property (T)) is sufficient to imply all the expansion properties that we need. In fact, our proof will apply with almost no changes to $\mathrm{SL}_m(\mathbf{Z})$ for $m\geqslant 3$, provided one takes as given the corresponding cases of Theorem 4.3.1.

The strategy of the proof is the following:

(1) First, we will show that the condition $X_n=g^m$ with $m\geqslant 2$ leads to a dichotomy depending on whether m is of size comparable to n (up to a multiplicative constant) or larger; this involves a nice geometric argument comparing the combinatorial distance in the Cayley graph and the usual norm of a matrix;

(2) If $m > n^2$, however large m may be, the structure arising from the first step is precise enough that the probability can be estimated by looking modulo a single suitable prime p, and this shows that

$$\mathbf{P}(\text{there exists } m > n^2 \text{ and } g \in \Gamma \text{ with } X_n = g^m) \ll e^{-\alpha n}.$$

(3) For each individual m with $2 \leqslant m \leqslant n^2$, we estimate

$$\mathbf{P}(\text{there exists } g \in \Gamma \text{ with } X_n = g^m)$$

using Theorem 5.3.1, the crucial point being that if $X_n = g^m$ is an m-th power in $\mathrm{SL}_3(\mathbf{Z})$, then it is also an m-th power modulo p for any prime p; using quite deep results on the distribution of primes in arithmetic progressions, this estimate is uniform enough to imply

$$\mathbf{P}(\text{there exists } m \text{ with } 2 \leqslant m \leqslant n^2 \text{ and } g \in \Gamma \text{ with } X_n = g^m) \ll e^{-\alpha n}.$$

(4) Combining (2) and (3), the theorem follows.

For Step 1, we will use the following terminology: if k is a field, then a matrix $g \in \mathrm{GL}_n(k)$ is *virtually unipotent* if all its eigenvalues (in an algebraic closure of k) are roots of unity. For $n = 3$ and $g \in \mathrm{SL}_3(k)$, we see that g is virtually unipotent if and only if all its eigenvalues are cube roots of unity (in an algebraic closure of k). In particular, any power of a virtually unipotent element is also virtually unipotent.

In Step 2, we will need the notion of *regular semisimple elements* in $\mathrm{GL}_n(k)$ (these will also play an importante role in Chapter 6, especially in Section 6.6): these are elements $g \in \mathrm{GL}_n(k)$ with n distinct eigenvalues (they are therefore diagonalizable over an algebraic closure of k).

We now fix a finite symmetric generating set S of $\mathrm{SL}_3(\mathbf{Z})$.

Lemma 5.3.4. *Let $g \in \mathrm{SL}_3(\mathbf{Z})$ and let $n = \ell_S(g) \geqslant 0$ be the distance to the identity in $\mathcal{C}(\mathrm{SL}_3(\mathbf{Z}), S)$. Assume that $h \in \mathrm{SL}_3(\mathbf{Z})$ and $m \geqslant 2$ are such that $g = h^m$. Then either g is virtually unipotent, or $m \ll n$, where the implied constant depends only on S.*

Proof. If h is virtually unipotent, then so is g, as we observed earlier. Assume then that h is *not* virtually unipotent.

We use the norm $\|\cdot\|$ on $\mathrm{SL}_3(\mathbf{C})$ and its subgroup $\mathrm{SL}_3(\mathbf{Z})$. Let

$$C = \max_{s \in S} \|s\|.$$

Since $\ell_S(g) = n$, the multiplicativity properties of the norm (see Lemma C.1.2 in Appendix C) shows that

$$\|g\| \leqslant C^n.$$

On the other hand, the fact that h is not virtually unipotent shows by Corollary C.3.4 that h has an eigenvalue λ such that $|\lambda| \geqslant 1 + \eta$, where $\eta > 0$ is an absolute constant. In that case, we have $\|g\| = \|h^m\| \geqslant (1 + \eta)^m$, and hence by comparison we get

$$(1 + \eta)^m \leqslant \|g\| \leqslant C^n,$$

which implies in passing that $C > 1$, and then gives $m \leqslant C_1 n$ with

$$C_1 = \frac{\log(1 + \eta)}{\log C} > 0.$$

□

Since $\ell_S(X_n) \leqslant n$ for a random walk (X_n) on $\mathcal{C}(\mathrm{SL}_3(\mathbf{Z}), S)$, it follows from this lemma that if n is large enough (in terms of S only) and if X_n is an m-th power for some $m > n^2$, then X_n is virtually unipotent. We deal with this possibility in the next lemma:

Lemma 5.3.5. *There exist* $\alpha > 0$ *and* $C \geqslant 0$*, depending only on* S*, such that*

$$\mathbf{P}(X_n \text{ is virtually unipotent}) \leqslant Ce^{-\alpha n}$$

for $n \geqslant 1$*.*

Proof. We will use reduction modulo primes, similarly to Theorem 5.3.1, but the set of virtually unipotent elements is small enough that we can work with a single well-chosen prime instead of having to combine many primes as in the large sieve.

For any prime p, we have

$$\mathbf{P}(X_n \text{ is virtually unipotent}) \leqslant \mathbf{P}(\pi_p(X_n) \text{ is virtually unipotent}).$$

We claim that

$$\frac{1}{|\operatorname{SL}_3(\mathbf{F}_p)|}|\{g \in \operatorname{SL}_3(\mathbf{F}_p) \mid g \text{ is virtually unipotent}\}| \ll \frac{1}{p} \tag{5.7}$$

for $p \geqslant 2$. This is a special case of an extremely general fact (the Lang-Weil Theorem for the number of solutions of systems of polynomial equations over finite fields), but we will suggest a direct roundabout proof in Exercise 5.3.6 below. Given this fact, equidistribution of the random walk $(\pi_p(X_n))$ on $\mathcal{C}(\operatorname{SL}_3(\mathbf{F}_p), S)$ (Corollary 3.2.28) gives

$$\mathbf{P}(X_n \text{ is virtually unipotent}) \leqslant \frac{1}{|\operatorname{SL}_3(\mathbf{F}_p)|}|\{g \in \operatorname{SL}_3(\mathbf{F}_p) \mid g \text{ is virtually unipotent}\}| + O(p^9\varrho_p^n)$$

where ϱ_p is the equidistribution radius for $\mathcal{C}(\operatorname{SL}_3(\mathbf{F}_p), S)$ and the implied constant is absolute. Since (by Theorem 4.4.4) the family of Cayley graphs is an expander, there exists $\varrho < 1$ (depending only on S) such that $\varrho_p < \varrho$ for all primes p. Using (5.7), we deduce then

$$\mathbf{P}(X_n \text{ is virtually unipotent}) \ll \frac{1}{p} + p^9\varrho^n$$

where the implied constant is absolute.

Since $\varrho < 1$, we can find a real number $A > 1$ such that $A\varrho < 1$. If n is large enough, depending on A, we can then pick a prime p such that $A^{n/9} < p \leqslant 2A^{n/9}$; then we obtain

$$\mathbf{P}(X_n \text{ is virtually unipotent}) \ll A^{-n/9} + (A\varrho)^n.$$

By our choice of A, there exists $\alpha > 0$ such that this means that

$$\mathbf{P}(X_n \text{ is virtually unipotent}) \ll e^{-\alpha n}.$$

□

Exercise 5.3.6. The basic idea behind (5.7) is that the set of virtually unipotent elements is the union of finitely many subsets of $\operatorname{SL}_3(\mathbf{F}_p)$, each of which is defined by the vanishing of some polynomial (with variables the coefficients of a matrix in SL_3) that is independent of p, and these polynomials take roughly all values equally often, so that the "probability" that the value is 0 is about $1/p$. Here we explain how to check the bound "with bare hands".

(1) Show that the bound (5.7) follows if we have

$$\frac{1}{|\operatorname{SL}_3(\mathbf{F}_p)|}|\{g \in \operatorname{SL}_3(\mathbf{F}_p) \mid \operatorname{Tr}(g) = \alpha\}| \ll \frac{1}{p},$$

for any $\alpha \in \mathbf{F}_p$, where the implied constant is independent of α. [Hint: Consider what are the possible sets of eigenvalues of a virtually unipotent element $g \in \operatorname{SL}_3(\mathbf{F}_p)$.]

(3) Show that

$$\frac{1}{|\operatorname{SL}_3(\mathbf{F}_p)|}|\{g \in \operatorname{SL}_3(\mathbf{F}_p) \mid g_{3,3} \neq 0 \text{ and } \operatorname{Tr}(g) = \alpha\}| \ll \frac{1}{p}$$

where the implied constant is independent of α. [Hint: Consider the products

$$g\begin{pmatrix} 1 & 0 & t \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

for $t \in \mathbf{F}_p$.]

(4) Show that

$$\frac{1}{|\operatorname{SL}_3(\mathbf{F}_p)|}|\{g \in \operatorname{SL}_3(\mathbf{F}_p) \mid g_{3,3} = 0 \text{ and } \operatorname{Tr}(g) = \alpha\}| \ll \frac{1}{p}$$

where the implied constant is independent of α, and conclude.

We now fix an integer $m \geqslant 2$, and we consider the probability that X_n is an m-th power, for n such that $m \leqslant n^2$. We can attempt to use Theorem 5.3.1: if $X_n = g^m$ for some $g \in \operatorname{SL}_3(\mathbf{Z})$, then $\pi_p(X_n)$ is also an m-th power in $\operatorname{SL}_3(\mathbf{F}_p)$ for all primes p, so does not belong to the set Ω_p of non-m-th powers. Our goal is to show that this is a non-trivial condition, in the sense that the set of non m-th powers is not too small.

Lemma 5.3.7. *Let $m \geqslant 2$ be an integer. Let $p \geqslant 10$ be a prime such that $p \equiv 1 \pmod m$. Then we have*

$$\frac{1}{|\operatorname{SL}_3(\mathbf{F}_p)|}|\{g \in \operatorname{SL}_3(\mathbf{F}_p) \mid g \textit{ is not an } m\textit{-th power}\}| \geqslant \frac{5}{72}.$$

Proof. This can also be interpreted as a very special case of a general statement, but we present a full proof. The idea is roughly the following: (1) "many" elements of $\operatorname{SL}_3(\mathbf{F}_p)$ are regular semisimple elements with eigenvalues in $\mathbf{F}_p$, i.e., they are diagonalizable with distinct eigenvalues lying all in $\mathbf{F}_p$; (2) among these regular semisimple elements, there is a high proportion that are *not* m-th powers. The first property is quite general (as soon as p is not too small, which here means simply $p \geqslant 10$), whereas the second ultimately boils down to the fact that, for $p \equiv 1 \pmod m$, only $(p-1)/m$ elements of $\mathbf{F}_p^\times$ are m-th powers.

Let $G = \operatorname{SL}_3(\mathbf{F}_p)$, and let $T \subset G$ be the subgroup of diagonal matrices. Let T_{reg} denote the subset of T of elements whose diagonal coefficients are distinct. We have $|T| = (p-1)^2$ (two diagonal coefficients can be chosen freely, and the last is then fixed by the requirement that the determinant is equal to 1) and

$$|T_{reg}| \geqslant (p-1)^2 - 3(p-1) \geqslant \frac{2}{3}|T| \tag{5.8}$$

for $p \geqslant 10$, since an element of T is not in T_{reg} if and only if two of its coefficients at least are equal (there are 3 possibilities for which coefficients $g_{i,i}$ are the same, and $(p-1)$ diagonal matrices with determinant 1 and $g_{i,i} = g_{j,j}$ for $i \neq j$.)

Let now $X \subset G$ be the subset of elements conjugate (in G) to some $t \in T_{reg}$. We can write

$$X = \bigcup_{r \in G} rT_{reg}r^{-1}.$$

This union is not disjoint, but if we take elements r which are distinct modulo the normalizer $N(T)$ of T in G, the union becomes disjoint. Indeed, if r_1 and r_2 are elements of G such that $r_1T_{reg}r_1^{-1} \cap r_2T_{reg}r_2^{-1} \neq \emptyset$, then we find t_1 and $t_2 \in T_{reg}$ such that

$$r_1t_1r_1^{-1} = r_2t_2r_2^{-1}.$$

In particular $(r_2^{-1}r_1)t_1(r_2^{-1}r_1)^{-1}$ is a conjugate of t_1 that belongs to T_{reg}. However, it is an elementary computation (or fact from linear algebra) that the normalizer in G of a regular semisimple element like t_1 is equal to $N(T)$. So we deduce that $r_2^{-1}r_1 \in N(T)$.

Another elementary computation is that the quotient $N(T)/T$ is isomorphic to the symmetric group $\mathfrak{S}_3$, so that $|N(T)| = 6|T|$. We deduce from this and (5.8) that

$$|X| = \frac{|G|}{|N(T)|}|T_{reg}| \geqslant \frac{|G|}{9}$$

for $p \geqslant 10$ (this is the precise form of (1) in the idea of the proof).

We now consider the set Y of matrices $g \in X$ that are m-th powers in G. We have a disjoint union

$$Y = \bigcup_{r \in G/N(T)} rY_rr^{-1},$$

where $Y_r \subset T_{reg}$ is the set of all $t \in T_{reg}$ such that rtr^{-1} is an m-th power. For such t, there exists $h \in G$ with $rtr^{-1} = h^m$, hence $t = (r^{-1}hr)^m$. The element $x = r^{-1}hr$ is then in the centralizer of t, which is equal to T because $t \in T_{reg}$. So $t = x^m$ for some $x \in T$. This shows that the size of Y_r is (at most) the number of m-th powers in T; since $t \mapsto t^m$ is a homomorphism of T with kernel equal to the subgroup of T where all coefficients are m-th roots of 1, and since $\mathbf{F}_p$ contains all the m-th roots of 1 (because $p \equiv 1 \pmod{m}$), we obtain

$$|Y| \leqslant \frac{|G|}{|N(T)|}\frac{|T|}{m^2} \leqslant \frac{|G|}{6m^2} \leqslant \frac{|G|}{24}.$$

We conclude that the set of elements of G which are not m-th powers has size

$$\geqslant |X| - |Y| = \Big(\frac{1}{9} - \frac{1}{24}\Big)|G| = \frac{5}{72}|G|.$$

□

We are now ready to finish the proof of the theorem of Lubotzky and Meiri. We can apply the large sieve bound of Theorem 5.3.1 for each fixed m with $m \geqslant 2$ using the set of primes

$$\mathcal{P}_m = \{p \geqslant 10 \mid p \equiv 1 \pmod{m}\}.$$

The necessary expansion follows immediately from Theorem 4.3.1 in the case of the family (5.4). It also does for the family (5.5), using the fact that for any primes $p_1 \neq p_2$, the simultaneous reduction homomorphism

$$\mathrm{SL}_3(\mathbf{Z}) \to \mathrm{SL}_3(\mathbf{F}_{p_1}) \times \mathrm{SL}_3(\mathbf{F}_{p_2})$$

induces an isomorphism

$$\mathrm{SL}_3(\mathbf{Z}/p_1p_2\mathbf{Z}) \simeq \mathrm{SL}_3(\mathbf{F}_{p_1}) \times \mathrm{SL}_3(\mathbf{F}_{p_2}),$$

by the Chinese Remainder Theorem. Since reduction modulo p_1p_2 is surjective from $\mathrm{SL}_3(\mathbf{Z})$ to $\mathrm{SL}_3(\mathbf{Z}/p_1p_2\mathbf{Z})$ (Proposition B.2.6), we have

$$\mathcal{C}(\Gamma_{p_1} \times \Gamma_{p_2}, S)) = \mathcal{C}(\mathrm{SL}_3(\mathbf{Z}/p_1p_2\mathbf{Z}), S),$$

so that the family (5.5) is part of the family of relative Cayley graphs of finite quotients of $\mathrm{SL}_3(\mathbf{Z})$ with respect to the generating set S, and hence is also an expander family.

The outcome of Theorem 5.3.1 for a fixed $m \geqslant 2$ is the upper bound

$$\mathbf{P}(X_n \text{ is an } m\text{-th power in } \mathrm{SL}_3(\mathbf{Z})) \ll H_m^{-1}$$

where

$$H_m = \sum_{\substack{10 \leqslant p \leqslant A^n \\ p \equiv 1 \,(\mathrm{mod}\, m)}} 1$$

for some $A > 1$. The parameter A is independent of m (in view of Lemma 5.3.7). Hence we have

$$\mathbf{P}(X_n \text{ is an } m\text{-th power in } \mathrm{SL}_3(\mathbf{Z}) \text{ for some } m \text{ with } 2 \leqslant m \leqslant n^2) \ll \sum_{m=2}^{n^2} H_m^{-1}.$$

The last crucial step is then a lower bound for H_m that is uniform for all $m \leqslant n^2$. This is provided by the famous Siegel–Walfisz Theorem:

Theorem 5.3.8 (Siegel–Walfisz Theorem). *Let $D > 0$ be a parameter. For any $X \geqslant 2$, for any integer $q \geqslant 1$ and any integer a coprime to q, we have*

$$\sum_{\substack{p \leqslant X \\ p \equiv a \,(\mathrm{mod}\, q)}} 1 = \frac{1}{\varphi(q)} \sum_{p \leqslant X} 1 + O\Big(\frac{X}{(\log X)^D}\Big)$$

where the implied constant depends only on D.

See for instance [58, Th. 5.29] for the proof. The point is that the main term is of size roughly $q^{-1}X/\log(X)$ by the Prime Number Theorem (1.1), hence the formula is a true asymptotic formula provided q is smaller than $(\log X)^{D-2}$, say.

In our situation, we consider primes $\leqslant X = A^n$ (the restriction $p \geqslant 10$ being irrelevant) modulo $m \leqslant n^2 \ll (\log X)^2$; so if we take $D = 4$ in the Siegel–Walfisz Theorem, we obtain the lower bound

$$H_m \gg \frac{1}{n^2} \frac{A^n}{(\log A^n)}$$

for $m \leqslant n^2$, where the implied constant is absolute. This translates to

$$\mathbf{P}(X_n \text{ is an } m\text{-th power in } \mathrm{SL}_3(\mathbf{Z}) \text{ for some } m \text{ with } 2 \leqslant m \leqslant n^2) \ll n^3 A^{-n} \ll e^{-\alpha n}$$

for some $\alpha > 0$. On the other hand, Lemmas 5.3.4 and 5.3.5 together imply

$$\mathbf{P}(X_n \text{ is an } m\text{-th power in } \mathrm{SL}_3(\mathbf{Z}) \text{ for some } m \text{ with } m > n^2) \ll e^{-\alpha_1 n}$$

for some $\alpha_1 > 0$. Taken the smallest of the two numbers α and α_1, say α, leads finally to the desired bound

$$\mathbf{P}(X_n \text{ is a proper power}) \ll e^{-\alpha n}$$

for $n \geqslant 1$. This concludes the proof of the special case $\Gamma = \mathrm{SL}_3(\mathbf{Z})$ of Theorem 5.3.3.

5.4. Geometric applications

Historically, the first applications of expander graphs outside of "discrete" mathematics were related to properties of towers of coverings of manifolds in the setting of Riemannian geometry, and especially of the spectral geometry of the (Riemannian) Laplace operator. Since then, many more applications of this type have been obtained, and we will present some of them in this section, without complete proofs but with precise references. It is remarkable to see how each statement may look strikingly different from the others – and how each of them may seem absolutely unexpected.

And yet, all these applications have a common feature in common, as they illustrate in a variety of situations what one may term the “expander philosophy”:

> *If the notion of Galois covering makes sense for certain mathematical objects then assuming that we have a sequence of Galois coverings $\mathcal{X}_n \rightarrow \mathcal{X}$ of such objects, where the Galois groups are finite groups whose Cayley graphs (with respect to a suitable generating set) form an expander family, then the objects $\mathcal{X}_n$ will be “very complicated” when n gets large – very often, they will look very much like generic (or random) objects of their type.*

Before discussing instances of this principle in geometry, the reader should observe that, in some sense, the equivalence between the various definitions of expansion are already examples of the expander philosophy, applied to graphs themselves.

By now, one can state precise incarnations of the principle in Riemannian geometry (comparison of Riemannian and discrete invariants, due to Brooks and Burger), algebraic geometry (lower-bounds for the gonality of algebraic curves, due to Ellenberg, Hall and Kowalski [34]), arithmetic geometry (finiteness of rational points of bounded degree on certain algebraic curves, obtained by combining the gonality bounds with results of Faltings and Frey), and differential geometry (lower-bounds for the distorsion of certain knots, due to Gromov and Guth [48], lower-bounds for the Heegaard genus of a compact 3-manifold, due to Lackenby [73]).

We will survey these results, concentrating on the statements of the gonality lower bounds, and of the distorsion lower bound of Gromov and Guth. The proof of both involves Laplace eigenvalue comparison, and we will explain the proof of the basic result of Brooks and Burger in that direction.

Example 5.4.1 (Gonality lower bounds). Our first example concerns a geometric invariant of compact Riemann surfaces of genus $g \geqslant 2$. We first recall a characterization of these surfaces which does not require much background. Details and references are supplied in many places, for instance the textbooks [87] of Miranda and [39] of Farkas–Kra, or the book [22] of Buser.

Consider the Poincaré upper half-plane

$$\mathbf{H} = \{z \in \mathbf{C} \mid \mathrm{Im}(z) > 0\} \subset \mathbf{C}.$$

This is an open subset of $\mathbf{C}$, and therefore it makes sense to speak of holomorphic (or meromorphic) functions on $\mathbf{H}$. The group $\mathrm{SL}_2(\mathbf{R})$ acts on $\mathbf{H}$ by the formula

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z = \frac{az+b}{cz+d}.$$

Indeed, a simple computation (using the fact that $ad - bc = 1$) shows that

$$\mathrm{Im}\Bigl(\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z\Bigr) = \mathrm{Im}\Bigl(\frac{az+b}{cz+d}\Bigr) = \frac{\mathrm{Im}(z)}{|cz+d|^2},$$

which shows that $\mathbf{H}$ is preserved by this action. The action is transitive: for instance, for any $x + iy \in \mathbf{H}$, we have

$$\begin{pmatrix} y^{1/2} & y^{-1/2}x \\ 0 & y^{-1/2} \end{pmatrix} \cdot i = x + iy,$$

but it is not faithful, in the sense that some matrices $g \neq 1$ act by the identity. In fact, it is a simple computation that only $g = -\mathrm{Id}$ has this property.

We then consider discrete subgroups $\Gamma \subset \mathrm{SL}_2(\mathbf{R})$ that act *without fixed points* and such that the space $X_\Gamma = \Gamma \backslash \mathbf{H}$ of orbits is a *compact* topological space with the quotient topology. (The fixed-point condition means that if $g \in \Gamma$ and $z \in \mathbf{H}$ satisfy $g \cdot z = z$, then $g = \pm \mathrm{Id}$). The compactness is a non-trivial condition, and completely explicit examples are not so easy to describe. However, it is known that such groups exist in great abundance. The corresponding quotients X_Γ are exactly the compact (connected) Riemann surfaces of genus $g \geqslant 2$ (see, e.g., [39, Ch. IV, IV.6.5]). The group Γ is (isomorphic to) the *fundamental group* $\pi_1(X_\Gamma, x_0)$ of the topological space X_Γ, where $x_0 \in X_\Gamma$ is an arbitrary basepoint.

The (topological) invariant g can be recovered from X_Γ, or from Γ, as follows: abstractly, the group Γ can be shown to be isomorphic to the finitely generated group generated by $2g$ elements $(a_1, \ldots, a_g, b_1, \ldots, b_g)$, subject to the unique relation

$$[a_1, b_1] \cdots [a_g, b_g] = 1$$

where $[x, y] = xyx^{-1}y^{-1}$ is the commutator of two elements of a group. In particular, the abelianization $\Gamma^{ab} = \Gamma/[\Gamma, \Gamma]$ is generated by $2g$ elements with no relation (since the image of each commutator is trivial in Γ^{ab}), i.e., it is isomorphic to $\mathbf{Z}^{2g}$.

Since X_Γ arises from $\mathbf{H}$, one can speak of holomorphic maps from X_Γ to $\mathbf{C}$, or of meromorphic functions: by definition $f \colon X_\Gamma \to \mathbf{C}$ has such a property if and only if the composition

$$\mathbf{H} \to X_\Gamma \to \mathbf{C}$$

is holomorphic (resp. meromorphic). (Similarly, it makes sense to speak of differentiable, or C^2, or smooth functions on X).

From this point of view, it is easy to see that if $\Gamma_1 \subset \Gamma$ is a finite index subgroup, the subgroup Γ_1 is also discrete and the quotient $\Gamma_1 \backslash \mathbf{H}$ is also compact. Hence X_{Γ_1} is defined, and the identity map of $\mathbf{H}$ induces a holomorphic map

$$f \colon X_{\Gamma_1} \to X_\Gamma.$$

This is a topological covering map; for any $z \in X_\Gamma$, the pre-image $f^{-1}(z)$ is a set of $[\Gamma : \Gamma_1]$ distinct points on X_{Γ_1}, so the map f has degree $[\Gamma : \Gamma_1]$. The Riemann surface X_{Γ_1} has genus $g(X_{\Gamma_1}) \geqslant 2$, related to the genus g of X_Γ by the Hurwitz formula

$$(2 - 2g(X_{\Gamma_1})) = [\Gamma : \Gamma_1](2 - 2g) \tag{5.9}$$

(see, e.g., [87, Th. 4.16]). In particular, if Γ_1 varies in Γ, then the genus of X_{Γ_1} grows linearly with the index of Γ_1 in Γ.

For a given genus $g \geqslant 2$, although all spaces X_Γ are the same from the topological point of view, they may be geometrically very different (in the sense of holomorphic equivalence, or of Riemannian geometry). In fact, Riemann already understood that the space of all possible surfaces X_Γ of genus $g \geqslant 2$, up to holomorphic isomorphism, depends on $3g - 3 \geqslant 3$ complex parameters, which are called “moduli” (there is an intuitive explanation of this number in [87, §VII.2], and a precise description of this space in [22, Ch. 6]).

Definition 5.4.2 (Gonality). Let X be a compact connected Riemann surface (not necessarily of genus $\geqslant 2$). The *gonality* $\gamma(X)$ is the smallest integer $\gamma \geqslant 1$ such that there exists a meromorphic map $f \colon X \to \mathbf{C}$ of degree γ, in the sense that for all but finitely many $z \in \mathbf{C}$, the fiber $f^{-1}(z)$ has γ elements.

The gonality is a geometric invariant that is finer (and much more complicated) than the genus. One can show the following facts:

- The gonality is finite (this is by no means obvious with our presentation, as it amounts to the existence of at least one non-constant meromorphic function on X, which is a priori an analytic problem; see [39, Cor. to Th. II.5.3]). In fact, one can prove that

$$\gamma(X) \leqslant \left\lfloor \frac{g+3}{2} \right\rfloor. \tag{5.10}$$

- The gonality is 1 if and only if X is holomorphically isomorphic to the Riemann sphere (i.e., to the projective line over $\mathbf{C}$; this means that the genus is 0).
- If $g \geqslant 2$, then all possible integers

$$2 \leqslant \gamma \leqslant \left\lfloor \frac{g+3}{2} \right\rfloor$$

are achieved as the gonality of some compact Riemann surface of genus $g \geqslant 2$; "generically" (in a precise sense involving the parameters describing compact Riemann surfaces of genus g), the gonality takes the maximal possible value; see the references in [38].
- Riemann surfaces with gonality 2 are called *hyperelliptic*; for $g = 1$ or $g = 2$, all Riemann surfaces of genus g have gonality 2, but this is false if $g \geqslant 3$, although it is true that there always exist hyperelliptic Riemann surfaces of genus g (see, e.g., [39, III.7] or [87, III.1]).

The first two facts are (fairly elementary) consequences of the Riemann-Roch Theorem, which is probably the most important analytic fact about compact Riemann surfaces (see, e.g., [87, VI.3 and VII] or [39, III.4]).

Because it is a geometric invariant, the gonality is typically rather difficult to compute. Proving lower bounds is far from obvious since it requires the proof that certain objects (holomorphic maps with small degree) do not exist. However, the following theorem from [34] shows how to obtain many families where the gonality grows:

THEOREM 5.4.3 (Ellenberg–Hall–Kowalski). *Let X_Γ be a compact connected Riemann surface of genus $g \geqslant 2$ and S a finite symmetric generating set of Γ. For $n \geqslant 1$, let $\Gamma_n \lhd \Gamma$ be a normal subgroup of finite index with $[\Gamma : \Gamma_n] \to +\infty$ as $n \to +\infty$. If the family of relative Cayley graphs*

$$(\mathcal{C}(\Gamma/\Gamma_n, S))_{n\geqslant 1}$$

is an expander family, then $\gamma(X_{\Gamma_n})$ tends to infinity as $n \to +\infty$. In fact, we then have

$$\gamma(X_{\Gamma_n}) \gg [\Gamma : \Gamma_n],$$

where the implied constant depends on S and the expansion parameters of the family.

Note that from (5.9) and (5.10), it follows easily that (for any family as in the theorem) we have

$$\gamma(X_{\Gamma_n}) \leqslant (g+2)[\Gamma : \Gamma_n],$$

and hence the growth of gonality given by Theorem 5.4.3 is as fast as possible (in terms of the order of magnitude). Moreover, the result does not hold for base curves of genus $g = 0$ or $g = 1$, since in that case the genus of X_{Γ_n} is also equal to g for all n by the Hurwitz formula, and we mentioned that the gonality of these Riemann surfaces is always equal to 1 or 2, respectively.

The strategy of the proof of Theorem 5.4.3 has two steps. First, a beautiful inequality of Li and Yau [77] connects the gonality of a compact Riemann surface $X = X_\Gamma$ of genus

$\geqslant 2$ with the spectral gap for the *Riemannian* Laplace operator on X. The latter is defined as follows for $X = X_\Gamma$. The differential operator

$$\Delta_{\mathbf{H}} = -y^2\Big(\frac{\partial^2}{\partial^2 x} + \frac{\partial^2}{\partial^2 y}\Big)$$

acting on smooth (or simply C^2) functions on $\mathbf{H}$ (not holomorphic functions!) has the property that

$$\Delta_{\mathbf{H}}(f(g\cdot z)) = g\cdot(\Delta_{\mathbf{H}}f)(z),$$

for all $g \in \mathrm{SL}_2(\mathbf{R})$, which means that it defines a differential operator Δ_X acting on smooth (or C^2) functions on X. Similarly, the measure

$$\mu_{\mathbf{H}} = \frac{dxdy}{y^2} \tag{5.11}$$

on $\mathbf{H}$ satisfies $g_*\mu_{\mathbf{H}} = \mu_{\mathbf{H}}$ for any $g \in \mathrm{SL}_2(\mathbf{R})$, from which it follows (somewhat less formally) that it induces a measure μ_X on $X = \Gamma\backslash\mathbf{H}$. One can therefore also define a space $L^2(X, \mu_X)$ of L^2-functions on X.

The measure μ_X is finite (because the measure $\mu_{\mathbf{H}}$ is finite on compact sets and there is a compact set in $\mathbf{H}$ that surjects to X), so that in particular the function 1 belongs to $L^2(X, \mu_X)$. This function also satisfies $\Delta_X(1) = 0$. The first non-zero eigenvalue, or *spectral gap*, of the Laplace operator of X, is then given by

$$\lambda_1(X) = \inf\Big\{\frac{\langle\Delta\varphi,\varphi\rangle}{\|\varphi\|^2} \mid \varphi \in L^2(X,\mu_X) \text{ smooth and } \int_N \varphi(x)d\mu(x) = 0\Big\}$$

$$= \min\Big\{\frac{\displaystyle\int_X \|\nabla\varphi\|^2 d\mu}{\|\varphi\|^2} \mid \varphi \in L^2(X,\mu_X) \text{ smooth and } \int_X \varphi(x)d\mu(x) = 0\Big\}, \tag{5.12}$$

where $\nabla\varphi$ refers to the gradient of φ with respect to the hyperbolic metric, namely $\nabla\varphi = (y\partial_x\varphi, y\partial_y\varphi)$. The fact that the minimum is achieved is non-trivial, but it is then relatively elementary that a function achieving this minimum is an eigenfunction of Δ with eigenvalue λ_1 (basic facts of spectral geometry can be found, e.g., in the book of Chavel [27, Ch. 1]).

Note that this formula is clearly analogue to the formula of Proposition 3.4.3 (2) for the first non-zero eigenvalue of the discrete Laplace operator on a finite graph.

We now have:

Theorem 5.4.4 (Li–Yau). *Let X be a compact connected Riemann surface of genus $\geqslant 2$. We have*

$$\gamma(X) \geqslant \frac{1}{4\pi}\mu_X(X)\lambda_1(X).$$

This follows from [77, Th. 1], which is a very nice argument involving the Brouwer fixed-point theorem, and [77, Fact 1].

To go from Theorem 5.4.4 to Theorem 5.4.3, we need a result that bridges the gap between the "continuous" world of Riemannian geometry and the discrete world of graphs. Different forms of this remarkable result were proved independently by Brooks [18] and Burger [19]. We will use Burger's version (which we state in greater generality for Riemannian manifolds of any dimension d; in the case of Riemann surfaces, we have $d = 2$).

Theorem 5.4.5 (Burger). *Let M be a compact connected oriented Riemannian manifold of dimension d. Let S be a fixed finite symmetric generating set of the fundamental* *group of M. There exists a constant $c_{S,M} > 0$ such that for any finite connected Galois covering $N \to M$ with Galois group G, we have*

$$\lambda_1(N) \geqslant c_{S,M}\, \underline{\lambda}_1(\mathcal{C}(G, S)).$$

In particular, if $X = X_\Gamma$ is a compact Riemann surface and Γ' is a normal subgroup with finite index in Γ, then

$$\lambda_1(X_{\Gamma'}) \geqslant c_{S,\Gamma}\, \underline{\lambda}_1(\mathcal{C}(\Gamma/\Gamma', S)).$$

Combining this result with the Li-Yau inequality, Theorem 5.4.3 follows: since X_{Γ_n} is a Galois covering of X with Galois group Γ/Γ_n, we get

$$\begin{aligned}\gamma(X_{\Gamma_n}) &\geqslant \frac{1}{4\pi}\mu_{X_{\Gamma_n}}(X_{\Gamma_n})\lambda_1(X_{\Gamma_n})\\ &\gg \mu_{X_{\Gamma_n}}(X_{\Gamma_n})\lambda_1(\mathcal{C}(\Gamma/\Gamma_n, S))\\ &\gg [\Gamma : \Gamma_n]\end{aligned}$$

since $\mu_{X_{\Gamma_n}}(X_{\Gamma_n}) = [\Gamma : \Gamma_n]\mu_X(X)$, and since the spectral gap of the Cayley graphs $\mathcal{C}(\Gamma/\Gamma_n, S)$ is bounded away from 0 by the expansion assumption.

Remark 5.4.6. Interestingly, since Theorem 5.4.5 is posterior (by a few years) to the Li-Yau inequality, the gonality growth result of Theorem 5.4.3 could have been proved at the same time!

It should be mentioned however that the main results of Ellenberg, Hall and Kowalski are the *arithmetic applications* of the gonality lower bounds, which arise from deep results of Faltings and Frey, and from the developments in the theory of expander graphs discussed in Section 4.3, especially those related to Theorem 4.3.2. These applications require the extensions of the Li-Yau inequality and of Theorem 5.4.5 to certain possibly non-compact Riemann surfaces, and to possibly non-Galois coverings; see [34] for the details.

We will give, after the discussion of the next example, a sketch of the proof of this comparison theorem, in order to illustrate the key geometric idea that explains how the continuous and discrete Laplace operators can be compared.

Example 5.4.7 (Distorsion lower bounds). The second example is maybe even more remarkable than the previous one. In fact, the main statement, due to Gromov and Guth, may seem to lie as far away as possible from the world of graphs, and it does not even directly mention Riemannian geometry to suggest the use of results like Theorem 5.4.5. This is the distorsion problem for knots that we mentioned already briefly in Chapter 1. We will only be able to sketch the results here, and we refer to the original paper [48], as well as to the Bourbaki seminar of N. Bergeron [8] for more details.

The underlying question is due to Gromov [47, p. 114], who defined a certain geometric invariant of knots, and asked whether it is unbounded.

Definition 5.4.8 (Knots). (1) A *physical knot* in $\mathbf{S}^3$ is a smooth *injective* map $k\colon \mathbf{S}^1 \to \mathbf{S}^3$.

(2) A *knot* in $\mathbf{S}^3$ is an equivalence class of physical knots for the equivalence relation where $k_1 \sim k_2$ if and only if there exists a smooth map $K\colon [0,1] \times \mathbf{S}^3 \to \mathbf{S}^3$ such that for each t, the map $x \mapsto K(t,x)$ is a diffeomorphism of $\mathbf{S}^3$, and moreover $K(0,x) = x$ and $K(1, k_1(x)) = k_2(x)$ for all $x \in \mathbf{S}^1$.

The map K is called an "ambient isotopy" of k_1 and k_2. From the definition, we see that if $k_1 \sim k_2$, then there exists a diffeomorphism $\varphi\colon \mathbf{S}^3 \to \mathbf{S}^3$ such that $k_2 = \varphi \circ k_1$, namely we can put $\varphi(x) = K(1,x)$.

Now we define Gromov's invariant, called the *distorsion*. For this we first recall that one can define the length of a smooth curve $\gamma\colon [0,1] \to \mathbf{R}^3$ by

$$\ell(\gamma) = \int_0^1 \|\ell'(t)\| dt,$$

and this is invariant under reparameterization of the curve. One can then define an intrinsic distance on a physical knot k by

$$d_k(x,y) = \inf_{\substack{\gamma(0)=x \\ \gamma(1)=y}} \ell(\gamma),$$

for x and y on k, where γ runs over smooth curves with image in k.

Definition 5.4.9 (Distorsion of a knot). (1) Let k be a physical knot in $\mathbf{R}^3$. The *distorsion* of k is

$$\mathrm{dist}(k) = \sup_{\substack{x,y\in k \\ x\neq y}} \frac{d_k(x,y)}{\|x-y\|}.$$

(2) Let $[k]$ be a knot in $\mathbf{R}^3$. The *intrinsic distorsion* of $[k]$ is

$$\mathrm{idist}([k]) = \inf_{k_1 \sim k} \mathrm{dist}(k_1).$$

Question (Gromov). Do there exist knots $[k]$ with $\mathrm{idist}([k])$ arbitrarily large?

Intuitively, the question of Gromov asks whether there exist knots in the space $\mathbf{R}^3$ that are geometrically arbitrarily complicated, in the sense that one can always find points on the knot that are "close" in $\mathbf{R}^3$ and yet far away on the knot itself. As Gromov explains, the difficulty in answering this question is that the known knot invariants usually have the property that they are *unbounded* when restricted to knots with (say) $\mathrm{idist}(k) \leqslant 100$. So it is not possible to get knots with large distorsion by looking for knots where these other invariants are themselves large.

This question was first answered by Pardon [93], who showed that certain "torus knots" have large distorsion[3]. But these knots are very special and one can ask how general the phenomenon is. Gromov and Guth use expanders to construct many more examples.

Their starting point is a "construction" of knots from classical differential topology (see [53] and [88] for the original papers).

Proposition 5.4.10 (Hilden; Montesinos). *Let M be a compact connected oriented 3-manifold. There exists a smooth map $f\colon M \to \mathbf{S}^3$ and a physical knot k in $\mathbf{S}^3$ such that f induces a covering of degree 3*

$$f^{-1}(\mathbf{S}^3 - k) \to \mathbf{S}^3 - k$$

of the complement of k.

For convenience, we will say that such a knot k is an HM-knot of M. There is no reason for it to be unique.

We can now state the remarkable theorem proved by Gromov and Guth:

[3] For readers who understand French, a very nice online presentation of the problem and of Pardon's work is available at images.math.cnrs.fr/Des-Noeuds-Indetordables.html

Theorem 5.4.11 (Gromov–Guth). *Let M be a compact hyperbolic 3-manifold, and let S be a fixed finite symmetric generating set of its fundamental group. Let $(M_n)_{n\geqslant 1}$ be a sequence of finite Galois coverings $M_n \to M$ with Galois groups Γ_n such that $(\mathcal{C}(\Gamma_n, S))_{n\geqslant 1}$ is an expander family. Let $(k_n)_{n\geqslant 1}$ be a sequence of* HM*-knots of M_n. Then* $\mathrm{idist}([k_n])$ *tends to infinity. In fact, we have*

$$\mathrm{idist}([k_n]) \gg |\Gamma_n|,$$

where the implied constant depends only on M.

In order to answer the question of Gromov, it is therefore required to know whether sequences of coverings (M_n) exist as requested in the statement. This is by no means obvious! Gromov and Guth used the special case of so-called arithmetic hyperbolic 3-manifolds, for which a generalization of the theorem of Selberg mentioned in Remark 4.3.3 (2), combined with Theorem 5.4.5, shows that congruence coverings have the desired property. However, the argument applies to any compact hyperbolic 3-manifold, provided one uses the Bourgain–Gamburd Theorem 6.1.1, together with the following fact concerning hyperbolic 3-manifolds:

Proposition 5.4.12. *Let M be a compact hyperbolic 3-manifold. There exists a number field $E/\mathbf{Q}$ and a subring $\mathcal{O}$ of E, obtained from the ring of integers by inverting finitely many primes, such that the fundamental group of M is isomorphic to a Zariski-dense subgroup of* $\mathrm{SL}_2(\mathcal{O})$. *In particular, there exists an infinite sequence of primes p such that the fundamental group of M admits a quotient isomorphic to* $\mathrm{SL}_2(\mathbf{F}_p)$.

We refer to the book [4, C.7, C.29] of Aschenbrenner, Friedl and Wilton for references about this fact – and for a fascinating description of many other remarkable properties of fundamental groups of 3-manifolds.

We briefly sketch the strategy of Gromov and Guth. The key geometric step is the following proposition:

Proposition 5.4.13. *There exists a constant $c_2 > 0$ such that for any compact hyperbolic 3-manifold M and any* HM*-knot k of M, we have*

$$\mathrm{dist}(k) \geqslant c_2 \,\mathrm{Vol}(M) h(M)$$

where $h(M)$ is the Riemannian Cheeger constant of M, defined by

$$h(M) = \inf_S \frac{\mathrm{Area}(S)}{\min(\mathrm{Vol}(A), \mathrm{Vol}(B))}$$

where S runs over all smooth surfaces $S \subset M$ such that $M - S = A \cup B$ with A and B open.

This result is quite strong! Indeed, in the setting of this result, since any knot $k_1 \sim k$ is also an HM-knot of M, and the right-hand side is independent of k, we get the stronger conclusion

$$\mathrm{idist}([k]) \geqslant c_2 \,\mathrm{Vol}(M) h(M)$$

for any HM-knot k of M. In particular, to prove Theorem 5.4.11 using this proposition, it suffices to show that for a sequence of Galois coverings $(M_n \to M)$ whose Cayley graphs are expanders, the Cheeger constant $h(M_n)$ does not tend to 0. This follows from Theorem 5.4.5 and the *Cheeger inequality*

$$\lambda_1(M) \leqslant \frac{1}{4} h(M)^2$$

for any compact connected oriented Riemannian manifold M (see [28]). This inequality is of course the geometric analogue of the bound (3.32) of Proposition 3.3.6.

We conclude this section with the promised sketch of the proof of Theorem 5.4.5, following [19, Ch. 6] and Appendix A in [34].

Sketch of the proof of Theorem 5.4.5. The comparison of eigenvalues relies on their variational characterizations, given by Proposition 3.4.3 (2) in the combinatorial case, and by the analogue

$$\lambda_1(N) = \min\Big\{ \frac{\displaystyle\int_N \|\nabla\varphi\|^2 d\mu}{\|\varphi\|^2} \mid \varphi \in L^2(N,\mu) \text{ smooth and } \int_N \varphi(x)d\mu(x) = 0 \Big\},$$

of (5.12) for the Riemannian Laplace operator Δ on N (where μ is the Riemannian measure on N). The reader may assume that we are in the setting of compact Riemann surfaces for simplicity, in which case Δ and μ have the meaning described above.

We denote $\lambda_1 = \lambda_1(N)$, which is > 0 by spectral geometry of compact Riemannian manifolds, and fix a function ψ on N of norm 1 such that $\Delta\psi = \lambda_1\psi$.

The starting point of the argument is that one can "embed" a Cayley graph of the Galois group G of the covering $M \to N$ in the surface N, although possibly with respect to a different generating set of the fundamental group of M. First we fix $x_0 \in M$. Let P be the set of points in N above x_0 and $\widetilde{P}$ the set of those in the universal cover $\widetilde{M}$ of M (which is $\mathbf{H}$ if M is a compact Riemann surface of genus $\geqslant 2$). The Galois group G of the covering acts simply transitively on P (because it is a Galois covering), and in particular we have $|P| = |G|$. For $x \in P$, $\widetilde{x} \in \widetilde{P}$, let

$$\mathcal{F}(x) = \{u \in N \mid d(u,x) < d(u,x') \text{ for all } x' \in P,\ x' \neq x\},$$
$$\widetilde{\mathcal{F}}(\widetilde{x}) = \{u \in \widetilde{M} \mid d(u,\widetilde{x}) < d(u,x') \text{ for all } x' \in \widetilde{P},\ x' \neq \widetilde{x}\}.$$

It is known that each $\widetilde{\mathcal{F}}(\widetilde{x}) \subset \widetilde{M}$ is a fundamental domain for the action of $\pi_1(M,x_0)$ on $\widetilde{M}$ (in other words every point of $\widetilde{M}$ lies in an orbit of a point in the closure of $\widetilde{\mathcal{F}}(\widetilde{x})$, and the orbits of elements of $\widetilde{\mathcal{F}}(\widetilde{x})$ are disjoint). Similarly, $\mathcal{F}(x) \subset N$ is a fundamental domain for the covering $N \to M$.

When x (resp. $\widetilde{x}$) varies, the sets $\mathcal{F}(x)$ (resp. $\mathcal{F}(\widetilde{x})$) are disjoint. The sets $\overline{\mathcal{F}(x)}$ for $x \in P$ cover N, with the complement having measure 0, and in particular we have $\mu(\mathcal{F}(x)) = \mu(M)$ for all $x \in P$.

Let

$$T = \{g \in \pi_1(M,x_0) \mid g\overline{\widetilde{\mathcal{F}}(\widetilde{x})} \cap \overline{\widetilde{\mathcal{F}}(\widetilde{x})} \neq \emptyset\}.$$

This is a finite symmetric subset of $\pi_1(M,x_0)$, containing 1, and one can check that it is a generating set of $\pi_1(M,x_0)$. By Proposition 3.5.1, we need only prove Theorem 5.4.5 when the generating set S is replaced by T. We denote $d = |T|$, and remark that T (and d) only depend on M, not on N.

We consider the simple graph Γ with vertex set P and edges joining x and x' in P if and only if

$$\overline{\mathcal{F}(x)} \cap \overline{\mathcal{F}(x')} \neq \emptyset$$

(in particular, with a loop at each vertex $x \in P$). Because the Galois group G of the covering permutes simply transitively the sets $\mathcal{F}(x)$ (which comes from the fact that the covering $N \to M$ is a Galois covering), we see that Γ is isomorphic to the Cayley graph

$\mathcal{C}(G,T)$, and in particular is d-regular. Our goal becomes to compare λ_1 and $\lambda_1(\Gamma)$. (See also [19, Ch.3, §4] for details about this construction.)

Now we must relate (smooth) functions on N to discrete functions on the graph Γ. The idea is quite simple: the space $L^2(\Gamma)$ (as a vector space) can be identified with the space of function on N which are constant on each piece $\mathcal{F}(x)$ for $x \in P$ (this is because the complement of the union of these pieces has measure 0). We consider then the orthogonal projection Φ from $L^2(N,\mu)$ to $L^2(\Gamma)$. Concretely, to a function $f \in L^2(N,\mu)$, we associate the function

$$x \mapsto \frac{1}{\mu(\mathcal{F}(x))}\int_{\mathcal{F}(x)} f d\mu = \frac{1}{\mu(M)}\int_{\mathcal{F}(x)} f d\mu$$

in $L^2(\Gamma)$. The linear map Φ is continuous since

$$\|\Phi(f)\|^2 = \sum_{x\in P}\frac{1}{\mu(M)^2}\Bigl|\int_{\mathcal{F}(x)} f d\mu\Bigr|^2 \leqslant \mu(M)^{-1}\|f\|^2 \tag{5.13}$$

by the Cauchy-Schwarz inequality and the fact that the sets $\mathcal{F}(x)$ are disjoint of measure $\mu(M)$.

For $x \in P$, we define

$$\mathcal{G}(x) = \bigcup_{y\sim x}\overline{\mathcal{F}(y)} \subset N,$$

where y runs over elements of P adjacent to x in Γ. We have $\mu(\mathcal{G}(x)) = d\mu(M)$.

We will use the following fact from spectral geometry, and will explain it below:

Fact 1. There exists a constant $\eta > 0$, depending only on M, such that, for all $x \in P$ and for $\mathcal{D} = \overline{\mathcal{F}(x)}$ and $\mathcal{D} = \overline{\mathcal{G}(x)}$, we have

$$\inf\Bigl\{\frac{\displaystyle\int_{\mathcal{D}}\|\nabla\varphi\|^2 d\mu}{\int_{\mathcal{D}}|\varphi|^2 d\mu} \mid 0 \neq \varphi \text{ smooth and } \int_{\mathcal{D}}\varphi(x)d\mu(x) = 0\Bigr\} \geqslant \eta. \tag{5.14}$$

Assuming this, consider a non-zero function of the type

$$f = \alpha + \beta\psi$$

on N with α, $\beta \in \mathbf{R}$, and ψ the λ_1-eigenfunction of norm 1 that we fixed at the beginning. We have an obvious inequality

$$\int_N \|\nabla f\|^2 d\mu = \lambda_1\beta^2 \leqslant \lambda_1\|f\|^2. \tag{5.15}$$

On the other hand, we can estimate the left-hand side from below using the pieces $\mathcal{G}(x)$. For any $x \in P$, the function

$$\varphi = f - \frac{1}{\mu(\mathcal{G}(x))}\int_{\mathcal{G}(x)} f d\mu$$

satisfies $\nabla\varphi = \nabla f$, and can be used to "test" (5.14). It follows that

$$\begin{aligned}\int_{\mathcal{G}(x)}\|\nabla f\|^2 &\geqslant \eta\int_{\mathcal{G}(x)}\Bigl|f - \frac{1}{\mu(\mathcal{G}(x))}\int_{\mathcal{G}(x)} f d\mu\Bigr|^2 d\mu \\ &= \eta\Bigl\{\int_{\mathcal{G}(x)}|f|^2 d\mu - \frac{1}{\mu(\mathcal{G}(x))}\Bigl|\int_{\mathcal{G}(x)} f d\mu\Bigr|^2\Bigr\}.\end{aligned}$$

The set $\mathcal{G}(x)$ is the union of the d subsets $\mathcal{F}(y)$ for $y \sim x$, and has measure $d\mu(M)$. Summing the above inequality over $x \in P$ and dividing by d we obtain

$$\|\nabla f\|^2 \geqslant \eta\Big\{\|f\|^2 - \frac{\mu(M)}{d^2}\sum_{x\in P}\Big|\sum_{y\sim x}\Phi(f)(y)\Big|^2\Big\}.$$

Using (5.15), this becomes

$$\lambda_1 \geqslant \eta\Big\{1 - \frac{\mu(M)}{d^2\|f\|^2}\sum_{x\in P}\Big|\sum_{y\sim x}\Phi(f)(y)\Big|^2\Big\}$$

which by (5.13) implies

$$\lambda_1 \geqslant \eta\Big\{1 - \frac{1}{d^2}\frac{\langle A_\Gamma\Phi(f), A_\Gamma\Phi(f)\rangle}{\|\Phi(f)\|^2}\Big\}.$$

Since the adjacency operator A_Γ is self-adjoint, this means that

$$\lambda_1 \geqslant \eta\frac{\langle B_\Gamma\Phi(f), \Phi(f)\rangle}{\|\Phi(f)\|^2},$$

where

$$B_\Gamma = 1 - \frac{1}{d^2}A_\Gamma^2 = \frac{1}{d^2}(d^2 - A_\Gamma^2) = \frac{1}{d^2}\Delta_\Gamma(2d - \Delta_\Gamma).$$

The operator B_Γ is positive and admits the eigenvalue 0 with multiplicity 1, with constant eigenfunction. Let $\lambda_1'(\Gamma) > 0$ denote the smallest positive eigenvalue of B_Γ. We claim:

Fact 2. There exists $c > 0$, depending only on M, such that either $\lambda_1 \geqslant c$ or else there exists α, $\beta \in \mathbf{R}$ not both zero for which $\Phi(\alpha + \beta\psi)$ is non-zero and has mean zero on P.

If this is the case, then either we have $\lambda_1 \geqslant c$, and we are done, or else we can construct the test function $f = \alpha + \beta\psi$; since $\langle\Phi(f), 1\rangle = 0$, the variational inequality for the spectrum of B_Γ leads to

$$\frac{\langle B_\Gamma\Phi(f), \Phi(f)\rangle}{\|\Phi(f)\|^2} \geqslant \lambda_1'(\Gamma).$$

However, we can compare $\lambda_1'(\Gamma)$ and $\lambda_1(\Gamma)$, namely we have $\lambda_1'(\Gamma) \geqslant d^{-1}\lambda_1(\Gamma)$. Indeed, the subspace $L_0^2 = (\mathbf{C}\cdot 1)^\perp \subset L^2(\Gamma)$ is stable under the linear maps B_Γ, Δ_Γ and $2d - \Delta_\Gamma$, and each of these is invertible on this subspace. We have then

$$\begin{aligned}\frac{1}{\lambda_1'(\Gamma)} = \|(B_\Gamma|L_0^2)^{-1}\| &\leqslant d^2\|(\Delta_\Gamma|L_0^2)^{-1}\|\ \|((2d - \Delta_\Gamma)|L_0^2)^{-1}\| \\ &\leqslant d\|(\Delta_\Gamma|L_0^2)^{-1}\| = \frac{d}{\lambda_1(\Gamma)},\end{aligned}$$

as claimed. Combining these inequalities, we conclude that

$$\lambda_1 \geqslant \min\Big(c, \frac{\eta}{d}\lambda_1(\Gamma)\Big),$$

which finishes the proof, up to the two facts stated above.

We now justify these two claims. For Fact 1, we note that the infimum considered, say $\eta(\mathcal{D})$, is nothing but the second eigenvalue for the Laplace operator with *Neumann* boundary condition on $\mathcal{D}$ (see, e.g., [22, 8.3.1]). This operator is positive and has first eigenvalue 0 with multiplicity 1, so $\eta(\mathcal{D}) > 0$. To see that $\eta(\mathcal{D})$ is in fact bounded below by a positive constant *that depends only on* M (and not on N), we fix some $\widetilde{x} \in \widetilde{M}$ above x_0. The reasoning in [19, p. 71] shows that $\mathcal{G}(x)$ is isometric to a quotient of the domain

$$\widetilde{N} = \bigcup_{s \in T} s\widetilde{\mathcal{F}}(\widetilde{x}_1) \subset \widetilde{M}$$

which depends only on M, under an equivalence relation $\sim$ of congruence modulo a subset of T^3. Each such quotient $\widetilde{N}/\sim$ is a compact domain in $\widetilde{M}$, hence has a second Neumann eigenvalue that is > 0. Since T is finite, there are only finitely many quotients $\widetilde{N}/\sim$ to consider, where the number depends only on M. It follows that the smallest of their second Neumann eigenvalue, say η, is > 0, and we have $\eta(\mathcal{D}) \geqslant \eta$ for all $\mathcal{D}$, proving Fact 1.

For Fact 2, we note that to find the required test function f it is enough to know that the map Φ is injective on the $\mathbf{R}$-span of 1 and ψ. Indeed, we can then find a non-zero $f = \alpha + \beta\psi$ in the kernel of the linear form $f \mapsto \langle \Phi(f), 1\rangle$, since this linear form is non-trivial (it maps 1 to $|P|$), and this f will satisfy the required conditions.

Now we have two cases. If $\lambda_1 \geqslant \eta$, where η is given by Fact 1, we are done (and take $c = \eta$). Otherwise, we have

$$0 < \lambda_1 < \eta, \tag{5.16}$$

and we now show that this implies that Φ is injective on the (real) span of 1 and ψ, which thus concludes the proof.

Thus, let α, $\beta \in \mathbf{R}$ be such that $\Phi(f) = \Phi(\alpha + \beta\psi) = 0$. Then, for all $x \in P$, we have

$$\int_{\mathcal{F}(x)} \|\nabla f\|^2 d\mu \geqslant \eta \int_{\mathcal{F}(x)} f^2 d\mu,$$

since $\Phi(f) = 0$ means that f restricted to $\mathcal{F}(x)$ can be used to test (5.14). Summing over x, we get $\|\nabla f\|^2 \geqslant \eta\|f\|^2$, but $f = \alpha + \beta\psi$ implies then that

$$\eta\|f\|^2 \leqslant \|\nabla f\|^2 = \beta^2\lambda_1 \leqslant \lambda_1\|f\|^2,$$

and by comparing with (5.16), we see that $f = 0$. □

5.5. Diophantine applications

We finish this chapter with a survey of some of the arithmetic consequences of (variants of) the gonality lower bound of Theorem 5.4.3. In this section, we will assume some familiarity with basic concepts of algebraic number theory and arithmetic geometry (for instance, the elementary arithmetic properties of elliptic curves over number fields, for which the standard reference is Silverman's book [105], while a quick survey can be found, e.g., in [67, §1]). We do not explain the most general statements, which can be found in the original paper of Ellenberg, Hall and Kowalski [34]. Another readable account is in the survey of Ellenberg [35].

The "base" Riemann surfaces we consider are very simple, but they are not compact. Precisely, X is the complement in $\mathbf{C}$ of k points points $\{s_1, \ldots, s_k\}$ (with $k \geqslant 2$, for simplicity). We wish the singularities to be defined over $\mathbf{Q}$, which means that $\{s_1, \ldots, s_k\}$ is the zero set of a polynomial with *rational* coefficients.

We consider some family of finite coverings $\pi_n \colon X_n \to X$, where we assume that X_n has the structure of an algebraic variety defined over $\mathbf{Q}$, and that the covering map π_n is also algebraic and defined over $\mathbf{Q}$. This means that, for some integers $N \geqslant 1$ and $m \geqslant 1$ (that may depend on n), the Riemann surface X_n can be viewed as the set of common zeros in $\mathbf{C}^N$ of m polynomials equations (in N variables), and that π_n is itself the restriction to X_n of a polynomial in $\mathbf{Q}[z_1, \ldots, z_N]$ (which happens to never take any of the values $\{s_1, \ldots, s_k\}$ when evaluated on X_n, so that the image is really contained in X). We also assume, as before, that the degree of X_n over X tends to infinity as $n \to +\infty$.

One can still speak of the gonality of X_n, with the same definition as in the previous section. This is a geometric invariant but it turns out to have arithmetic significance, as the next theorem will show. To state it, for any field extension $k/\mathbf{Q}$ and $n \geqslant 1$ we denote by $X_n(k)$ the set of all points $x \in X_n$ such that all coordinates of x belong to k (one says that "x is defined over k").

Theorem 5.5.1 (Faltings; Frey). *Let $d \geqslant 1$ be an integer. Suppose that X_n is connected and that the gonality γ of X_n is $\geqslant 2d$. Then the union*

$$\bigcup_{\substack{k/\mathbf{Q} \\ [k:\mathbf{Q}]\leqslant d}} X_n(k)$$

is finite. In other words, there are only finitely many $x \in X_n$ such that x is defined over an extension $k/\mathbf{Q}$ of degree $\leqslant d$.

In particular, there exist only finitely many algebraic numbers $t \in X$ which are of the form $t = \pi_n(x)$ for some $x \in X_n$ defined over k with $[k : \mathbf{Q}] \leqslant d$.

Remark 5.5.2. (1) To be precise, we have specialized to the current situation and notation the actual result of Faltings and Frey, which is more general. The basic ingredient of the proof is a deep theorem of Faltings [37], from which Frey deduces the above result [40] (see also the paper [2] of Abramovich and Voloch).

(2) It is easy to see that if X_n has gonality d, then it will have infinitely points with coordinates in fields of degree $\leqslant d$ over $\mathbf{Q}$. For instance, suppose that X_n is given by a so-called superelliptic equation $y^d = f(x)$, with $d \geqslant 2$ and $f \in \mathbf{Q}[X]$ non-constant. The map $X_n \to \mathbf{C}$ defined by $(x, y) \mapsto x$ has degree d, since the pre-images of any $x \in \mathbf{C}$ are the roots of the equation $y^d = f(x)$, and there are exactly d of them for most values of x. So the gonality of X_n is at most d. But for the same reason, any choice of $x \in \mathbf{Q}$ such that $f(x) \neq 0$ will give rise to d elements $(x, y) \in X_n(\mathbf{Q}(\sqrt[d]{x}))$. Since $\mathbf{Q}(\sqrt[d]{f(x)})$ is an extension of degree $\leqslant d$, we see that X_n has infinitely many points in extensions of degree $\leqslant d$.

The very deep content of the theorem is that this elementary remark "almost" accounts for *all* cases where an algebraic curve like X_n has infinitely many points of small degree.

To apply the theorem of Faltings–Frey, one needs to have growth of gonality. For this, we will use expansion. The fundamental group G of $X = \mathbf{C} - \{s_1, \ldots, s_k\}$ is a free group with k generators (see, e.g., [10, Ch. 4, Exemple 3, p. 419]). In the general setting described above, the covering $X_n \to X$ is not necessarily a Galois covering, but by considering the Galois closure $Y_n \to X$ and its Galois group over X, we obtain a sequence of finite groups $G_n = \mathrm{Gal}(Y_n/X)$ and of subgroups $\mathrm{Gal}(Y_n/X_n) \subset G_n$ such that $[G_n : H_n] = \deg(X_n \to X)$, each of which is a quotient of the fundamental group G. These families of groups satisfy a generalization of Theorem 5.4.3:

Theorem 5.5.3 (Ellenberg, Hall, Kowalski). *Suppose that there exists a finite symmetric generating set S of G such that the family of Schreier graphs $(\mathcal{A}(H_n\backslash G_n, S))_n$ is an expander family. Then we have*

$$\gamma(X_n) \gg [G_n : H_n]$$

for all n, where the implied constant depends only on X and S.

Combining this growth bound with the last statement in the theorem of Faltings and Frey, we deduce:

Corollary 5.5.4. *Suppose that there exists a finite symmetric generating set S of G such that the family of Schreier graphs $(\mathcal{A}(H_n\backslash G_n, S))_n$ is an expander family. Then, for all $d \geqslant 1$, and for all n large enough depending on d, there exist only finitely many algebraic numbers $t \in X$ which are of the form $t = \pi_n(x)$ for some $x \in X_n$ defined over $k/\mathbf{Q}$ with $[k : \mathbf{Q}] \leqslant d$.*

Remark 5.5.5. (1) In fact, inspection of the proof of Theorem 5.5.3 shows that for many applications (including those we will present in Examples 5.5.6 and 5.5.9 below), the expansion condition on the family of Cayley graphs may be relaxed to the weaker one that they form an *esperantist* family, in the sense of Definition 3.5.5. Although there is currently no “natural” application (in the context of this section) where this condition holds without the family being an expander, it is worth noting that it is often significantly easier to prove that a family is esperantist than that it is to prove that it forms an expander. For instance, in the context of Chapter 6, the esperantist condition corresponds to Helfgott's Growth Theorem 6.6.1, which is the first step in the proof of the Bourgain–Gamburd Theorem 6.1.1. So, for many diophantine applications, one does not require the more difficult steps of the expansion proofs.

(2) M. Orr [91, Th. 1.4] has recently derived other arithmetic consequences from the growth of gonality in suitable families that do not involve Theorem 5.5.1.

We will conclude this section with two illustrative applications.

Example 5.5.6. The following concrete example already exhibits quite well some of the general features of this topic. Before beginning, it is also worth pointing out that *in this special case*, there are more direct approaches to the growth of gonality, using the theory of modular curves and work of Abramovich [1], Zograf [121] and Poonen [96].

Let $X = \mathbf{C} - \{0, 1\}$. For $t \in X$, we denote by E_t the Legendre elliptic curve with affine equation

$$y^2 = x(x-1)(x-t)$$

(we again recommend [105] for an introduction to elliptic curves). Recall that the set of points of E_t, together with a point at infinity (that we denote simply ∞, although one should think that it depends on t), is an abelian group. One knows that for each integer n, the set $E_t[n]$ of solutions $(x, y) \in E_t$ of the equation $n(x, y) = 0$ is a finite group isomorphic to $(\mathbf{Z}/n\mathbf{Z})^2$ (these points are called n-torsion points on E_t; the reason for this isomorphism is that the whole groupe of $\mathbf{C}$-valued solutions is isomorphic to $(\mathbf{R}/\mathbf{Z})^2$, where the n-torsion points are of the form $(a/n, b/n)$ for a and b modulo n). Moreover, because the formulas for the addition of points of E_t are rational functions with rational coefficients, the solutions $(x, y) \in E_t[n]$ have algebraic coordinates. One can deduce that putting

$$X_n = \{(t, (x, y)) \mid (x, y) \in E_t[n]\}$$

and defining $\pi_n(t, (x, y)) = t$, one obtains coverings $\pi_n \colon X_n \to X$ of the type described in the beginning of this section. The fiber $\pi_n^{-1}(t)$ over t is then $E_t[n]$.

The type of conclusions that one can *hope* to get from Theorem 5.5.1 in that case is that, for most choices of t, solving the equation $n(x, y) = 0$ in E_t will require introducing new algebraic numbers that generate a high-degree extension of $\mathbf{Q}(t)$. This is not unexpected, since a priori the relevant equation is of degree n^2, but one must exclude “miracles” where this equation would factor and reduce to much simpler ones.

We can easily understand some exceptions in that special case. Indeed, it is a standard fact that $(x, y) \in E_t[2]$ if and only if either $(x, y) = \infty$ (which is in $E_t(\mathbf{Q})$) or $y = 0$. So X_2 is simply a (disconnected) product

$$X_2 = X \times (\mathbf{Z}/2\mathbf{Z})^2$$

with $E_t[2] = \{(0, 0), (1, 0), (t, 0), \infty\}$. In this case, solving $2(x, y) = 0$ requires no extension of the field $\mathbf{Q}(t)$ containing the coefficient t. However, if one writes down the equations for $(x, y) \in E_t[3]$, for instance, one doesn't see any obvious solution. In fact, one can show (this is a theorem of Igusa) that X_n is connected if n is odd, and that the Galois group of the Galois closure $Y_n \to X$ of the covering $X_n \to X$ is $\mathrm{SL}_2(\mathbf{Z}/n\mathbf{Z})$.

We now want to apply Theorem 5.5.3. We first check for which subfamilies of the coverings $(X_n \to X)$ the result is applicable. If n is odd, we just mentioned that the Galois group of the Galois closure of X_n is $G_n = \mathrm{SL}_2(\mathbf{Z}/n\mathbf{Z})$. It turns out that, in this case (and many similar situations), there exists a single infinite Galois covering $Y \to X$ with Galois group a finite index subgroup Γ of $\mathrm{SL}_2(\mathbf{Z})$ such that each Y_n arises as the covering associated to the normal subgroup

$$\ker(\Gamma \to \mathrm{SL}_2(\mathbf{Z}/n\mathbf{Z})) \subset \Gamma$$

(the kernel of reduction modulo n). If we pick a finite symmetric generating set S of Γ and consider the Cayley graphs of $G_n = \mathrm{SL}_2(\mathbf{Z}/n\mathbf{Z})$ with respect to the image of S, then:

(1) Taking all odd prime values of n, we are in the situation of the Bourgain–Gamburd Theorem 6.1.1 (proved in Chapter 6), and therefore the corresponding family is an expander.

(2) If we take all odd squarefree values of n, then we can appeal to the Bourgain–Gamburd–Sarnak Theorem (the case $\mathbf{G} = \mathrm{SL}_2$ of Theorem 4.3.8) instead, and have the same conclusion.

(3) Finally, taking all n odd, we are in the situation of the Bourgain–Varjú Theorem (the case $\mathbf{G} = \mathrm{SL}_2$ of the final statement of Theorem 4.3.8), hence the corresponding family is again an expander.

We've mentioned all three cases, although of course the last encompasses the other, because the analogue of the Bourgain–Varjú Theorem might not be known in other situations, and sometimes it is not needed for interesting applications, as we will now see.

We apply Theorem 5.5.3 first to the family $(X_n \to X)_{n \text{ odd}}$. This shows that the gonality of X_n tends for infinity. By the Faltings–Frey Theorem, we conclude that if n is large enough, the set

$$\bigcup_{[k:\mathbf{Q}]\leqslant d} \pi_n(X_n(k))$$

is finite. In other words, having fixed d, if $n \geqslant 1$ is large enough in terms of d, there are only finitely many Legendre curves E_t with $[\mathbf{Q}(t) : \mathbf{Q}] \leqslant d$ such that $E_t(\mathbf{Q}(t))$ contains the n-torsion points of E_t. A finer argument (see [34, Th. 7]) shows that in fact the same conclusion holds for those t where $E_t(\mathbf{Q}(t))$ contains a single non-zero n-torsion point.

We now discuss an even more interesting application of these results. We recall that an endomorphism u of an elliptic curve E is a group morphism $E \to E$ that is also algebraic. In particular, the endomorphisms $(x, y) \mapsto n(x, y)$ of multiplication by $n \in \mathbf{Z}$ are always of this type. Elliptic curves over a field of characteristic zero are of two kinds: those curves E that have no other endomorphisms than multiplication by integers, and those that do. These last special curves are called "curves with complex multiplication", or CM curves for short.

The CM curves are indeed extremely special. If we restrict to the Legendre curves E_t, it is not difficult to prove that any t for which E_t has CM is algebraic. One example is

$$E_{-1}\colon y^2 = x(x-1)(x+1) = x^3 - x$$

(where an "extra" endomorphism, if we take the base field to contain i in addition to $\mathbf{Q}$, is $(x, y) \mapsto (-x, iy)$).

A fascinating problem with a long history, although in the disguise of counting imaginary quadratic fields with bounded class numbers, is to try to count how many CM curves E_t there are with $\mathbf{Q}(t)$ fixed (e.g., with $t \in \mathbf{Q}$) or with $[\mathbf{Q}(t) : \mathbf{Q}]$ bounded. (A wonderful discussion of this problem under many aspects is in the book [30] of Cox).

The key observation is that, for any n, the Galois action on torsion points $(x, y) \in E_t[n]$ must commute with any endomorphism of E_t that is defined over $\mathbf{Q}(t)$. If the curve has CM, then this becomes a strong restriction on the possible Galois group, hence it implies that computing torsion points is "not as complicated" as it is for curves without CM.

Following this idea, one can construct from the family $(Y_n \to X)$ some other auxiliary families of coverings $(\sigma_n\colon Z_n \to X)$ with the property that, if E_t has CM, then it follows that $t = \sigma_n(x)$ for some $x \in Z_n(\mathbf{Q}(t))$. Using expansion, one can prove again that the gonality of Z_n is increasing for n odd. Hence, as before, if we fix $d \geqslant 1$, then the set

$$\bigcup_{[k:\mathbf{Q}]\leqslant d} \sigma_n(Z_n(k))$$

is finite if n is large enough. Picking one such n (depending on d), the previous observation allows us to conclude:

Corollary 5.5.7. *Let $d \geqslant 1$ be an integer. The set of $t \in \bar{\mathbf{Q}}$ with $[\mathbf{Q}(t) : \mathbf{Q}] \leqslant d$ such that E_t has CM is* finite.

Observe that, in this application, the final conclusion does not mention any family of coverings: these only played an auxiliary role. Moreover, since we are free to select n as we wish, we may take for instance a large enough prime, which means that the conclusion only depends on the (somewhat) easier Bourgain–Gamburd Theorem, and doesn't require its extension to squarefree n or to all (odd) n.

Here is a final re-interpretation of this conclusion. The theory of Complex Multiplication (see [30, Ch. 3]) is actually very precise concerning the fields $k/\mathbf{Q}$ such that there is some Legendre curve E_t with CM and $t \in k$. In particular, this theory implies that $[\mathbf{Q}(t) : \mathbf{Q}]$ is bounded in terms of the *class number* of the endomorphism ring of E_t, which one can show is a finite index subring of the ring of integers in an imaginary quadratic field. Conversely, for any imaginary quadratic field $E/\mathbf{Q}$, there exists a Legendre curve E_t with CM by E, and with $[\mathbf{Q}(t) : \mathbf{Q}]$ bounded (explicitly) in terms of the class number of E. In particular, we observe that Corollary 5.5.7 proves that if we fix some integer $d \geqslant 1$, then there can only be finitely many imaginary quadratic extensions of $\mathbf{Q}$ with class number $\leqslant d$ (since each of these would give rise to at least one Legendre curve E_t with CM and $[\mathbf{Q}(t) : \mathbf{Q}] \leqslant d$.)

Thus we have deduced a famous result of Deuring and Heilbronn, first conjectured by Gauss:

Theorem 5.5.8 (Deuring, Heilbronn). *As the absolute value of the discriminant tends to infinity, the class number of an imaginary quadratic field tends to infinity.*

We refer, e.g., to [58, p. 174, Ch. 22, Ch. 23] for the derivation of the "classical" proof of this result, based on Dirichlet's Class Number Formula and analytic properties of Dirichlet L-functions, and for further study of the (still fascinating) problem of making the statement *effective*, namely to give an effective "explicit" lower bound for the class number of an imaginary quadratic field with discriminant d. Whereas the Generalized Riemann Hypothesis swiftly implies that the class number is roughly of size $|d|^{1/2}$, the best known result (due to Goldfeld and Gross–Zagier, see, e.g., [58, Ch. 23]) is rather less impressive, since it is of size merely $\log |d|$ (or even a bit smaller!). This is also closely related to the Siegel–Walfisz Theorem 5.3.8.

Example 5.5.9. Our final example (see [34, Th. 6]) is still related to Legendre curves, but we present it separately because now (in contrast to what was stated at the beginning of Example 5.5.6) there will be no way to avoid using expansion properties of Cayley graphs of the type proved by Helfgott, Bourgain and Gamburd (and others).

We have seen that, to construct the torsion points $E_t[n]$ of order n in a Legendre curve with t algebraic, one needs to solve polynomial equations of large degree. We can now ask if, for *two* Legendre curves, these equations are "the same". More precisely, how often can one find $t_1 \neq t_2$ in the same number field $k/\mathbf{Q}$ such that $E_{t_1}[n]$ and $E_{t_2}[n]$ are isomorphic as finite abelian groups with the action of the Galois group of k? Note that this would in particular mean that the coordinates of n-torsion points of E_{t_2} can be expressed in terms of those of E_{t_1} without introducing further extensions.

For concreteness, we consider the specific case where $t_2 = 2t_1$. Let $Y = X - \{1/2\} = \mathbf{C} - \{0, 1, 1/2\}$. One can construct a family of finite coverings $\tau_n \colon T_n \to Y$ such that, for any $t \in k$, if $E_t[n]$ and $E_{2t}[n]$ are isomorphic, then there exists $x \in T_n(k)$ such that $t = \tau_n(x)$. The coverings $T_n \to Y$ arise as quotients of an infinite covering $T \to Y$, whose Galois group was shown by Nori to be an *infinite index* subgroup of $\mathrm{SL}_2(\mathbf{Z}) \times \mathrm{SL}_2(\mathbf{Z})$, although it is Zariski-dense in $\mathrm{SL}_2(\mathbf{C}) \times \mathrm{SL}_2(\mathbf{C})$.

It is then from the most general work of Salehi-Golsefidy and Varjú [99](see Theorem 4.3.8) that we deduce that Theorem 5.5.3 is applicable to the family $(T_\ell \to Y)$, where ℓ runs over large enough primes. The arithmetic conclusion is the following:

Corollary 5.5.10. *For any integer $d \geqslant 1$, there exists ℓ_0 such that if $\ell \geqslant \ell_0$ is a prime number, then there are only finitely many t with $[\mathbf{Q}(t) : \mathbf{Q}] \leqslant d$ such that $E_t[\ell]$ and $E_{2t}[\ell]$ are isomorphic as finite groups with Galois action.*

Note that if $\ell = 2$, then $E_t[\ell]$ and $E_{2t}[\ell]$ are always isomorphic (since $E_t[2]$ is contained in $E_t(\mathbf{Q}(t))$, the Galois group acts trivially in that case).

CHAPTER 6

Expanders from $\mathrm{SL}_2(\mathbf{F}_p)$

6.1. Introduction

In this chapter, we set out to prove Theorem 4.3.2 on expansion of Cayley graphs of $\mathrm{SL}_m(\mathbf{F}_p)$, in the special case $m = 2$ (which is due to Bourgain and Gamburd [12]). We recall and rephrase the statement in a convenient way:

Theorem 6.1.1 (Expansion in subgroups of $\mathrm{SL}_2(\mathbf{Z})$). *Let $S \subset \mathrm{SL}_2(\mathbf{Z})$ be any finite symmetric subset and let G be the subgroup generated by S. For prime numbers p, let $\Gamma_p = \mathcal{C}(\mathrm{SL}_2(\mathbf{F}_p), S)$ be the Cayley graph of the finite group $\mathrm{SL}_2(\mathbf{F}_p)$ with respect to the reduction modulo p of the set S. Then $(\Gamma_p)_{p\geqslant p_0}$ is an expander family if and only if Γ_p is connected for all p large enough.*

Although the arguments of the proof are, to a very large extent, elementary, they are quite subtle and involved. In the first section, we present the strategy and state some group-theoretic properties of $\mathrm{SL}_2(\mathbf{F}_p)$ which underlie the whole argument. Originally, the work of Bourgain and Gamburd was spurred by Helfgott's growth theorem (see Theorem 6.6.1 below, which implies Theorem 4.3.5) for the same groups (which suffices to show the weaker esperantist property). However, we present those two steps backwards: we explain first how Bourgain and Gamburd reduced the expansion property to Helfgott's theorem, and then prove the latter. Our justification for this is simply that this seems to the author to involve the gentlest learning curve.

As we stated in Section 4.3, Theorem 6.1.1 as well as Theorem 6.6.1 have been considerably generalized in recent years. We refer to Tao's book [109] for a complete account of the proof of much more general results.

Remark 6.1.2. (1) Both Helfgott's Theorem and the Bourgain-Gamburd method are completely effective and explicit, and one can use them to compute *actual* numerical bounds for the spectral gap of an explicitly given subgroup. This requires rather strenuous and tedious bookkeeping however; the reader may check in the paper [70] the quality of the resulting bounds.

(2) The proof involves a number of properties of the groups $\mathrm{SL}_2(\mathbf{F}_p)$. All these can be proved elementarily, but the full details are quite long, and we have chosen not to include all of them. Roughly speaking, we will prove most of what is needed to the growth theorem, but only state (with proper references) the results used in the Bourgain–Gamburd argument. This compromise is of course somewhat arbitrary...

By convention, throughout this chapter, when we consider Cayley graphs $\mathcal{C}(G, S)$, we assume that S is not empty (so that in particular the Cayley graph has no isolated vertices, and we may speak of its Markov operator).

6.2. Preliminaries and strategy

Theorem 6.1.1 will be proved using the spectral definition of expanders, and in fact by appealing to ideas involving random walks. The fundamental idea is to try to detect the spectral gap by looking at the spectral decomposition of a certain specific quantity built using the Markov operator. The following choice turns out to be very efficient (maybe because of its simplicity from the spectral point of view?)

Lemma 6.2.1 (Counting cycles). *Let $\Gamma = (V, E, \mathrm{ep})$ be a finite non-empty connected graph without isolated vertices. Let M be the Markov operator of Γ and let $(X_n)_{n\geqslant 0}$ be a random walk on Γ. For any integer $m \geqslant 0$, we have*

$$\mathrm{Tr}(M^m) = \sum_{x\in V} \mathbf{P}(X_m = x \mid X_0 = x).$$

In particular, if $\Gamma = \mathcal{C}(G, S)$ is a Cayley graph and the random walk starts at $X_0 = 1$, we have

$$\frac{1}{|G|}\,\mathrm{Tr}(M^m) = \mathbf{P}(X_m = 1), \tag{6.1}$$

the probability of returning to the identity after m steps.

Proof. For the sake of variety, we use a relatively non-standard proof (the reader is invited to do this more straightforwardly!). Let (φ_i) be an orthonormal basis of eigenfunctions of M in $L^2(\Gamma)$, with $M\varphi_i = \lambda_i\varphi_i$. Then the trace of M^m is equal to the sum of the λ_i^m. By orthonormality, we can write

$$\sum_i \lambda_i^m = \sum_i \langle\varphi_i, \varphi_i\rangle \lambda_i^m = \sum_i \Bigl(\frac{1}{N}\sum_{x\in V} \mathrm{val}(x)|\varphi_i(x)|^2\Bigr)\lambda_i^m,$$

where N is the sum of the valencies, as in Definition 3.2.11.

After exchanging the sums, the inner summand is the diagonal value $g(x, x)$ of the function

$$g(x, y) = \frac{\mathrm{val}(x)}{N}\sum_i \lambda_i^m \overline{\varphi_i(x)}\varphi_i(y).$$

But, for any $x \in V$, the expression

$$\frac{\mathrm{val}(x)}{N}\sum_i \overline{\varphi_i(x)}\varphi_i$$

is the spectral expansion in the basis (φ_i) of the characteristic function δ_x of the single point $x \in V$. Since the φ_i are eigenfunctions of M with eigenvalue λ_i, linearity implies that

$$M^m\delta_x = \frac{\mathrm{val}(x)}{N}\sum_i \lambda_i^m \overline{\varphi_i(x)}\varphi_i.$$

Evaluating this expression at x, we deduce that

$$\sum_i \lambda_i^m = \sum_{x\in V} (M^m\delta_x)(x).$$

By the basic property of the Markov operator (Lemma 3.2.16), we know that

$$(M^m\delta_x)(x) = \mathbf{P}(X_m^{(x)} = x),$$

where $(X_m^{(x)})$ refers to a random walk started at x. By the Markov property, this is the same as $\mathbf{P}(X_m = x \mid X_0 = x)$ for an arbitrary random walk, and hence

$$\sum_i \lambda_i^m = \sum_{x\in V} \mathbf{P}(X_m = x \mid X_0 = x).$$

When Γ is a Cayley graph, the probability $\mathbf{P}(X_m = x \mid X_0 = x)$ is independent of the starting point x, by homogeneity. Hence, selecting $x = 1$, we get

$$\mathbf{P}(X_m = 1) = \frac{1}{|G|}\sum_i \lambda_i^m$$

for a random walk starting at $X_0 = 1$. □

Example 6.2.2. Let $\Gamma = \mathcal{C}(G, S)$ be a Cayley graph, and consider the random walk starting at the identity, i.e., $X_0 = 1$. Since the steps of this random walk are obtained by multiplication with a generator $s \in S$ which is uniformly chosen, we see that we have a concrete combinatorial description

$$\frac{1}{|G|}\operatorname{Tr}(M^m) = \mathbf{P}(X_m = 1) = \frac{1}{|S|^m}|\{(s_1, \ldots, s_m) \in S^m \mid s_1 \cdots s_m = 1\}|,$$

which is the number of "relations" in G of length m when presenting the group using the generators from S.

Using the expression of the trace as a sum of eigenvalues, and the non-negativity of squares of real numbers, this lemma leads to the following corollary:

Corollary 6.2.3. *Let $\Gamma = \mathcal{C}(G, S)$ be a finite connected Cayley graph with Markov operator M, and let $(X_n)_{n\geqslant 0}$ be a random walk on Γ with $X_0 = 1$ fixed. Let $\Lambda \subset [-1, 1]$ be the eigenvalues of M, with multiplicity $n(\lambda) \geqslant 1$ for $\lambda \in \Lambda$. For any subset $\Lambda_1 \subset \Lambda$ and for any integer $m \geqslant 0$, we have*

$$\frac{1}{|G|}\sum_{\lambda\in\Lambda_1} n(\lambda)\lambda^{2m} \leqslant \mathbf{P}(X_{2m} = 1),$$

with equality if $\Lambda = \Lambda_1$.

This gives an upper bound for any eigenvalue λ, and in particular for the equidistribution radius ϱ_Γ, *provided* one can usefully estimate $\mathbf{P}(X_{2m} = 1)$. The latter is the trickiest part, and the argument would not work if the multiplicities $n(\lambda)$ did not bring a little help. However, they do for groups which are "complicated", in a certain specific sense, for the following simple reason: the group G, through its action on its Cayley graphs, also act on each eigenspace of M, and can not have invariant vectors except for the 1-eigenspace.

Proposition 6.2.4 (Representation of G on $L^2(G)$). *Let G be a finite group, $S \subset G$ a finite symmetric generating set.*

(1) *The group G acts by linear automorphisms on $L^2(G)$ by means of the left-regular representation*

$$\operatorname{reg}(g)\varphi(x) = \varphi(g^{-1}x)$$

for all $x \in G$ and $g \in G$. This is a unitary representation

$$\operatorname{reg} : G \longrightarrow \mathrm{U}(L^2(G)).$$

(2) *The regular representation commutes with the Markov operator of $\mathcal{C}(G, S)$, i.e.,*

$$M(\operatorname{reg}(g)\varphi) = \operatorname{reg}(g)(M\varphi)$$

for every $g \in G$ and $\varphi \in L^2(G)$. In particular, each eigenspace $\ker(M - \lambda) \subset L^2(G)$ is a subrepresentation of the regular representation. This subrepresentation contains an invariant vector, i.e., some non-zero $\varphi \in L^2(G)$ such that $\operatorname{reg}(g)\varphi = \varphi$ for all $g \in G$, if and only if $\lambda = 1$.

Proof. The first part is a formal computation, which we leave to the reader. The beginning of Part (2) is due to the fact that the regular representation involves multiplication on the left, while we walk on $\mathcal{C}(G,S)$ by multiplying on the right with elements of S, and the two sides of multiplication commute. Precisely, we have

$$M(\mathrm{reg}(g)\varphi)(x)=\frac{1}{|S|}\sum_{s\in S}\varphi(g^{-1}xs)=\mathrm{reg}(g)(M\varphi)(x).$$

From this, the stability of $\ker(M-\lambda)$ is easy: if $M\varphi=\lambda\varphi$, we have also

$$M(\mathrm{reg}(g)\varphi)=\mathrm{reg}(g)(M\varphi)=\mathrm{reg}(g)(\lambda\varphi)=\lambda\,\mathrm{reg}(g)\varphi,$$

i.e., $\mathrm{reg}(g)\varphi\in\ker(M-\lambda)$ for all $g\in G$, as claimed. Now if φ is any non-zero function invariant under the action of G, we get

$$\varphi(x)=(\mathrm{reg}(x)\varphi)(1)=\varphi(1)$$

for all x, which means that φ is constant. But then $M\varphi=\varphi$, so this only happens when $\lambda=1$. □

The innocuous-looking corollary is the following:

Corollary 6.2.5 (Bounding ϱ_Γ from return probability). *Let G be a finite group, and define $d(G)$ to be the minimal dimension of a* non-trivial *unitary representation of G. Let $\Gamma=\mathcal{C}(G,S)$ be a finite connected Cayley graph with Markov operator M, and let $(X_n)_{n\geqslant 0}$ be a random walk on Γ with $X_0=1$ fixed. Then*

$$\varrho_\Gamma\leqslant\Big(\frac{|G|}{d(G)}\,\mathbf{P}(X_{2m}=1)\Big)^{1/2m}$$

for all $m\geqslant 0$.

Proof. Indeed, ϱ_Γ is an eigenvalue $\neq 1$ of M, and hence the representation of G on the ϱ_Γ-eigenspace of M is non-trivial. Thus the multiplicity of λ as an eigenvalue is at least equal to $d(G)$, and we can apply the previous corollary with $\Lambda_1=\{\varrho_\Gamma\}$. □

The point of this result is that there exist groups for which $d(G)$ is indeed rather large in comparison with their size. We will apply this, in Section 6.4, to one such family, and we state here the corresponding result (for a proof, see Proposition B.2.1, (3)):

Theorem 6.2.6 (Frobenius). *Let $\mathbf{F}_q$ be a finite field with q elements. If q is odd, we have*

$$d(\mathrm{SL}_2(\mathbf{F}_q))=\frac{q-1}{2}.$$

In particular, since $|\,\mathrm{SL}_2(\mathbf{F}_q)|=q(q-1)(q+1)$, we have

$$d(\mathrm{SL}_2(\mathbf{F}_q))\sim|\,\mathrm{SL}_2(\mathbf{F}_q)|^{1/3}$$

as $q\to+\infty$.

This will be absolutely crucial for the Bourgain-Gamburd theorem. It should be noted that this property had already been used in situations involving spectral gaps for the hyperbolic Laplace operator. This goes back (although one must pay close attention to this paper to detect it!) to a work of Huxley [56, §4], and was first applied by Sarnak and Xue [100] in contexts where other techniques to prove spectral gaps for the Riemannian Laplace operator (specifically, Fourier expansions around cusps and estimates for Kloosterman sums) were not available. Another very relevant application is found in Gamburd's thesis [43]. In fact, we will use again this property of $\mathrm{SL}_2(\mathbf{F}_p)$ at a further step, in order to exploit Gowers's notion of *quasi-random groups* (see Section 6.5).

Remark 6.2.7. Note that $d(G) \neq 1$ implies in particular that there can not be any surjective homomorphism $G \longrightarrow \{\pm 1\}$, and therefore that any Cayley graph of G is non-bipartite (by Proposition 2.3.6).

The corollary leads to expander-quality bounds in the following situation: if we have

$$\mathbf{P}(X_{2m} = 1) \leqslant |G|^{-1+\varepsilon}, \tag{6.2}$$

for a fixed $\varepsilon > 0$, when m is relatively small, of size $m \leqslant c \log |G|$, then we get

$$\varrho_\Gamma \leqslant \exp\Bigl(\frac{\varepsilon \log |G| - \log d(G)}{2c \log |G|}\Bigr).$$

If, as happens for $G = \mathrm{SL}_2(\mathbf{F}_q)$, the group is such that $\log d(G) \geqslant d \log |G|$ with a fixed $d > 0$, and if one can select $c > 0$ so that the bound (6.2) holds with ε arbitrarily small, in particular $\varepsilon < d$, this leads to a uniform upper-bound for ϱ_Γ, namely

$$\varrho_G \leqslant \varrho = \exp\Bigl(-\frac{d-\varepsilon}{2c}\Bigr) < 1.$$

Here is an intermediate summary of this discussion:

Proposition 6.2.8. *For $d > 0$, $\varepsilon > 0$ and $c > 0$, let $\mathcal{G}_1(d, \varepsilon, c)$ be the family of all finite connected Cayley graphs $\mathcal{C}(G, S)$ where $d(G) \geqslant |G|^d$ and*

$$\mathbf{P}(X_{2m} = 1) \leqslant |G|^{-1+\varepsilon}$$

for some $m \leqslant c \log |G|$. Then for any $\varepsilon < d$ and any graph $\Gamma \in \mathcal{G}_1(d, \varepsilon, c)$, the equidistribution radius satisfies

$$\varrho_\Gamma \leqslant \exp\Bigl(-\frac{d-\varepsilon}{2c}\Bigr).$$

In particular, if $0 < \varepsilon < d$ and $c > 0$ are such that the family $\mathcal{G}_1(d, \varepsilon, c)$ contains graphs $\mathcal{C}(G, S)$ with arbitrarily large vertex sets G and bounded generating sets S, these form an expander family.

We may now ask two questions: What have we gained in attacking the problem this way? And is there a reasonable chance to make this work?

The first answer is that we have reduced the question of proving the asymptotic formula for $\mathbf{P}(X_{2m} = 1)$, which would follow from a bound on the equidistribution radius, to that of proving an *upper bound*, roughly of the right order of magnitude, for the same quantity. Indeed, we know that $\mathbf{P}(X_{2m} = 1)$ converges to $1/|G|$ as $m \to +\infty$ (or $2/|G|$ if we want to keep track of the bipartite case) and in view of the effective equidistribution statement, the probability should be already of the right order of magnitude when m is a fixed (possibly large) multiple of $\log |G|$ (see Example 3.2.30).

It may be helpful to keep in mind the concrete interpretation of the probability $\mathbf{P}(X_{2m} = 1)$: we want to prove that

$$\frac{1}{|S|^{2m}} |\{(s_1, \ldots, s_{2m}) \in S^{2m} \mid s_1 \cdots s_{2m} = 1\}| \leqslant |G|^{-1+\varepsilon}$$

for m of size $c \log |G|$.

In effect, the result of Bourgain-Gamburd can then be stated as follows:

Theorem 6.2.9 (Bourgain-Gamburd). *Fix $S \subset \mathrm{SL}_2(\mathbf{Z})$ a finite symmetric subset such that the projection of S modulo p generates $\mathrm{SL}_2(\mathbf{F}_p)$ for $p \geqslant p_0$. Then, for any $\varepsilon > 0$, there exists $c > 0$, depending on S and ε, such that*

$$\mathcal{C}(\mathrm{SL}_2(\mathbf{F}_p), S) \in \mathcal{G}_1(\tfrac{1}{3}, \varepsilon, c)$$

for all $p \geqslant p_0$.

6.3. The Bourgain-Gamburd argument

This section contains the great new ingredient discovered by Bourgain and Gamburd that out turns to open the door to implementing the general strategy discussed in the previous section. This is called the "L^2-flattening lemma" by Bourgain, Gamburd and Sarnak.

In rough outline – and probabilistic language –, the idea is to show that if a $\mathrm{SL}_2(\mathbf{F}_p)$-valued symmetrically distributed random variable X is not too concentrated, but also not very uniformly distributed on $\mathrm{SL}_2(\mathbf{F}_p)$, then a product of two independent "copies" of X will be significantly more uniformly distributed, *unless* there are obvious reasons why this should fail to hold. These exceptional possibilities can then be handled separately.

Applying this to some suitable step X_k of the random walk, the result of Bourgain-Gamburd leads to successive great improvements of the uniformity of the distribution for X_{2k}, X_{4k}, ..., X_{2^jk}, until the assumptions of the lemma fail. In that situation, it will be seen that $m = 2^jk$ is of size $\ll \log|G|$, and that $\mathbf{P}(X_m = 1)$ satisfies (6.2), and one can conclude.

To begin, we introduce an invariant which measures the qualitative property of concentration and uniformity mentioned in the informal description above.

Definition 6.3.1 (Return probabilities). Let G be a finite group and let X be a G-valued random variable which is symmetrically distributed, i.e.,

$$\mathbf{P}(X = g) = \mathbf{P}(X = g^{-1})$$

for all $g \in G$. The *return probability* $\mathrm{rp}(X)$ is defined as

$$\mathrm{rp}(X) = \mathbf{P}(X_1X_2 = 1),$$

where (X_1, X_2) are independent random variables with the same distribution as X. Equivalently, we have

$$\mathrm{rp}(X) = \sum_{g\in G} \mathbf{P}(X = g)^2.$$

If X, Y are two G-valued random variables, we denote

$$\mathrm{rp}^+(X, Y) = \max(\mathrm{rp}(X), \mathrm{rp}(Y)).$$

Example 6.3.2 (Uniform measures). If $A \subset G$ is any subset and X is uniformly distributed on A (equivalently, if we consider the measure ν given by

$$\nu(g) = \begin{cases} \frac{1}{|A|} & \text{if } g \in A \\ 0 & \text{otherwise} \end{cases}$$

for $g \in G$) then we have $\mathrm{rp}(X) = \frac{1}{|A|}$.

We will start by proving a general inequality which follows from the method of Bourgain and Gamburd, but still applies to general Cayley graphs (see Theorem 6.3.5 below, which it would be awkward to state now). This provides an approach to estimate $\mathrm{rp}(X_1X_2)$ in terms of $\mathrm{rp}^+(X_1, X_2)$ for fairly general independent symmetric G-valued random variables X_1 and X_2. Only afterwards will we use special features of the groups $\mathrm{SL}_2(\mathbf{F}_p)$.

Thus let X_1 and X_2 be symmetrically-distributed and independent. (The original method of Bourgain and Gamburd corresponds to situations where X_1 and X_2 are identically distributed; we then have $\mathrm{rp}^+(X_1, X_2) = \mathrm{rp}(X_1) = \mathrm{rp}(X_2)$).

By definition, we have

$$\mathrm{rp}(X_1X_2) = \sum_{g\in G} \mathbf{P}(X_1X_2 = g)^2.$$

To estimate this, we observe that it is partly the lack of uniformity of the distribution of the random variables which makes it difficult to understand what happens. To get some control on this lack of uniformity, a common strategy in analysis is to decompose the range of values of the density functions

$$\nu_i(x) = \mathbf{P}(X_i = x)$$

into intervals where their variation is within by a fixed (multiplicative) factor. Because we only attempt to estimate $\mathrm{rp}(X_1X_2)$ (and not to find an asymptotic formula), losing control of such a fixed factor is typically not a catastrophic loss.

It is most usual to consider dyadic intervals, i.e., intervals of the form $]a, 2a]$. One also wishes to avoid considering too many dyadic intervals, because they will be handled separately, and one must be able to afford losing a factor equal to the number of intervals. This means one should treat separately the very small values of the densities. We therefore consider a parameter $I \geqslant 1$, to be chosen later, and decompose

$$[\min \mathbf{P}(X = x), \max \mathbf{P}(X = x)] \subset [0, 1] = \mathcal{I}_0 \cup \mathcal{I}_1 \cup \cdots \cup \mathcal{I}_I$$

where $\mathcal{I}_i$ is, for $0 \leqslant i < I$, the dyadic interval

$$\mathcal{I}_i =]2^{-i-1}, 2^{-i}],$$

and the final complementary interval is $\mathcal{I}_I = [0, 2^{-I}]$, to account for the small values. This gives corresponding partitions of G in subsets

$$A_{1,i} = \{x \in G \mid \nu_1(x) = \mathbf{P}(X_1 = x) \in \mathcal{I}_i\},$$
$$A_{2,i} = \{x \in G \mid \nu_2(x) = \mathbf{P}(X_2 = x) \in \mathcal{I}_i\},$$

about which we note right now that, for $0 \leqslant i < I$, we have a rough size estimate

$$|A_{j,i}| \leqslant 2^{i+1}, \qquad j = 1, 2. \tag{6.3}$$

Note also that

$$\mathbf{P}(X_1 \in A_{1,I}) = \sum_{x\in A_{1,I}} \mathbf{P}(X_1 = x) \leqslant \frac{|A_{1,I}|}{2^I} \leqslant \frac{|G|}{2^I},$$

and similarly $\mathbf{P}(X_2 \in A_{2,I}) \leqslant |G|2^{-I}$. Using the partition above and the definition of $\mathrm{rp}(X_1X_2)$, we obtain

$$\begin{aligned}\mathrm{rp}(X_1X_2) &= \sum_{g\in G}\Bigl(\sum_{0\leqslant i,j\leqslant I} \mathbf{P}(X_1X_2 = g, X_1 \in A_{1,i},\ X_2 \in A_{2,j})\Bigr)^2 \\ &\leqslant 8|G|^3 2^{-2I} + 2\sum_{g\in G}\Bigl(\sum_{0\leqslant i,j<I} \mathbf{P}(X_1X_2 = g, X_1 \in A_{1,i},\ X_2 \in A_{2,j})\Bigr)^2 \\ &\leqslant 2^{3-2I}|G|^3 + 2I^2 \sum_{0\leqslant i,j<I}\sum_{g\in G} \mathbf{P}(X_1X_2 = g, X_1 \in A_{1,i},\ X_2 \in A_{2,j})^2\end{aligned}$$

by the Cauchy-Schwarz inequality. Furthermore, the inner sums in the second term, say $B(A_{1,i}, A_{2,j})$ are given by

$$\begin{aligned}
B(A_{1,i}, A_{2,j}) &= \sum_{g\in G} \mathbf{P}(X_1X_2 = g,\ X_1 \in A_{1,i},\ X_2 \in A_{2,j})^2 \\
&= \sum_{g\in G} \Bigl(\sum_{\substack{(x,y)\in A_{1,i}\times A_{2,j} \\ xy=g}} \mathbf{P}(X_1 = x)\,\mathbf{P}(X_2 = y) \Bigr)^2 \\
&= \sum_{\substack{x_1,x_2\in A_{1,i}, y_1,y_2\in A_{2,j} \\ x_1y_1=x_2y_2}} \nu_1(x_1)\nu_1(x_2)\nu_2(y_1)\nu_2(y_2) \\
&\leqslant 2^{-2i-2j} |\{(x_1, x_2, y_1, y_2) \in A_{1,i}^2 \times A_{2,j}^2 \mid x_1y_1 = x_2y_2\}|,
\end{aligned}$$

using the independence of X_1 and X_2 and the dyadic decomposition. The last quantity has a name, going back to Gowers at least:

Definition 6.3.3 (Multiplicative energy). Let G be a finite group and A, B subsets of G. The *multiplicative energy* $E(A, B)$ is given by

$$E(A, B) = |\{(x_1, x_2, y_1, y_2) \in A^2 \times B^2 \mid x_1y_1 = x_2y_2\}|,$$

and the *normalized multiplicative energy* is either 0 if A or B is empty, and otherwise is given by

$$e(A, B) = \frac{|E(A, B)|}{(|A||B|)^{3/2}}.$$

It may not be obvious that the normalization is the "correct" one, but this will become clear very soon. In any case, for the moment, we have shown that

$$\mathrm{rp}(X_1X_2) \leqslant 2^{3-2I}|G|^3 + 2I^2 \sum_{0\leqslant i,j<I} 2^{-2(i+j)} E(A_{1,i}, A_{2,j})$$

We now want to insert, for comparison, the return probability $\mathrm{rp}^+(X_1, X_2)$ itself in the right-hand side. This is done in different ways, depending on the size of the subsets involved; the "very small" and "very large" subsets can be handled with rather easy bounds, and we can concentrate on the "medium" range. Precisely, we have the following lemma:

Lemma 6.3.4. (1) *For any finite group G, and any subsets A, $B \subset G$, we have*

$$E(A, B) \leqslant \min(|A|^2|B|, |A||B|^2). \tag{6.4}$$

(2) *With notation as above, for all i and j, we have*

$$2^{-2(i+j)} E(A_{1,i}, A_{2,j}) \leqslant 2^4 \mathrm{rp}^+(X_1, X_2) e(A_{1,i}, A_{2,j}),$$

and for all $\alpha \geqslant 1$, we have

$$2^{-2(i+j)} E(A_{1,i}, A_{2,j}) \leqslant \alpha^{-1} \mathrm{rp}^+(X_1, X_2),$$

unless

$$\frac{|A_{1,i}|}{2^i} \geqslant \frac{1}{2\sqrt{\alpha}}, \qquad \frac{|A_{2,j}|}{2^j} \geqslant \frac{1}{2\sqrt{\alpha}}. \tag{6.5}$$

(3) *If $|A_{1,i}| \geqslant \alpha^{-1}|G|$ for some $\alpha \geqslant 1$, then*

$$2^{-2(i+j)} E(A_{1,i}, A_{2,j}) \leqslant 2^4 \alpha |G|^{-1}.$$

Proof. (1) follows from the definition

$$E(A,B) = |\{(x_1,x_2,y_1,y_2) \in A^2 \times B^2 \mid x_1y_1 = x_2y_2\}|,$$

since (x_1,y_1,x_2,y_2) is determined uniquely by (x_1,x_2,y_1), or by (x_2,y_1,y_2); the former means that $E(A,B) \leqslant |A|^2|B|$, and the second gives $E(A,B) \leqslant |A||B|^2$.

(2) This is in some sense the crucial point of the whole argument, and yet it is surprisingly simple. We remark that

$$\begin{aligned}\mathrm{rp}^+(X_1,X_2) &= \max(\mathrm{rp}(X_1),\mathrm{rp}(X_2)) \\ &\geqslant \frac{1}{2}\Big(\sum_{g\in G}(\mathbf{P}(X_1=g)^2+\mathbf{P}(X_2=g)^2)\Big) \\ &\geqslant \frac{1}{2}\Big(\frac{|A_{1,i}|}{2^{2+2i}}+\frac{|A_{2,j}|}{2^{2+2j}}\Big) \geqslant \frac{1}{4}\frac{(|A_i||A_j|)^{1/2}}{2^{i+j}}.\end{aligned}$$

for any choice of i and j. Hence we get

$$\begin{aligned}2^{-2(i+j)}E(A_{1,i},A_{2,j}) &= 2^{-2(i+j)}e(A_{1,i},A_{2,j})(|A_{1,i}||A_{2,j}|)^{3/2} \\ &\leqslant 4\mathrm{rp}^+(X_1,X_2)e(A_{1,i},A_{2,j})\frac{|A_{1,i}||A_{2,j}|}{2^{i+j}} \\ &\leqslant 16\mathrm{rp}^+(X_1,X_2)e(A_{1,i},A_{2,j})\end{aligned}$$

by (6.3), which was the first goal.

If instead we assume that $2^{-2(i+j)}E(A_{1,i},A_{2,j}) > \alpha^{-1}\mathrm{rp}^+(X_1,X_2)$, then we write simply

$$2^{-2(i+j)}|A_{1,i}|^2|A_{2,j}| \geqslant 2^{-2(i+j)}E(A_{1,i},A_{2,j}) \geqslant \alpha^{-1}\mathrm{rp}(X_2) \geqslant \alpha^{-1}\frac{|A_{2,j}|}{2^{2+2j}},$$

and get the first inequality of (6.5), the second being obtained symmetrically.

(3) This is also elementary: using (6.5) twice, we have

$$2^{-2(i+j)}E(A_{1,i},A_{2,j}) \leqslant \frac{|A_{1,i}||A_{2,j}|^2}{2^{2i+2j}} \leqslant 2^{-i+3} \leqslant 2^4|A_{1,i}|^{-1}$$

and hence the assumption leads directly to

$$2^{-2(i+j)}E(A_{1,i},A_{2,j}) \leqslant 2^4\alpha|G|^{-1}.$$

□

We refer to Lemma A.1.3 in Appendix A for some other elementary properties of the multiplicative energy.

We will now fix some parameters α, $\beta \geqslant 1$, and define

$$Q_\alpha = \{(i,j) \mid 0 \leqslant i,j < I, \quad |A_{1,i}| < 2^{i-1}\alpha^{-1} \text{ or } |A_{2,j}| < 2^{j-1}\alpha^{-1}\}, \tag{6.6}$$

$$\widetilde{Q}_\beta = \{(i,j) \mid 0 \leqslant i < I, \quad |A_{1,i}| \geqslant \beta^{-1}|G|\}, \tag{6.7}$$

and denote by P or $P_{\alpha,\beta}$ the complement of the union of these two sets. Thus P corresponds intuitively (for suitable values of the parameters) to those (i,j) for which trivial estimates are not enough to obtain a useful bound on $\mathrm{rp}(X_1X_2)$.

For $(i,j) \in Q_\alpha$, the second part of the lemma gives us

$$2^{-2(i+j)}E(A_{1,i},A_{2,j}) \leqslant \alpha^{-2}\mathrm{rp}^+(X_1,X_2),$$

and for $(i,j) \in \widetilde{Q}_\beta$, the third part gives

$$2^{-2(i+j)}E(A_{1,i},A_{2,j}) \leqslant 2^4\beta|G|^{-1},$$

and thus we have now shown that

$$\mathrm{rp}(X_1X_2) \leqslant 2^{3-2I}|G|^3 + 2^4\beta I^4|G|^{-1} + 2\alpha^{-2}\mathrm{rp}(X)I^4 \\ + 2^5\mathrm{rp}^+(X_1,X_2)I^2 \sum_{(i,j)\in P} e(A_{1,i},A_{2,j}),$$

where we just estimated the size of Q_α and $\widetilde{Q}_\beta$ by I^2. We now select

$$I = \left\lceil \frac{2\log 2|G|}{\log 2} \right\rceil \leqslant 3\log(3|G|),$$

and hence obtain a first basic inequality:

Theorem 6.3.5 (Towards L^2-flattening). *Let G be a finite group, X_1 and X_2 two symmetric independent G-valued random variables. With notation as above, for any α, $\beta \geqslant 1$, we have*

$$\mathrm{rp}(X_1X_2) \ll \frac{(\log 3|G|)^4\beta}{|G|} + \mathrm{rp}^+(X_1,X_2)(\log 3|G|)^4\Big\{\frac{1}{\alpha^2} + \sum_{(i,j)\in P} e(A_{1,i},A_{2,j})\Big\}, \tag{6.8}$$

where the implied constant is absolute.

Intuitively, β will be a small power of $|G|$, and the first term here is then close to $|G|^{-1}$. It is then essentially optimal, and can neither be improved or removed.

So if we want to understand how to use this inequality to obtain an improvement in the return probability for X_1X_2 in terms of $\mathrm{rp}(X_1)$ and $\mathrm{rp}(X_2)$, with the parameters α and β at our disposal, we have to show that $e(A_{1,i},A_{2,j})$ is rather small when $(i,j)\in P$.

The next lemma explains quite clearly what is at stake. It is not needed for the actual Bourgain-Gamburd argument (we will need stronger tools), but is certainly instructive in a first reading.

Lemma 6.3.6 (Extreme of the normalized energy). *Let G be a finite group and A, $B\subset G$ non-empty subsets. We have $e(A,B)\leqslant 1$, with equality if and only if there exists a subgroup $H\subset G$, and elements x, $y\in G$ such that*

$$A = xH, \qquad B = Hy.$$

Moreover, if $e(A,B)\geqslant \alpha^{-1}$ with $\alpha\geqslant 1$, we have

$$\alpha^{-2}|A| \leqslant |B| \leqslant \alpha^2|A|. \tag{6.9}$$

Proof. Let $a=|A|$ and $b=|B|$ for simplicity. We already know that

$$E(A,B)\leqslant \min(a^2b,ab^2),$$

(by (6.4)), and to deduce $e(A,B)\leqslant 1$, we observe that

$$\min(a^2b,ab^2)\leqslant (ab)^{3/2}, \tag{6.10}$$

if need be by considering the cases $a\leqslant b$ and $b\leqslant a$ separately (for example, in the first case, we have

$$\min(a^2b,ab^2) = a^2b = a^{3/2}a^{1/2}b \leqslant (ab)^{3/2}$$

and the other case is symmetric.)

We now prove (6.9). If $e(A,B)\geqslant\alpha^{-1}$, with $\alpha\geqslant 1$, and if $b\leqslant a$, we deduce

$$\alpha^{-1}(ab)^{3/2}\leqslant E(A,B)\leqslant \min(ab^2,a^2b) = ab^2,$$

so that

$$\alpha^{-2}a \leqslant b \leqslant a \leqslant \alpha^2 a,$$

which is (6.9) in that case. Of course, assuming $b \leqslant a$ leads to the same result.

We now attempt to characterize the sets with $e(A, B) = 1$. One direction is clear: if $H \subset G$ is a subgroup and $A = xH$, $B = Hy$, we have

$$\{(x_1, x_2, y_1, y_2) \in A^2 \times B^2 \mid x_1y_1 = x_2y_2\} = \{(h_1, h_2, h_3, h_4) \in H^4 \mid xh_1h_3y = xh_2h_4y\}$$

which contains $|H|^3 = (|A||B|)^{3/2}$ elements.

For the converse, we note first that (6.9) with $\alpha = 1$ shows that $a = b$ if $e(A, B) = 1$. Then we define what will turn out to be the necessary subgroup H, namely

$$H = \{g \in G \mid Ag = A\}$$

(which is indeed a subgroup of G). Fixing a single element $x_1 \in A$, we get $x_1h \in A$ for all $h \in H$, i.e., $x_1H \subset A$, and in particular $|H| \leqslant a$. We will now prove that $b \leqslant |H|$: since $|E(A, B)| = ab^2$, we see that for all x_1 in A and y_1, $y_2 \in B$, the element

$$x_1y_1y_2^{-1}$$

must also be in A. Since x_1 is arbitrary, this means that $y_1y_2^{-1} \in H$, hence that $y_1 \in Hy_2$. Taking y_2 to be fixed and varying y_1, we obtain $B \subset Hy_2$.

This gives therefore

$$|H| \leqslant a = b \leqslant |Hy_2| = |H|,$$

so there must be equality $|H| = a = b$, and then there must also be equalities in the inclusions we used, i.e.,

$$x_1H = A, \qquad Hy_2 = B,$$

which was our desired conclusion. □

Comparing this statement with the inequality (6.8), we can see that each term $e(A_{1,i}, A_{2,j})$ in the sum on the right-hand side over i and j is at most 1. We can easily understand when one of them is equal to 1, by using the lemma, and the reader may also want to first solve the next exercise.

Exercise 6.3.7 (Baby case). Show that, with notation as in (6.8), there exists an (explicit) absolute constant $\delta > 0$ such that, if there exists $(i, j) \in P$ with $e(A_{1,i}, A_{j,2}) = 1$, there also exists a subgroup $H \subset G$ and $x \in G$ for which

$$\mathbf{P}(X \in xH) \geqslant 4\alpha^{-1}.$$

Show that H is a proper subgroup of G unless

$$\mathrm{rp}(X) \leqslant 4|G|^{-1}.$$

If we assume that α is a small power of $|G|$, this means that if there is a term in (6.8) with $e(A_{1,i}, A_{j,2}) = 1$, and if X is not very uniformly distributed, the random variable X has a rather large probability of being in a proper subgroup.

However, simply knowing that each term in (6.8) is less than 1 is not particularly useful,[1] and we want a significantly better estimate on $e(A_{1,i}, A_{2,j})$. Precisely, we are looking for a structural understanding of pairs of sets A, $B \subset G$ such that $e(A, B) \geqslant \alpha^{-1}$ with α as large as possible (as a function of $|G|$). The ideal goal is that one should be able to describe such sets in terms similar to the characterization of the condition $e(A, B) = 1$, i.e., in terms of cosets of a common subgroup.

[1] Because it is not difficult to check that $\mathrm{rp}(X_1X_2) \leqslant \mathrm{rp}(X)$ anyway.

One can indeed do this for many groups, such as $\mathrm{SL}_2(\mathbf{F}_p)$. More precisely, the argument involves two steps. In the first one, which still applies to all finite groups G, one shows how to control sets with $e(A, B)$ rather large in terms of certain subsets $\mathsf{H} \subset G$ which are called *approximate subgroups*, and which play the role of the subgroup H in the case $e(A, B) = 1$.

In the second step, which is much more involved, we must classify approximate subgroups for certain families of finite groups G. For the groups $\mathrm{SL}_2(\mathbf{F}_p)$, such a classification is equivalent to Helfgott's theorem, which informally will show that all approximate subgroups of $\mathrm{SL}_2(\mathbf{F}_p)$ are essentially controlled by actual subgroups. Such a classification has now been established in much greater generality (indeed, in some sense, in full generality, by work of Breuillard, Green and Tao [17], although that general result is less precise than Helfgott's Theorem in the specific context of $\mathrm{SL}_2(\mathbf{F}_p)$).

We start with defining approximate subgroups, following Tao (see [108, Def. 3.8]).

Definition 6.3.8 (Approximate subgroup). Let G be a finite group and $\alpha \geqslant 1$. A non-empty subset $\mathsf{H} \subset G$ is an α-approximate subgroup if $1 \in \mathsf{H}$, $\mathsf{H} = \mathsf{H}^{-1}$ and there exists a symmetric subset $X \subset G$ of order at most α such that

$$\mathsf{H} \cdot \mathsf{H} \subset X \cdot \mathsf{H}, \tag{6.11}$$

which implies also $\mathsf{H} \cdot \mathsf{H} \subset \mathsf{H} \cdot X$. The *tripling constant* of H is defined by

$$\mathrm{trp}(\mathsf{H}) = \frac{|\mathsf{H} \cdot \mathsf{H} \cdot \mathsf{H}|}{|\mathsf{H}|}.$$

Remark 6.3.9. (1) Early works concerning approximate subgroups sometimes include further conditions. For instance, in [108, Def. 3.8], it is also asked that $X \subset H \cdot H$, and in [110], it is further necessary that $X \cdot H \subset H \cdot X \cdot X$. These two conditions are now thought to be extraneous. In any case, they do not play any role in the proof of the next result.

(2) By (6.11) we have an immediate bound for the tripling constant:

$$\mathsf{H} \cdot \mathsf{H} \cdot \mathsf{H} \subset (X \cdot \mathsf{H}) \cdot \mathsf{H} \subset (X \cdot X) \cdot \mathsf{H}$$

leads to

$$\mathrm{trp}(\mathsf{H}) \leqslant |X|^2 \leqslant \alpha^2.$$

However, if one is concerned with explicit upper bounds, one may well know a better bound than this, as in the next theorem.

We now state the generalization of Lemma 6.3.6 where the condition $e(A, B) = 1$ is relaxed.

Theorem 6.3.10 (Sets with large multiplicative energy). *Let G be a finite group and $\alpha \geqslant 1$. If A and B are subsets of G such that $e(A, B) \geqslant \alpha^{-1}$, there exist constants β_1, β_2, $\beta_3 \geqslant 1$, such that $\beta_i \leqslant c_1 \alpha^{c_2}$ for some constants c_1, $c_2 > 0$, and there exist a β_1-approximate subgroup $\mathsf{H} \subset G$ and elements x, $y \in G$ with*

$$|\mathsf{H}| \leqslant \beta_2 |A| \leqslant \beta_2 \alpha^2 |B|,$$

$$|A \cap x\mathsf{H}| \geqslant \frac{1}{\beta_3}|A|, \qquad |B \cap \mathsf{H}y| \geqslant \frac{1}{\beta_3}|B|,$$

$$\mathrm{trp}(\mathsf{H}) \leqslant \beta_4.$$

This is proved by Tao in [108, Th. 5.4, (i) implies (iv)] and quoted in [110, Th. 2.48]. We give a proof in Appendix A (see A.3.7), with explicit values of the constants (see also [70, Th. 2.1, Appendix A]).

Exercise 6.3.11. (1) Show that a 1-approximate subgroup of G is just a subgroup.

(2) Let $G = \mathbf{Z}/m\mathbf{Z}$ where $m \geqslant 1$ is an integer, and let $\mathsf{H} \subset G$ be the reduction modulo m of the integers

$$-k, -k+1, \ldots, -1, 0, 1, \ldots, k-1, k$$

for some $k < m/2$. Show that H is a 2-approximate subgroup of G. (The point of this example is that if k is very small compared with m, then H is not "close" to an ordinary subgroup.)

(3) [Helfgott] Let p be a prime number, $k < p/2$ an integer and fix two elements r, $s \in \mathbf{F}_p^{\times}$. Let

$$\mathsf{H} = \left\{ \begin{pmatrix} r^n & x & y \\ 0 & s^n & z \\ 0 & 0 & (rs)^{-n} \end{pmatrix} \mid x, y, z \in \mathbf{F}_p,\ -k < n < k \right\} \subset \mathrm{SL}_3(\mathbf{F}_p).$$

Show that H is an α-approximate subgroup of $\mathrm{SL}_3(\mathbf{F}_p)$ for some α independent of p and k.

We can now combine Theorem 6.3.10 with the Bourgain-Gamburd inequality, while still remaining at a level of great generality. We define for this purpose a type of groups where approximate subgroups are under control (this is not a standard definition, but it will turn out to be convenient.)

Definition 6.3.12. For $\delta > 0$, a finite group G is *δ-flourishing* if any symmetric subset $H \subset G$, containing 1, which generates G and has tripling constant $\mathrm{trp}(H) < |H|^{\delta}$ satisfies

$$H \cdot H \cdot H = G.$$

One could use variants of this definition, but it will be convenient. The motivation is, historically, one of opportunity: the theorem of Helfgott that we already mentioned, and which will be proved in Section 6.6, can be stated as follows: there exists $\delta > 0$, an absolute constant, such that $\mathrm{SL}_2(\mathbf{F}_p)$ is δ-flourishing for all primes p. The following exercise gives some useful basic insight on the nature of this property.

Exercise 6.3.13. (1) Show that there exists *no* $\delta > 0$ such that $\mathbf{Z}/p\mathbf{Z}$ is δ-flourishing for all primes p.

(2) Show that if G is δ-flourishing, there exists some explicit $\delta_1 > 0$ (depending on δ) such that any symmetric generating subset H containing 1 of size $|H| \geqslant |G|^{1-\delta_1}$ satisfies $H \cdot H \cdot H = G$.

(3) Show that if G is δ-flourishing, there exists $A > 0$, depending only on δ, such that

$$\mathrm{diam}(\mathcal{C}(G, S)) \leqslant (\log |G|)^A$$

for any symmetric generating set S of G.

We now establish a general form of L^2-flattening.

Theorem 6.3.14 (L^2-flattening conditions). *Let G be a finite group which is δ-flourishing for some δ with $0 < \delta \leqslant 1$. Let X_1, X_2 be G-valued independent symmetric random variables. Let $0 < \gamma < 1$ be given, and assume that*

$$\mathbf{P}(X_1 \in xH) \leqslant |G|^{-\gamma} \tag{6.12}$$

for all proper subgroups $H \subset G$ and $x \in G$.

Then for any $\varepsilon > 0$, there exists $\delta_1 > 0$, depending only on ε, δ and γ, such that

$$\mathrm{rp}(X_1X_2) \ll \frac{1}{|G|^{1-\varepsilon}} + \frac{\mathrm{rp}^+(X_1,X_2)}{|G|^{\delta_1}}$$

where the implied constant depends only on $(\varepsilon, \delta, \gamma)$.

Remark 6.3.15. If X_1 and X_2 are identically distributed, we obtain the case which was considered by Bourgain and Gamburd.

Proof. Fix some $\varepsilon > 0$. We apply (6.8) with $\alpha = |G|^{\delta_1-\varepsilon}$, for some $\delta_1 > 0$ to be chosen later, keeping β free for the moment. We obtain

$$\mathrm{rp}(X_1X_2) \ll \Bigl\{|G|^{-1+\varepsilon} + \mathrm{rp}^+(X_1,X_2)|G|^{-2\delta_1} + \mathrm{rp}^+(X_1,X_2)\sum_{(i,j)\in P} e(A_{1,i},A_{2,j})\Bigr\}$$

where the implied constant depends only on ε.

Let then

$$R = R_\alpha = \{(i,j) \in P \mid e(A_{1,i},A_{2,j}) \geqslant \alpha^{-1}\} \subset P,$$

so that the contribution of those $(i,j) \in P$ which are not in R_α, together with the middle term, can be bounded by

$$\ll |G|^{-\delta_1+\varepsilon}\mathrm{rp}^+(X_1,X_2),$$

where the implied constant depends only on ε.

This is of the right shape. We will now analyze the set R_α and show that it is empty when δ_1 is chosen small enough, and β is well-chosen. By Theorem 6.3.10, for each $(i,j) \in R$, there exists a β_1-approximate subgroup $\mathsf{H}_{i,j}$ and elements $(x_i,y_j) \in A_{1,i} \times A_{2,j}$ such that

$$|\mathsf{H}_{i,j}| \leqslant \beta_2|A_{1,i}|, \qquad |A_{1,i} \cap x_i\mathsf{H}_{i,j}| \geqslant \beta_3^{-1}|A_{1,i}|, \qquad |A_{2,j} \cap \mathsf{H}_{i,j}y_j| \geqslant \beta_3^{-1}|A_{2,j}|,$$

and with tripling constant bounded by β_4, where

$$\beta_i \leqslant c_1|G|^{c_2\delta_1}$$

for some absolute constants c_1, $c_2 > 0$. We then note first that if $H_{i,j}$ denotes the "ordinary" subgroup generated by $\mathsf{H}_{i,j}$, we have

$$\begin{aligned}\mathbf{P}(X \in x_iH_{i,j}) &\geqslant \mathbf{P}(X \in x_i\mathsf{H}_{i,j})\\ &\geqslant \mathbf{P}(X \in A_{1,i} \cap x_i\mathsf{H}_{i,j}) \geqslant \frac{1}{\beta_3}\frac{|A_{1,i}|}{2^{i+1}} \geqslant \frac{1}{4\beta_3\alpha} \gg \frac{1}{|G|^{(1+c_2)\delta_1}},\end{aligned}$$

with an absolute implied constant, using the fact that elements of P are not in the set (6.6). If δ_1 is small enough that

$$(1+c_2)\delta_1 < \gamma, \tag{6.13}$$

and if $|G|$ is large enough, this is not compatible with (6.12), and hence, for such a choice of δ_1, we deduce that each $\mathsf{H}_{i,j}$ (if any!) generates the group G.

We next observe that $\mathsf{H}_{i,j}$ can not be extremely small, which will allow us to relate the tripling constant to the size of $\mathsf{H}_{i,j}$ instead of that of G. Indeed, we have

$$|\mathsf{H}_{i,j}| \geqslant |x_i\mathsf{H}_{i,j} \cap A_{1,i}| \geqslant \beta_3^{-1}|A_{1,i}|,$$

on the one hand, and by applying (6.12) with $H = 1$, we can see that $A_{1,i}$ is not too small, namely

$$|A_{1,i}| \geqslant \frac{\mathbf{P}(X \in A_{1,i})}{\max_{g\in G}\mathbf{P}(X = g)} \geqslant |G|^\gamma\,\mathbf{P}(X \in A_{1,i}) \geqslant \frac{|G|^\gamma|A_{1,i}|}{2^{i+1}} \geqslant \frac{|G|^\gamma}{4\alpha}$$

using again the definition of P.

This gives the lower bound

$$|\mathsf{H}_{i,j}| \geqslant \frac{|G|^{\gamma}}{4\alpha\beta_3} \gg |G|^{\gamma_1}$$

with $\gamma_1 = \gamma - \delta_1(1+c_2)$ (which is > 0 by (6.13)), for some absolute implied constant. This leads to control of the tripling constant, namely

$$\mathrm{trp}(\mathsf{H}_{i,j}) \leqslant \beta_4 \leqslant c_1|G|^{c_2\delta_1} \ll |\mathsf{H}_{i,j}|^{c_2\delta_1\gamma_1^{-1}}$$

where the implied constant depends on γ and δ_1.

Since we assumed that G is δ-flourishing, we see from Definition 6.3.12 that if δ_1 is such that

$$\frac{c_2\delta_1}{\gamma_1} = \frac{c_2\delta_1}{\gamma - (1+c_2)\delta_1} < \delta, \tag{6.14}$$

and again if $|G|$ is large enough, the approximate subgroup $\mathsf{H}_{i,j}$ must in fact be very large, specifically it must satisfy

$$\mathsf{H}_{i,j} \cdot \mathsf{H}_{i,j} \cdot \mathsf{H}_{i,j} = G,$$

and in particular

$$|\mathsf{H}_{i,j}| \geqslant \frac{|\mathsf{H}_{i,j} \cdot \mathsf{H}_{i,j} \cdot \mathsf{H}_{i,j}|}{\beta_4} = \frac{|G|}{\beta_4}$$

Then we get

$$|A_{1,i}| \geqslant |A_{1,i} \cap x_i\mathsf{H}_{i,j}| \geqslant \frac{|\mathsf{H}_{i,j}|}{\beta_2} \geqslant \frac{|G|}{\beta_2\beta_4} \gg |G|^{1-2c_2\delta_1}$$

where the implied constant is absolute. If we now select $\beta = M|G|^{2c_2\delta_1}$ for M large enough, this implies that $(i,j) \in \widetilde{Q}_{\beta}$. Since P is also in the complement of (6.7), this means that R is empty for these parameters.

We now conclude that for any $\varepsilon > 0$ and any $\delta_1 > 0$ small enough so that (6.13) and (6.14) are satisfied, we have

$$\mathrm{rp}(X_1X_2) \ll |G|^{-1+2c_2\delta_1+\varepsilon} + |G|^{-\delta_1}\mathrm{rp}^{+}(X_1,X_2),$$

where the implied constant depends on ε, δ_1 and γ. Fixing δ_1 small enough, we make the first exponent as close to -1 as possible; then the second is of the form $-\delta_2$, where $\delta_2 > 0$. Thus, renaming the constants, we obtain the conclusion as stated. □

In order to apply this theorem iteratively, we need also the following simple observation of "increase of uniformity".

Lemma 6.3.16 (Uniformity can only increase). *Let G be a finite group, S a symmetric generating set, and let (X_n) be the corresponding random walk on $\mathcal{C}(G,S)$. For any $n \geqslant 1$ and $m \geqslant n$ we have $\mathrm{rp}(X_m) \leqslant \mathrm{rp}(X_n)$.*

Proof. By the spectral interpretation (6.1) of the return probability, we have

$$\mathrm{rp}(X_m) = \mathbf{P}(X_{2m} = 1) = \frac{1}{|G|}\mathrm{Tr}(M^{2m}), \qquad \mathrm{rp}(X_{2n}) = \mathbf{P}(X_{2n} = 1) = \frac{1}{|G|}\mathrm{Tr}(M^{2n})$$

were M is the Markov operator. Since all eigenvalues of M^2 are non-negative and $\leqslant 1$, it follows that

$$\mathrm{Tr}(M^{2m}) \leqslant \mathrm{Tr}(M^{2n})$$

for $m \geqslant n$, as desired. □

We can summarize the conclusion of all this section as follows, in the spirit of Proposition 6.2.8.

Corollary 6.3.17 (The Bourgain-Gamburd expansion criterion). *Let $\underline{c} = (c, d, \delta, \gamma)$ be a tuple of positive real numbers, and let $\mathcal{G}_2(\underline{c})$ be the family of all finite connected Cayley graphs $\mathcal{C}(G, S)$ for which the following conditions hold:*

(1) *We have $d(G) \geqslant |G|^d$;*

(2) *The group G is δ-flourishing;*

(3) *For the random walk (X_n) on G with $X_0 = 1$, we have that*

$$\mathbf{P}(X_{2k} \in xH) \leqslant |G|^{-\gamma}$$

for some $k \leqslant c \log |G|$ and all $x \in G$ and proper subgroups $H \subset G$.

Then, if $\underline{c}$ is such that $\mathcal{G}_2(\underline{c})$ contains graphs $\mathcal{C}(G, S)$ with arbitrarily large vertex sets G and bounded generating sets S, these form an expander family.

The idea in this corollary is that $c > 0$ will be rather small, and Condition (3) means that, after k steps, the random walk on $\mathcal{C}(G, S)$ has begun spreading out, and escaping from proper subgroups, at least to some extent. In view of Proposition 6.2.8, the content of this corollary is therefore that these two conditions imply that, after $\leqslant m \log |G|$ steps, for some large m, the random walk will become "almost" uniform, allowing us to apply Corollary 6.2.5.

Proof. Let $\Gamma = \mathcal{C}(G, S)$ be a graph in $\mathcal{G}_2(\underline{c})$. We apply Theorem 6.3.14 with $0 < \varepsilon < d$, say $\varepsilon = d/2$. Let δ_1 be such that the L^2-flattening inequality holds for this value, so that

$$\mathrm{rp}(Y_1 Y_2) \ll \max\Bigl(\frac{1}{|G|^{1-d/2}}, \frac{\mathrm{rp}^+(Y_1, Y_2)}{|G|^{\delta_1}}\Bigr)$$

for random variables Y_1, Y_2 which satisfy the assumptions of this theorem.

Let $k = \lfloor c \log |G| \rfloor$ be given by (3). We apply the theorem to $Y_1 = X_{2^j k}$ and $Y_2 = X_{2^{(j+1)}k} Y_1^{-1}$ for $j \geqslant 0$. These are indeed independent and symmetric random variables, and Conditions (2) and (3) imply that we can apply Theorem 6.3.14 to these random variables for any $j \geqslant 2$. Since Y_1 and Y_2 are identically distributed, we have

$$\mathrm{rp}^+(Y_1, Y_2) = \mathrm{rp}(Y_1) = \mathrm{rp}(X_{2^j k}).$$

Thus, applying the theorem, we obtain by induction

$$\mathrm{rp}(X_{2^j k}) \ll \mathrm{rp}(X_k) |G|^{-j\delta_1/2} \ll |G|^{-j\delta_1/2}$$

when j is such that

$$|G|^{1-d/2} > |G|^{j\delta_1/2},$$

and for larger j, we get

$$\mathrm{rp}(X_{2^j k}) \ll |G|^{-1+d/2},$$

where the implied constants depend only on (d, γ). In particular, we obtain this last inequality for

$$j \ll \frac{1}{\delta_1}$$

which, by the "cycle-counting" Corollary (6.2.5), gives

$$\varrho_\Gamma \leqslant (|G|^{1-d} \mathrm{rp}(X_{2^j k}))^{1/(2^j k)} \leqslant \exp(-cd)$$

for some constant $c > 0$ which is independent of $\Gamma \in \mathcal{G}_2(\underline{c})$. This proves the theorem. □

6.4. Implementing the Bourgain-Gamburd argument

Theorem 6.1.1 will now be proved by applying the criterion of Corollary 6.3.17. Thus we will consider the groups $G_p = \mathrm{SL}_2(\mathbf{F}_p)$ for p prime, for which Condition (1) of the Bourgain-Gamburd criterion (which is purely a group-theoretic property) is known: this is Theorem 6.2.6 of Frobenius, which gives the value $\delta = 1/3$. Condition (2) is a much more delicate matter: it is Helfgott's Theorem, which is proved in Section 6.6. However, it is still purely a property of the groups $\mathrm{SL}_2(\mathbf{F}_p)$.

Condition (3), on the other hand, involves the choice of generating sets. The symmetric generating sets S_p in Theorem 6.1.1 are assumed to be obtained by reduction modulo p of a fixed symmetric subset $S \subset \mathrm{SL}_2(\mathbf{Z})$. What makes this situation special, and in particular makes it possible to check Condition (3) of the criterion, is the following special case: if $S \subset \mathrm{SL}_2(\mathbf{Z})$ generates a *free group*, the first steps of the random walks modulo p (up to a small multiple of $\log |G_p|$) are "the same" as those of a random walk on an infinite regular tree. We can then "see" the probabilities $\mathbf{P}(X_k = g)$ that we need to estimate at the level of this tree, where they are easier to analyze.

Since we work with symmetric sets in groups, the following definition will be useful:

Definition 6.4.1. Let G be a group and S a symmetric generating set of G. We say that G is *freely generated* by S if S has even cardinality and S is the disjoint union $S = T \cup T^{-1}$, where T generates a free group. In particular, G is then a free group on $|T|$ generators.

We begin with a classical proposition, whose idea goes back to Margulis. For the statement, we will use the norm

$$\|g\| = \max_{v,w\neq 0} \frac{|\langle gv, w\rangle|}{\|v\|\|w\|}$$

of matrices, as recalled in Appendix C.

Proposition 6.4.2 (Large girth for finite Cayley graphs). *Let $S \subset \mathrm{SL}_2(\mathbf{Z})$ be a symmetric set, and let $\Gamma = \mathcal{C}(G, S)$ be the corresponding Cayley graph. Let $\tau > 0$ be defined by*

$$\tau^{-1} = \log \max_{s\in S} \|s\| > 0, \tag{6.15}$$

which depends only on S.

(1) *For all primes p and all $r < \tau \log(p/2)$, where $G_p = \mathrm{SL}_2(\mathbf{F}_p)$, the subgraph Γ_r induced by the ball of radius r in Γ maps injectively to $\mathcal{C}(G_p, S)$.*

(2) *If G is freely generated by S, in particular $1 \notin S$, the Cayley graph $\mathcal{C}(G_p, S)$ contains no cycle of length $< 2\tau \log(p/2)$, i.e., its* girth *is at least equal to $2\tau \log(p/2)$.*

Proof. The main point is that if all coordinates of two matrices g_1, $g_2 \in \mathrm{SL}_2(\mathbf{Z})$ are less than $p/2$ in absolute value, a congruence $g_1 \equiv g_2 \pmod p$ is equivalent to the equality $g_1 = g_2$. And because G is freely generated by S, knowing a matrix in G is equivalent to knowing its expression as a word in the generators in S.

Thus, let x be an element in the ball of radius r centered at the origin. By definition, x can be expressed as

$$x = s_1 \cdots s_m$$

with $m \leqslant r$ and $s_i \in S$. Using the elementary properties (C.1) and (C.2) of the norm, we get

$$\max_{i,j} |x_{i,j}| \leqslant \|x\| \leqslant \|s_1\| \cdots \|s_m\| \leqslant e^{m/\tau} \leqslant e^{r/\tau}.$$

Applying the beginning remark and this fact to two elements x and y in $\mathcal{B}_1(r)$, for r such that $e^{r/\tau} < \frac{p}{2}$, it follows that $x \equiv y \,(\mathrm{mod}\, p)$ implies $x = y$, which is (1).

Then (2) follows because any embedding of a cycle $\gamma : C_m \to \mathcal{C}(G_p, S)$ such that $\gamma(0) = 1$ and such that

$$d(1, \gamma(i)) \leqslant m/2 < \tau \log(p/2)$$

for all i can be lifted to the cycle (of the same length) with image in the Cayley graph of G with respect to S, and if S generates freely G, the latter graph is a tree. Thus a cycle of length $m = \mathrm{girth(\mathcal{C}(G_p, S))}$ must satisfy $m/2 \geqslant \tau \log(p/2)$. □

We can now check Condition (3) in the Bourgain-Gamburd criterion, first for cosets of the trivial subgroup, i.e., for the probability that X_n be a fixed element when n is of size $c \log p$ for some fixed (but small) $c > 0$.

Corollary 6.4.3 (Decay of probabilities). *Let $S \subset \mathrm{SL}_2(\mathbf{Z})$ be a symmetric set, G the subgroup generated by S. Assume that S freely generates G in the sense of Definition 6.4.1. Let p be a prime such that the reduction S_p of S modulo p generates $G_p = \mathrm{SL}_2(\mathbf{F}_p)$, and let (X_n) be the random walk on $\mathcal{C}(G_p, S_p)$ with $X_0 = 1$.*

There exists $\gamma > 0$, depending only on S, such that for any prime p large enough and any $x \in \mathrm{SL}_2(\mathbf{F}_p)$, we have

$$\mathbf{P}(X_n = x) \leqslant |G_p|^{-\gamma} \tag{6.16}$$

for some

$$n \asymp \tau \log(p/2),$$

where τ is defined in Proposition 6.4.2.

Proof. There exists $\widetilde{x} \in G$ such that $\widetilde{x}$ reduces to x modulo p and $\widetilde{x}$ is at the same distance to 1 as x, and by Proposition 6.4.2, (2), we have

$$\mathbf{P}(X_n = x) = \mathbf{P}(\widetilde{X}_n = \widetilde{x}),$$

for $n \leqslant \tau \log(p/2)$, where $(\widetilde{X}_n)$ is the random walk starting at 1 on the infinite $|S|$-regular tree $\mathcal{C}(G, S)$. By Proposition 3.2.31, we have

$$\mathbf{P}(\widetilde{X}_n = \widetilde{x}) \leqslant r^{-n} \quad \text{with} \quad r = \frac{|S|}{2\sqrt{|S| - 1}},$$

for all $n \geqslant 1$ and all $\widetilde{x} \in G$. For p large enough, and n of size (say) $\frac{1}{2}\tau \log p$,this implies

$$\mathbf{P}(X_n = x) \leqslant |G_p|^{-\gamma}$$

where $\gamma > 0$ depends only on S. □

In order to deal with cosets of other proper subgroups of $\mathrm{SL}_2(\mathbf{F}_p)$, we will exploit the fact that those subgroups are very well understood, and in particular, there is no proper subgroup that is "both big and complicated". Precisely, by Corollary B.2.3 in Appendix B, we see that if $p \geqslant 5$ and $H \subset \mathrm{SL}_2(\mathbf{F}_p)$ is a proper subgroup, one of the following two properties holds:

(1) The order of H is at most 120;

(2) For all $(x_1, x_2, x_3, x_4) \in H$, we have

$$[[x_1, x_2], [x_3, x_4]] = 1. \tag{6.17}$$

The first ones are "small": if H is of this type and (6.16) holds, we get

$$\mathbf{P}(X \in xH) \leqslant 120|G_p|^{-\gamma}$$

immediately. The second are, from the group-theoretic point of view, not very complicated (their commutator subgroups are abelian). We handle them (as was done in [12]) at the level of the free group generated by S, although it is also possible to keep attention focused to the finite groups $\mathrm{SL}_2(\mathbf{F}_p)$ (indeed, for groups which are more complicated than SL_2, there exist proper "complicated" subgroups, and this second option is then most natural.) Precisely, we have the following *ad-hoc* lemma:

Proposition 6.4.4. *Let $k \geqslant 2$ be an integer and let $W \subset F_k$ be a subset of the free group on k generators $(a_1, \ldots, a_k)$ such that*

$$[[x_1, x_2], [x_3, x_4]] = 1 \tag{6.18}$$

for all $(x_1, x_2, x_3, x_4) \in W$. Then for any $m \geqslant 1$, we have

$$|\{x \in W \mid d_T(1, x) \leqslant m\}| \leqslant (4m+1)(8m+1) \leqslant 45m^2,$$

where T is the $(2k)$-regular tree $\mathcal{C}(F_k, S)$ where $S = \{a_i^{\pm 1}\}$.

Proof. The basic fact we need is that the condition $[x, y] = 1$ is very restrictive in F_k: precisely, for a fixed $x \neq 1$, we have $[x, y] = 1$ if and only if y belongs to the centralizer $C_{F_k}(x)$ of x in F_k, which is an infinite cyclic group by Proposition B.1.1, (3). Let z be a generator of this cyclic group. We find

$$|\{y \in \mathcal{B}_1(m) \mid [x, y] = 1\}| = |\{h \in \mathbf{Z} \mid d_{T_k}(1, z^h) \leqslant m\}| \leqslant 2m + 1 \tag{6.19}$$

since (Proposition B.1.1, (5)), we have $d_T(1, z^h) \geqslant |h|$.

Now we come back to a set W verifying the assumption (6.18), which we assume to be not reduced to 1, and we denote $W_m = W \cap \mathcal{B}_1(m)$, the set we want to estimate. If $[x, y] = 1$ for all x, $y \in W_m$, taking a fixed $x \neq 1$ in W_m, we get $W_m \subset C_{F_k}(x) \cap \mathcal{B}_1(m)$, and (6.19) gives the result.

Otherwise, fix x_0 and y_0 in W_m such that $a = [x_0, y_0] \neq 1$. Then, for all y in W_m we have $[a, [x_0, y]] = 1$. Noting that $d_T(1, [x_0, y]) \leqslant 4m$, it follows again from the above that the number of possible values of $[x_0, y]$ is at most $8m + 1$ for $y \in W_m$.

Now for one such value $b = [x_0, y]$, we consider how many $y_1 \in W_m$ may satisfy $[x_0, y_1] = b$. We have $[x_0, y] = [x_0, y_1]$ if and only if $\varphi(yy_1^{-1}) = yy_1^{-1}$, where $\varphi(y) = x_0 y x_0^{-1}$ denotes the inner automorphism of conjugation by x_0. Hence y_1 satisfies $[x_0, y_1] = b$ if and only if $\varphi(yy_1^{-1}) = yy_1^{-1}$, which is equivalent to $yy_1^{-1} \in C_{F_k}(x_0)$. Taking a generator z of this centralizer again (note $x_0 \neq 1$), we get

$$\begin{aligned} |\{y_1 \in \mathcal{B}_1(m) \mid [x_0, y_1] = [x_0, y]\}| &= |\{h \in \mathbf{Z} \mid yz^h \in \mathcal{B}_1(m)\}| \\ &\leqslant |\{h \in \mathbf{Z} \mid z^h \in \mathcal{B}_1(2m)\}| \leqslant 4m + 1, \end{aligned}$$

since

$$d_T(1, z^h) = d_T(y, yz^h) \leqslant d_T(1, y) + d_T(1, yz^h) \leqslant 2m$$

for $h \in \mathbf{Z}$ such that yz^h is in $\mathcal{B}_1(m)$.

Hence we have $|W_m| \leqslant (4m+1)(8m+1)$ in that case, which proves the result. □

Using Corollary 6.4.3, we finally verify fully Condition (3) in Corollary 6.3.17:

Corollary 6.4.5 (Decay of probabilities, II). *Let $S \subset \mathrm{SL}_2(\mathbf{Z})$ be a symmetric set, G the subgroup generated by S. Assume that S freely generates G. Let p be a prime such that the reduction S_p of S modulo p generates $G_p = \mathrm{SL}_2(\mathbf{F}_p)$, and let (X_n) be the random walk on $\mathcal{C}(G_p, S_p)$ with $X_0 = 1$.*

There exist $c > 0$ and $\gamma > 0$ such that for any prime p large enough, any $x \in \mathrm{SL}_2(\mathbf{F}_p)$ and any proper subgroup $H \subset \mathrm{SL}_2(\mathbf{F}_p)$, we have

$$\mathbf{P}(X_n \in xH) \leqslant |G_p|^{-\gamma} \tag{6.20}$$

for some

$$n \leqslant c \log |G_p|.$$

Proof. We start by noting that, for all $x \in G_p$ and $H \subset G_p$, we have

$$\begin{aligned}\mathbf{P}(X_{2n} \in H) &= \sum_{g \in G_p} \mathbf{P}(X_n = g \text{ and } X_n^{-1} X_{2n} \in H)\\ &\geqslant \sum_{h \in H} \mathbf{P}(X_n = xh)\, \mathbf{P}(h^{-1}x^{-1}X_n \in H)\end{aligned}$$

because $X_n^{-1}X_{2n}$ is independent of, and has the same distribution as, X_n (Proposition 3.2.9). Since

$$\mathbf{P}(h^{-1}x^{-1}X_n \in H) = \mathbf{P}(X_n \in xH),$$

this gives

$$\mathbf{P}(X_{2n} \in H) \geqslant \mathbf{P}(X_n \in xH) \sum_{h \in H} \mathbf{P}(X_n = xh) = \mathbf{P}(X_n \in xH)^2,$$

which means that it is enough to give an upper bound for $\mathbf{P}(X_{2n} \in H)$ to get one for $\mathbf{P}(X_n \in xH)$.

Consider first the case where H is "big", but (6.17) holds for H. Let $\widetilde{H} \subset G$ be the pre-image of H under reduction modulo p. If $2n \leqslant \tau \log(p/2)$, then as in the proof of Corollary 6.4.3, we get

$$\mathbf{P}(X_{2n} \in H) = \mathbf{P}(\widetilde{X}_{2n} \in \widetilde{H}).$$

Provided n also satisfies the stronger condition $n \leqslant m = \frac{1}{16}\tau \log(p/2)$, any commutator

$$[[x_1, x_2], [x_3, x_4]]$$

with $x_i \in \widetilde{H} \cap \mathcal{B}_1(n)$ is an element at distance at most $\tau \log(p/2)$ from 1 in the tree $\mathcal{C}(G, S)$, which reduces to the identity modulo p by (6.17), and therefore must be itself equal to 1. In other words, we can apply Proposition 6.4.4 to $W = \widetilde{H} \cap \mathcal{B}_1(m)$ to deduce the upper bound

$$|\widetilde{H} \cap \mathcal{B}_1(m)| \leqslant 45m^2.$$

We now take

$$n = \frac{1}{32} \lfloor \tau \log(p/2) \rfloor,$$

and we derive

$$\mathbf{P}(X_{2n} \in H) \leqslant |\widetilde{H} \cap \mathcal{B}_1(m)| r^{-2n} \leqslant 45m^2 |G_p|^{-\gamma/16}$$

where γ is as in Corollary 6.4.3, and hence

$$\mathbf{P}(X_n \in xH) \leqslant |G_p|^{-\gamma/32}$$

provided p is large enough, which gives the conclusion in that case.

On the other hand, if H is "small", i.e., if $|H| \leqslant 120$, then for the same value of n we get

$$\mathbf{P}(X_n \in xH) \leqslant 120 |G_p|^{-\gamma}$$

by Corollary 6.4.3, and this gives again the desired result for p large enough. □

We can now summarize what we have obtained concerning the Bourgain-Gamburd criterion (Corollary 6.3.17) in the situation of Corollary 6.4.5 for $G_p = \mathrm{SL}_2(\mathbf{F}_p)$. This will finally prove Theorem 6.1.1.

(1) We have

$$d(G_p) = \frac{p-1}{2}$$

for $p \geqslant 3$. In particular, $d(G_p) \geqslant |G_p|^d$ for any $d < 1/3$ provided p is large enough.

(2) The next section will show that these groups are δ-flourishing for some $\delta > 0$ independent of p.

(3) For the random walk (X_n) on G_p with $X_0 = 1$ (associated to the generating set S_p), we have

$$\mathbf{P}(X_{2k} \in xH) \leqslant |G|^{-\gamma}$$

when

$$k \leqslant c \log |G_p|$$

for some $c > 0$, $\gamma > 0$ and p large enough.

We conclude that, if $S \subset \mathrm{SL}_2(\mathbf{Z})$ freely generates a free subgroup of rank $\geqslant 2$ and γ is defined as above, then for $\underline{c}$ defined by

$$\underline{c} = (c, d, \delta, \gamma)$$

for any $d < 1/3$, the family $\mathcal{G}_2(\underline{c})$ contains all but finitely many Cayley graphs of $\mathrm{SL}_2(\mathbf{F}_p)$ with respect to S modulo p. In particular, we then deduce from Corollary 6.3.17 that these families are expander families, and we can even write down an explicit value for the spectral gap (for p large enough). This proves Theorem 6.1.1 for these sets S.

We can now finally explain the classical tool which is used to reduce the full statement of Theorem 6.1.1 to this first case.

Lemma 6.4.6. *Let $S \subset \mathrm{SL}_2(\mathbf{Z})$ be any finite symmetric subset and let G be the subgroup generated by S. If the reduction of S modulo p generates $\mathrm{SL}_2(\mathbf{F}_p)$ for all primes p large enough, then there exists a symmetric generating set $S_1 = \{s_1^{\pm 1}, s_2^{\pm 1}\} \subset G$ which freely generates a free subgroup $G_1 \subset G$ of rank 2. Moreover, for p large enough, S_1 modulo also generates $\mathrm{SL}_2(\mathbf{F}_p)$.*

Proof. This is a special case of a very general fact (often referred to as the "Tits alternative") about subgroups of linear groups. There is however a very simple argument in this case. We consider

$$\widetilde{G} = G \cap \Gamma(2),$$

where

$$\Gamma(2) = \ker(\mathrm{SL}_2(\mathbf{Z}) \longrightarrow \mathrm{SL}_2(\mathbf{Z}/2\mathbf{Z})).$$

By Proposition B.1.3, (3), the subgroup $\Gamma(2)$ of $\mathrm{SL}_2(\mathbf{Z})$ is a free group of rank 2, hence the intersection $\widetilde{G}$ is a free group (as a subgroup of the free group $\Gamma(2)$, see Proposition B.1.1 (1)). It cannot be of rank 1 (or 0) because it is of finite index in G, and hence (by the assumption on G) still surjects to $\mathrm{SL}_2(\mathbf{F}_p)$ modulo all primes which are large enough. Even if it were not of rank 2, one can take two arbitrary elements in a free generating set of $\widetilde{G}$, and the subgroup G_1 they generate. □

Using this, for a given $S \subset \mathrm{SL}_2(\mathbf{Z})$, we construct the free subgroup G_1 generated by $S_1 = \{s_1^{\pm 1}, s_2^{\pm 1}\}$. We are simply going to compare the expansion for the Cayley graphs of $\mathrm{SL}_2(\mathbf{F}_p)$ with respect to S and to S_1.

For p large enough so that $G_p = \mathrm{SL}_2(\mathbf{F}_p)$ is generated both by S modulo p and S_1 modulo p, we have

$$d(x,y) \leqslant C d_1(x,y)$$

where $d_1(\cdot,\cdot)$ is the distance in the Cayley graph $\Gamma_1 = \mathcal{C}(G_p, S_1)$, and $d(\cdot,\cdot)$ the distance in $\Gamma_2 = \mathcal{C}(G_p, S)$ and C is the maximum of the word length of s_1, s_2 with respect to S. Hence, by Lemma 3.1.17 (applied to Γ_1 and Γ_2 with f the identity), the expansion constants satisfy

$$h(\mathcal{C}(G_p, S)) = h(\Gamma_2) \geqslant w^{-1} h(\mathcal{C}(G_p, S_1))$$

with

$$w = 4 \sum_{j=1}^{\lfloor C \rfloor} |S|^{j-1}.$$

In particular, if Theorem 6.1.1 holds for G_1, it will also hold for G.

Example 6.4.7 (The Lubotzky group). The group L generated by

$$S = \left\{ \begin{pmatrix} 1 & \pm 3 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ \pm 3 & 1 \end{pmatrix} \right\} \subset \mathrm{SL}_2(\mathbf{Z})$$

is a free group on 2 generators, and is of infinite index in $\mathrm{SL}_2(\mathbf{Z})$ (see Proposition B.1.3 in Appendix B). For all $p \neq 3$, the reduction of S modulo p generates $\mathrm{SL}_2(\mathbf{F}_p)$, by Proposition B.2.1, (2).

6.5. Quasi-random groups

Our next goal is the prove that the groups $\mathrm{SL}_2(\mathbf{F}_p)$ satisfy Condition (2) of the Bourgain-Gamburd criterion: they are δ-flourishing for some $\delta > 0$ independent of p. The first step, however, will be to prove a weaker property (which would also follow from δ-flourishing) which turns out to be more accessible and useful to simplify the final arguments. This result is due to Gowers [46] and Nikolov-Pyber [90]:

Theorem 6.5.1 (Gowers; Nikolov-Pyber). *Let $p \geqslant 3$ be prime, and let $H \subset \mathrm{SL}_2(\mathbf{F}_p)$ be an arbitrary subset such that*

$$|H| \geqslant 2|\,\mathrm{SL}_2(\mathbf{F}_p)|^{8/9}.$$

Then we have $H \cdot H \cdot H = \mathrm{SL}_2(\mathbf{F}_p)$.

This should be compared with Exercise 6.3.13, (2), which shows that *if* $\mathrm{SL}_2(\mathbf{F}_p)$ is to be δ-flourishing, such a property must hold (possibly with an exponent closer to 1 than 8/9.)

In fact, a similar property can be stated for all finite groups, although it is only of special interest when the invariant $d(G)$ is relatively large.

Theorem 6.5.2. *Let G be a finite group, and let A, B, $C \subset G$ be any subsets of G such that*

$$\frac{|A||B||C|}{|G|^3} \geqslant \frac{1}{d(G)}.$$

Then we have

$$A \cdot B \cdot C = G.$$

In particular, if $|A| \geqslant |G|/d(G)^{1/3}$, then we have $A \cdot A \cdot A = G$.

Applying the theorem of Frobenius (Theorem 6.2.6), we see that Theorem 6.5.1 is a corollary of this general fact.

In some sense, Theorem 6.5.2 can be seen as a "triple product" version of the following well-known fact: if A is a subset of G such that $|A| > |G|/2$, then $A \cdot A = G$. The latter is proved by a nice trick: if $x \in G$ is any element, we have

$$|A| + |xA^{-1}| = 2|A| > |G|,$$

and hence the sets A and xA^{-1} can not be disjoint. Picking $a \in A$ such that $a \in xA^{-1}$, we can write $a = xb^{-1}$ with $b \in A$, and therefore $x = ab \in A \cdot A$.

In fact, the proof of Theorem 6.5.2 will begin by proving a weaker-looking result (this is the idea of "quasirandom groups" of Gowers), and then using a similarly clever trick of Nikolov-Pyber, will use this to conclude.

Proposition 6.5.3 (Gowers). *Let G be a finite group, and let A, B, $C \subset G$ be any subsets of G such that*

$$\frac{|A||B||C|}{|G|^3} \geqslant \frac{1}{d(G)}.$$

Then we have

$$(A \cdot B) \cap C \neq \emptyset.$$

Let us see quickly how to prove the theorem from this. Given A, B and C as in the statement and $x \in G$, we consider the subsets

$$A_1 = B^{-1}, \quad B_1 = A^{-1}x, \quad C_1 = C.$$

We have

$$\frac{|A_1||B_1||C_1|}{|G|^3} = \frac{|A||B||C|}{|G|^3} \geqslant \frac{1}{d(G)},$$

and hence $A_1 \cdot B_1 \cap C_1 = B^{-1} \cdot (A^{-1}x) \cap C \neq \emptyset$. Thus there exists $(a, b, c) \in A \times B \times C$ with

$$c = b^{-1}a^{-1}x,$$

or or $x = abc \in A \cdot B \cdot C$.

Proof of Proposition 6.5.3. The idea is to consider the function on G defined by

$$\varphi(g) = \frac{|C \cap gB|}{|G|} = \mu(C \cap gB),$$

(where μ is the uniform probability measure on G) and to show that it is non-zero on a "large" set by estimating its variance

$$V = \frac{1}{|G|} \sum_{g \in G} \Bigl(\varphi(g) - \langle \varphi, 1 \rangle\Bigr)^2.$$

Indeed, by positivity, the set $X = \{x \mid \varphi(x) = 0\}$ satisfies

$$\frac{|X|}{|G|}(\langle \varphi, 1 \rangle)^2 \leqslant V,$$

hence any set A which is large enough in the sense that

$$\mu(A) > \frac{V}{(\langle \varphi, 1 \rangle)^2} \tag{6.21}$$

can not be contained in X, i.e., there exists $g \in A$ with $C \cap gB \neq \emptyset$, which means $C \cap AB \neq \emptyset$.

We begin by computing the average $\langle\varphi, 1\rangle$, which is easy: we have

$$\langle\varphi, 1\rangle = \frac{1}{|G|^2}\sum_{g\in G}\sum_{x\in C\cap gB} 1 = \frac{1}{|G|^2}\sum_{x\in C}\sum_{x\in gB} 1 = \mu(B)\mu(C).$$

To bound the variance, we start by writing $\varphi = T(\delta_C)$ where δ_C is the characteristic function of C and T is the linear operator defined by

$$(T\psi)(g) = \frac{1}{|G|}\sum_{x\in B}\psi(gx).$$

We denote $\varphi_0 = \varphi - \langle\varphi, 1\rangle = T\widetilde{\delta}_C$ where

$$\widetilde{\delta}_C = \delta_C - \mu(C),$$

so that $\widetilde{\delta}_C$ is in the subspace $L^2_0(G)$ of functions of average 0. We note that T acts on $L^2_0(G)$. We have

$$V = \|\varphi_0\|^2 = \|T\widetilde{\delta}_C\|^2 = \langle T^*T\widetilde{\delta}_C, \widetilde{\delta}_C\rangle \leqslant \lambda^2\|\widetilde{\delta}_C\|^2,$$

where $\lambda^2 \geqslant 0$ is the largest eigenvalue of the positive self-adjoint operator T^*T acting on $L^2_0(G)$. The norm of $\widetilde{\delta}_C$ is easy to compute: we have

$$\|\widetilde{\delta}_C\|^2 \leqslant \|\delta_C\|^2 = \mu(C)(1-\mu(C)) < \mu(C).$$

We are therefore reduced to estimating from above the eigenvalue λ^2. This allows us to bring in representation theory, because T, and $T_2 = T^*T$, commute with the left-regular representation reg defined in Proposition 6.2.4, i.e., we have

$$\mathrm{reg}(x)(T_2\psi) = T_2(\mathrm{reg}(x)\psi)$$

for all $\psi \in L^2(G)$ and $x \in G$. As a consequence (and exactly as in Proposition 6.2.4, (2)), the λ^2-eigenspace (say W) of T_2 is a subrepresentation of G. Since it is orthogonal to the space of constant functions, this subrepresentation does not contain an invariant vector, and hence the dimension of W is at least $d(G)$.

As we did in the proof of Corollary 6.2.5, we now use this multiplicity and positivity to deduce that

$$\lambda^2 d(G) \leqslant \mathrm{Tr}(T_2) = \frac{1}{|G|^2}\sum_{x,y\in B}\mathrm{Tr}(\mathrm{reg}'(x^{-1}y)) = \mu(B),$$

where $\mathrm{reg}'(g)$ is the right-regular representation operator

$$\mathrm{reg}'(g)\psi(x) = \psi(xg),$$

which has trace 0 if $g \neq 1$, and $|G|$ otherwise.[2] We have therefore obtained

$$V \leqslant \lambda^2\|\widetilde{\delta}_C\|^2 < \frac{\mu(B)\mu(C)}{d(G)},$$

and the condition (6.21) which ensures that $C \cap AB \neq \emptyset$ is implied by

$$\mu(A) \geqslant \frac{\mu(B)\mu(C)}{d(G)}\frac{1}{\langle\varphi, 1\rangle^2} = \frac{1}{\mu(B)\mu(C)d(G)},$$

which is the statement of Proposition 6.5.3. □

[2] Compute the trace using the basis of $L^2(G)$ of characteristic functions.

6.6. Growth of generating subsets of $\mathrm{SL}_2(\mathbf{F}_p)$

Finally, in the remaining sections, we will prove Helfgott's growth theorem [52]:

Theorem 6.6.1 (Helfgott). *There exists a constant $\delta > 0$ such that for any prime p and any subset $H \subset \mathrm{SL}_2(\mathbf{F}_p)$ which generates $\mathrm{SL}_2(\mathbf{F}_p)$, we have either*

$$|H \cdot H \cdot H| \geqslant |H|^{1+\delta},$$

or

$$H \cdot H \cdot H = \mathrm{SL}_2(\mathbf{F}_p).$$

Remark 6.6.2. (1) In [70, Th. 1.2], we showed that one can take $\delta = 1/3024$, and a recent preprint of Rudnev and Shkredov [98] improves this very significantly to $1/40$. It is an interesting problem to determine what might be the best possible value of δ (and the analogue for other groups). Button and Roney-Dougal [23] have shown that δ must be $\geqslant \frac{1}{6}(\log(7)/\log(2) - 1) = 0.301\ldots$, and their argument suggests that this might be the correct value.

(2) Exercise 6.6.6 gives examples to show that the analogue statement fails if $H \cdot H \cdot H$ is replaced by $H \cdot H$.

The interpretation of this theorem is usually that a subset $H \subset \mathrm{SL}_2(\mathbf{F}_p)$ "grows" significantly under product, in the sense that

$$\mathrm{trp}(H) \geqslant |H|^{\delta},$$

unless it can not grow for relatively obvious reasons: either H is contained in a proper subgroup, or it is already so large that the triple product is all of $\mathrm{SL}_2(\mathbf{F}_p)$.

The argument that we present is essentially the one sketched by Pyber and Szabó in [97, §1.1], which is expanded in their paper to cover much more general situations. It is in many ways similar related to the reasoning of Breuillard, Green and Tao [16], and many ingredients are already visible in Helfgott's original paper [52].

We start, however, with some discussion of possible motivations for the proof and the ideas involved. Suppose you just wondered whether a statement like Theorem 6.6.1 holds or not, or suppose you didn't believe in it and wanted to find a counterexample. The following might be a plausible line of argument: a set H which *does not grow* at all in a finite group G is any proper subgroup. Of course, this is the reason why the statement of the theorem applies only to generating sets H, but suppose H is a subgroup of $\mathrm{SL}_2(\mathbf{F}_p)$ together with just one extra element (chosen – we assume this is possible – so that one obtains a generating set). Does such a set "grow"? To be specific, take

$$H = h \cup \left\{ \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \mid t \in \mathbf{F}_p \right\}$$

where h is any fixed matrix in $\mathrm{SL}_2(\mathbf{F}_p)$ with bottom-left coefficient non-zero, so $|H| = p + 1$. What can one say about the size of $H \cdot H \cdot H$? Why should it be significantly larger than $|H|$?

The naive idea is that one can, at least, write down many elements which are products of few elements of H and which "look different": in a non-abelian group, especially one which is rather complicated, there are often no obvious coincidences in the group of the values of "words" written using different generators. Here is an illustration of this idea, which we select because it is elementary, and yet closely related to ideas found later in the proof (indeed, we will use this statement at the end.)

Proposition 6.6.3 (Non-concentration example). *Let p be a prime number and let*

$$U = \left\{ \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \mid t \in \mathbf{F}_p \right\} \subset \mathrm{SL}_2(\mathbf{F}_p).$$

For any symmetric subset H of $\mathrm{SL}_2(\mathbf{F}_p)$ not contained in the subgroup B of upper-triangular matrices, in particular for any symmetric generating set H of $\mathrm{SL}_2(\mathbf{F}_p)$, we have

$$|U \cap H| \leqslant 2|H^{(5)}|^{1/3}$$

where $H^{(5)} = H \cdot H \cdot H \cdot H \cdot H$ is the 5-fold product set.

Such an inequality naturally leads to growth results: if U contains most of H (as in the example above!), it follows that the 5-fold product set must be much larger. In particular, in the example, we get

$$|H^{(5)}| \geqslant \frac{1}{2}p^3,$$

which shows that such a set is, in fact, rather fast-growing! (Since $|H^{(5)}| > \frac{1}{2}|\mathrm{SL}_2(\mathbf{F}_p)|$, the ten-fold product set will be all of $\mathrm{SL}_2(\mathbf{F}_p)$).

Proof. We try to implement this idea of making many products which look different, and in particular seem to escape from U. For this, we first observe that since H is not contained in the subgroup

$$B = \left\{ \begin{pmatrix} a & b \\ 0 & a^{-1} \end{pmatrix} \mid a \in \mathbf{F}_p^*,\ b \in \mathbf{F}_p \right\}$$

of upper-triangular matrices, there exists $h \in H$ of the form

$$h = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with $c \neq 0$. Since H is symmetric, we also have $h^{-1} \in H$.

Now consider $U^* = U - 1$, and define a map

$$\psi : \begin{cases} U^* \times U^* \times U^* & \longrightarrow & \mathrm{SL}_2(\mathbf{F}_p) \\ (u_1, u_2, u_3) & \mapsto & u_1 h u_2 h^{-1} u_3 \end{cases} \tag{6.22}$$

If we restrict ψ to $(U^* \cap H)^3$, we see that $\psi((U^* \cap H)^3) \subset H^{(5)}$. So we can estimate the size of $U^* \cap H$ by summing according to the values of ψ, namely

$$|U^* \cap H|^3 = \sum_{x \in \psi((U^* \cap H)^3)} |\psi^{-1}(x) \cap (U^* \cap H)^3|.$$

Here is the crucial point: for each $x \in \mathrm{SL}_2(\mathbf{F}_p)$, the inverse image $\psi^{-1}(x)$ is either empty or a single point (note that this is intuitively not un-reasonable, because the size of the domain of ψ is about the same as the size of $\mathrm{SL}_2(\mathbf{F}_p)$.) Using this, we get

$$|U^* \cap H|^3 \leqslant |\psi((U^* \cap H)^3)| \leqslant |H^{(5)}|,$$

and then we add again the element $1 \in U \cap H$ to get

$$|U \cap H| = 1 + |U^* \cap H| \leqslant 1 + |H^{(5)}|^{1/3} \leqslant 2|H^{(5)}|^{1/3},$$

finishing the proof.

To check the claim, we compute... Precisely, if

$$u_i = \begin{pmatrix} 1 & t_i \\ 0 & 1 \end{pmatrix} \in U^*,$$

a matrix multiplication leads to

$$\psi(u_1, u_2, u_3) = \begin{pmatrix} 1 - t_1 t_2 c^2 - t_2 ac & \star \\ -t_2 c^2 & \star \end{pmatrix},$$

and in order for this to be a fixed matrix $x \in \mathrm{SL}_2(\mathbf{F}_p)$, we see that t_2 (i.e., u_2) is uniquely determined (since $c \neq 0$). Moreover, since u_2 is in U^*, we have $t_2 \neq 0$ (we defined ψ using U^* in order to ensure this...) Thus t_1 (i.e. u_1) is also uniquely determined, and finally

$$u_3 = (u_1 h u_2 h^{-1})^{-1} x$$

is uniquely determined, if it exists... □

We now start the proof of Helfgott's Theorem, following [97]. The point of view towards $\mathrm{SL}_2(\mathbf{F}_p)$ is that, for the most part, it consists of elements which are *diagonalizable* with distinct eigenvalues (though not necessarily with eigenvalues in the field $\mathbf{F}_p$ itself). Such elements produce a certain amount of extra structure: they come in "packets", which are the sets of all elements of this type which are diagonalized in the same basis. As it turns out, sets with small tripling constant tend to be equitably distributed among such "packets"...

We begin with an important observation, which applies to all finite groups, and goes back to Ruzsa: to prove that the tripling constant of a generating set H is at least a small power of $|H|$, it is enough to prove that the growth ratio after an arbitrary (but fixed) number of products is of such order of magnitude. Note that Proposition 6.6.3 would suggest strongly that we look for such a relation, since we obtain the growth of the five-fold product set, not of $H \cdot H \cdot H$, and we would like to understand the relation between the two. But before stating Ruzsa's lemma, we take the opportunity to introduce more generally the notation for k-fold product sets.

Definition 6.6.4. Let H be a subset of a group G, and let $n \geqslant 0$ be an integer. We define the *n-fold symmetric product set*

$$H^{(n)} = \{x \in G \mid x = a_1 \cdots a_n \text{ for some } a_i \in H \cup H^{-1} \cup \{1\}\}.$$

Note the immediate relations

$$(H^{(n)})^{-1} = H^{(n)}, \quad (H^{(n)})^{(m)} = H^{(nm)}, \quad H^{(n+m)} = H^{(n)} \cdot H^{(m)}$$

for n, $m \geqslant 0$.

Proposition 6.6.5 (Ruzsa's Lemma). *Let G be a finite group, and let $H \subset G$ be a non-empty symmetric subset.*

(1) *Denoting $\alpha_n = |H^{(n)}|/|H|$, we have*

$$\alpha_n \leqslant \alpha_3^{n-2} = \mathrm{trp}(H)^{n-2} \tag{6.23}$$

for all $n \geqslant 3$.

(2) *We have $\mathrm{trp}(H^{(2)}) \leqslant \mathrm{trp}(H)^4$ and for $k \geqslant 3$, we have*

$$\mathrm{trp}(H^{(k)}) \leqslant \mathrm{trp}(H)^{3k-3}.$$

Proof. The first part is proved by induction on $n \geqslant 3$, with the initial case $n = 3$ being tautological. For the induction step, assuming (6.23) for some $n \geqslant 3$, we use the triangle inequality for the Ruzsa distance

$$d(A, B) = \log\Big(\frac{|A \cdot B^{-1}|}{\sqrt{|A||B|}}\Big) \leqslant d(A, C) + d(C, B)$$

between non-empty subsets of G (Lemma A.1.2 in Appendix A). We denote $h_n = |H^{(n)}|$ for simplicity, and then write

$$\alpha_{n+1} = \frac{h_{n+1}}{h_1} = \frac{|H^{(n-1)} \cdot H^{(2)}|}{h_1},$$

which we express in terms of the Ruzsa distance $d(H^{(n-1)}, H^{(2)})$ (exploiting the fact that H and its k-fold product sets are symmetric), namely

$$\begin{aligned}\alpha_{n+1} &= h_1^{-1} h_{n-1}^{1/2} h_2^{1/2} \exp(d(H^{(n-1)}, H^{(2)})) \\ &\leqslant h_1^{-1} h_{n-1}^{1/2} h_2^{1/2} \exp(d(H^{(n-1)}, H^{(1)}) + d(H^{(1)}, H^{(2)})) \\ &= h_1^{-1} h_{n-1}^{1/2} h_2^{1/2} (h_n h_{n-1}^{-1/2} h_1^{-1/2})(h_3 h_1^{-1/2} h_2^{-1/2}) \\ &= \frac{h_{n-1}}{h_1} \frac{h_3}{h_1} = \alpha_{n-1}\alpha_3 \leqslant \alpha_3^{n-2+1},\end{aligned}$$

using the induction assumption, and completing the proof of (6.23).

For (2), we have

$$\operatorname{trp}(H^{(k)}) = \frac{h_{3k}}{h_k} = \frac{\alpha_{3k}}{\alpha_k}.$$

Since $\alpha_k \geqslant \alpha_3$ for $k \geqslant 3$, we obtain $\operatorname{trp}(H^{(k)}) \leqslant \alpha_3^{3k-3}$ for $k \geqslant 3$ by (1), while for $k \geqslant 2$, we simply use $\alpha_2 \geqslant 1$ to get $\operatorname{trp}(H^{(2)}) \leqslant \alpha_3^4$. □

Exercise 6.6.6. This exercise shows that it is crucial in Ruzsa's Lemma to start with the 3-fold product set, and not – as one might naively hope – the 2-fold one. Similarly, Helfgott's Growth Theorem does not hold for the 2-fold product set.

Let p be a prime, $G = \mathrm{SL}_2(\mathbf{F}_p)$ and B the subgroup of upper-triangular matrices. Denote

$$w = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \notin B,$$

and define $H = B \cup \{w, w^{-1}\}$.

(1) Show that there exists no $A \geqslant 0$ such that

$$\operatorname{trp}(H) \leqslant \Bigl(\frac{|H^{(2)}|}{|H|}\Bigr)^A$$

for all p.

(2) Show that Helfgott's Theorem does not hold if $H^{(3)}$ is replaced with $H^{(2)}$ in the statement.

Our first use of Ruzsa's Lemma is the following:

Lemma 6.6.7 (Very small sets grow). *Let G be a finite group and let H be a symmetric generating set of G containing 1. If $H^{(3)} \neq G$, we have $|H^{(3)}| \geqslant 2^{1/2}|H|$.*

Proof. The argument is in fact a bit similar to that in Proposition 6.6.3. If the triple product set is not all of G, it follows that $H^{(3)} \neq H^{(2)}$. We fix some $x \in H^{(3)} - H^{(2)}$, and consider the injective map

$$i : \begin{cases} H & \longrightarrow & G \\ h & \mapsto & hx \end{cases}.$$

The image of this map is contained in $H^{(4)}$ and it is disjoint with H since $x \notin H^{(2)}$. Hence $H^{(4)}$, which contains H and the image of i, satisfies

$$|H^{(4)}| \geqslant 2|H|.$$

By Ruzsa's Lemma, we obtain

$$\operatorname{trp}(H) \geqslant \Big(\frac{|H^{(4)}|}{|H|}\Big)^{1/2} \geqslant 2^{1/2}.$$

□

In particular, if $|H|$ is bounded by an absolute constant (say, by 120), and p is not so small that $H^{(3)} = G$, then there exists $\delta > 0$ such that $|H^{(3)}| \geqslant |H|^{1+\delta}$. This means that in proving Helfgott's Theorem, we do not have to worry about small subsets H.

Combined with Proposition 6.6.5 and with Theorem 6.5.1, this shows that in order to prove Helfgott's Theorem 6.6.1, it is enough to exhibit a real number $x \geqslant 2$, an integer $m \geqslant 3$ and $\delta > 0$, all being absolute constants, such that for all primes $p > x$ and all symmetric generating subsets $H \subset \mathrm{SL}_2(\mathbf{F}_p)$, we have either

$$|H| \geqslant 2|\,\mathrm{SL}_2(\mathbf{F}_p)|^{8/9},$$

or

$$|H^{(m)}| \geqslant |H|^{1+\delta}.$$

Indeed, if $p \leqslant x$, we also have $|H| \leqslant x$, and the above remark applies. Otherwise, we get $H^{(3)} = \mathrm{SL}_2(\mathbf{F}_p)$ in the first case, by Theorem 6.5.1, and

$$\operatorname{trp}(H) = \frac{|H^{(3)}|}{|H|} \geqslant \Big(\frac{|H^{(m)}|}{|H|}\Big)^{\frac{1}{m-2}} \geqslant |H|^{\delta/(m-2)}$$

in the second, by Ruzsa's Lemma.

We now finish this section by introducing the crucial definitions for the Pyber-Szabó argument, which will be implemented in the next section. These are, again, quite classical notions in group theory. One point which is important in the general picture (though we do not make really essential use of it for our semi-ad-hoc proof) is to introduce the infinite groups $\mathrm{SL}_2(\bar{\mathbf{F}}_p)$ in addition to $\mathrm{SL}_2(\mathbf{F}_p)$. In the remainder of this chapter, we will denote $\mathbf{G}_p = \mathrm{SL}_2(\bar{\mathbf{F}}_p)$, and often simply write $\mathbf{G}$, and similarly we denote $G = G_p = \mathrm{SL}_2(\mathbf{F}_p)$.

We recall that the Frobenius automorphism σ of $\bar{\mathbf{F}}_p$ is the field automorphism $x \mapsto x^p$. We will also write x^σ instead of $\sigma(x)$.

The Frobenius automorphism has the crucial property that

$$\mathbf{F}_p = \{x \in \bar{\mathbf{F}}_p \mid x^\sigma = x\},$$

and more generally that the extension $\mathbf{F}_{p^n}$ of degree n of $\mathbf{F}_p$ contained in $\bar{\mathbf{F}}_p$ is

$$\mathbf{F}_{p^n} = \{x \in \bar{\mathbf{F}}_p \mid x^{\sigma^n} = x\}.$$

It is also important that σ acts on many other objects. For instance, we extend the definition to $\bar{\mathbf{F}}_p^n$, for any integer $n \geqslant 1$, by acting on each coordinate:

$$(x_1, \ldots, x_n)^\sigma = (x_1^\sigma, \ldots, x_n^\sigma),$$

in which case we can also write

$$\mathbf{F}_p^n = \{x \in \mathbf{F}_p^n \mid x^\sigma = x\}.$$

The Frobenius also acts on $\mathbf{G}_p$ by

$$x^\sigma = \begin{pmatrix} a^\sigma & b^\sigma \\ c^\sigma & d^\sigma \end{pmatrix} \text{ for } x = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Then, we have again the fixed-point property

$$G_p = \{x \in \mathbf{G}_p \mid x^\sigma = x\},$$

and we can also observe relations like

$$(xy)^\sigma = x^\sigma y^\sigma, \qquad \det(x^\sigma) = \det(x)^\sigma,$$

and if $x \in \mathbf{G}_p$ and $e \in \bar{\mathbf{F}}_p^2$, then

$$(xe)^\sigma = x^\sigma e^\sigma.$$

And now we can begin...

Definition 6.6.8 (Regular semisimple elements; maximal tori). Fix a prime number p and let $G = \mathrm{SL}_2(\mathbf{F}_p)$, $\mathbf{G} = \mathrm{SL}_2(\bar{\mathbf{F}}_p)$.

(1) An element $x \in \mathbf{G}$ is *semisimple* if it is *diagonalizable* (in some basis). It is *regular semisimple* if in addition the eigenvalues α, β of x, which are elements of $\bar{\mathbf{F}}_p$, are distinct. For any subset $H \subset \mathbf{G}$, we write H_{reg} for the set of regular semisimple elements in H.[3]

(2) A *maximal torus* $\mathbf{T}$ in $\mathbf{G}$ is the centralizer of a regular semisimple element x, i.e., a subgroup of the form

$$\mathbf{T} = C_{\mathbf{G}}(x)$$

for some $x \in \mathbf{G}_{reg}$. A *maximal torus* T in G is a subgroup of the form $T = \mathbf{T} \cap G$, where $\mathbf{T} \subset \mathbf{G}$ is a maximal torus of the infinite group $\mathbf{G}$ which is σ-invariant, i.e., such that $x^\sigma \in \mathbf{T}$ for all $x \in \mathbf{T}$.

The basic properties of regular semisimple elements, of their centralizers, and of maximal tori are stated in the next propositions. The first concerns statements about the infinite group, whereas the second deals with results about the finite ones. Most of these extend, with suitable definitions, to much more general groups.

Proposition 6.6.9. *Fix a prime number p and let $\mathbf{G} = \mathrm{SL}_2(\bar{\mathbf{F}}_p)$.*

(1) *A subgroup $\mathbf{T} \subset \mathbf{G}$ is a maximal torus if and only if there exists an ordered basis (e_1, e_2) of $\bar{\mathbf{F}}_p^2$ such that $\mathbf{T}$ is the set of elements $x \in \mathbf{G}$ which are diagonal with respect to the basis (e_1, e_2).*

(2) *A regular semisimple element $x \in \mathbf{G}$ is contained in a* unique *maximal torus $\mathbf{T}$, namely its centralizer $\mathbf{T} = C_{\mathbf{G}}(x)$. In particular, if $\mathbf{T}_1 \neq \mathbf{T}_2$ are two maximal tori, we have*

$$\mathbf{T}_{1,reg} \cap \mathbf{T}_{2,reg} = \emptyset, \tag{6.24}$$

and $\mathbf{T} \cap \mathbf{T}^\sigma = \{\pm 1\}$ if $\mathbf{T}$ is a maximal torus with $\mathbf{T}^\sigma \neq \mathbf{T}$.

(3) *For any maximal torus $\mathbf{T} \subset \mathbf{G}$ and any $y \in \mathbf{G}$, the conjugate subgroup $y\mathbf{T}y^{-1}$ is a maximal torus.*

(4) *If $\mathbf{T} \subset \mathbf{G}$ is a maximal torus, we have*

$$|\mathbf{T}_{nreg}| = |\mathbf{T} - \mathbf{T}_{reg}| \leqslant 2,$$

with equality if p is odd.

(5) *For any maximal torus $\mathbf{T}$, the normalizer $N_{\mathbf{G}}(\mathbf{T})$ contains $\mathbf{T}$ as a subgroup of index 2.*

(6) *The conjugacy class $\mathbf{Cl}(g)$ of a regular semisimple element $g \in \mathbf{G}$ is the set of all $x \in \mathbf{G}$ such that $\mathrm{Tr}(x) = \mathrm{Tr}(g)$. The set of elements in $\mathbf{G}$ which are not regular semisimple is the set of all $x \in \mathbf{G}$ such that $\mathrm{Tr}(x)^2 = 4$.*

To a large extent, this is just linear algebra, but we provide the proofs. Many readers will probably want to skip them or to find arguments on their own.

[3] Regular semisimple elements, for SL_3, already appeared in Section 5.3.

Proof. (1) Let $\mathbf{T}$ be a maximal torus, and x a regular semisimple element such that $\mathbf{T} = C_{\mathbf{G}}(x)$. Let (e_1, e_2) be an ordered basis of eigenvectors of x. Then any $y \in \mathbf{G}$ which is diagonal in the basis (e_1, e_2) commutes with x, hence belongs to $\mathbf{T}$. Conversely, if $y \in C_{\mathbf{G}}(x)$, then $y(e_1)$ and $y(e_2)$ are eigenvectors of x with the same eigenvalues as e_1 and e_2 respectively. *Because* x is regular, this means that $y(e_1)$ is proportional to e_1, and $y(e_2)$ is proportional to e_2, which means that y is diagonal in the basis (e_1, e_2). Hence $\mathbf{T}$ is the subgroup of elements diagonal in the basis (e_1, e_2).

Conversely, let (e_1, e_2) be a fixed basis. Let x in $\mathbf{G}$ be an element such that x is diagonal with different eigenvalues in this basis. Then x is regular semisimple. By the previous reasoning, its centralizer $\mathbf{T} = C_{\mathbf{G}}(x)$ is a maximal torus, and it coincides with the elements that are diagonal in the basis (e_1, e_2).

(2) Let x be a regular semisimple element. It is of course contained in its centralizer. By (1), any maximal torus $\mathbf{T}$ that contains x is the set of elements that are diagonal in some ordered basis (e_1, e_2). The only other bases for which this is true are $(\alpha e_1, \beta e_2)$ and $(\beta e_2, \alpha e_1)$, where α and β are non-zero elements of $\bar{\mathbf{F}}_p$ (again because the distinct eigenvalues pinpoint the basis vectors as proportional to e_1 or e_2), and these define the same maximal torus.

The last assertions follow: for instance, if $\mathbf{T}$ is a maximal torus that is not equal to $\mathbf{T}^{\sigma}$, then $\mathbf{T} \cap \mathbf{T}^{\sigma}$ only contains diagonalizable elements which are not regular, and only 1 and -1 satisfy this condition.

(3) This is straightforward, for instance because if $\mathbf{T} = C_{\mathbf{G}}(x)$, then $y\mathbf{T}y^{-1} = C_{\mathbf{G}}(yxy^{-1})$, and yxy^{-1} has the same eigenvalues as x.

(4) If we view $\mathbf{T}$ as the group of elements that are diagonal in a basis (e_1, e_2), then the set $\mathbf{T} - \mathbf{T}_{reg}$ is the set of elements x of $\mathbf{G}$ that are diagonal in that basis, but with a repeated eigenvalue, say α. Since the determinant of x is equal to 1, we have $\alpha^2 = 1$, hence α is either 1 or -1. There is only one diagonal elements with eigenvalues 1 and one with eigenvalues -1, so $|\mathbf{T} - \mathbf{T}_{reg}| \leqslant 2$, with equality if $p \geqslant 3$.

(5) Suppose that $\mathbf{T}$ is the group of elements diagonal in the basis (e_1, e_2). Then a straightforward computation shows that the normalizer of $\mathbf{T}$ is the group of elements that *permute* the lines generated by e_1 and e_2. The elements of $N_{\mathbf{G}}(\mathbf{T})$ that are not in $\mathbf{T}$ are of the form

$$\begin{pmatrix} 0 & a \\ -a^{-1} & 0 \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} w, \qquad \text{where} \quad w = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

with $a \in \bar{\mathbf{F}}_p^{\times}$, which shows that $N_{\mathbf{G}}(\mathbf{T})/\mathbf{T}$ is of order 2 (with the coset of w representing the non-trivial element).

(6) Let g be a regular semisimple element. Any element conjugate to g has the same trace. Conversely, any element x with trace $\mathrm{Tr}(g)$ has characteristic polynomial $X^2 - \mathrm{Tr}(g)X + 1$ equal to that of g; since it has distinct roots, the element x has two distinct eigenvalues equal to those of g. This implies that x is conjugate to g (conjugation being given by a matrix mapping a basis of eigenvectors of one matrix to a basis of eigenvectors of the other).

Finally, for any $x \in \mathbf{G}$, the characteristic polynomial is $X^2 - \mathrm{Tr}(x)X + 1$. Therefore x fails to have distinct roots in $\bar{\mathbf{F}}_p$ if and only if the discriminant $\mathrm{Tr}(x)^2 - 4$ is 0. □

For the finite groups, our arguments will be sometimes more ad-hoc.

Proposition 6.6.10. *Fix a prime number p and let $G = \mathrm{SL}_2(\mathbf{F}_p)$, $\mathbf{G} = \mathrm{SL}_2(\bar{\mathbf{F}}_p)$. Let $\varepsilon \in \mathbf{F}_p$ be a fixed element that is not a square, and fix an element $\sqrt{\varepsilon} \in \mathbf{F}_{p^2}$ with square ε.*

(1) *A maximal torus $T \subset G$ is either conjugate to the group*

$$T_s = \left\{ \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} \mid a \in \mathbf{F}_p^{\times} \right\}, \tag{6.25}$$

or to the subgroup

$$T_{ns} = \left\{ \begin{pmatrix} a & b \\ b\varepsilon & a \end{pmatrix} \mid a^2 - \varepsilon b^2 = 1 \right\}. \tag{6.26}$$

(2) *For any maximal torus $T \subset G$ and any $y \in G$, the subgroup yTy^{-1} is a maximal torus of G.*

(3) *For any maximal torus $T \subset G$, the normalizer $N_G(T)$ contains T as a subgroup of index 2; it is conjugate to either*

$$N_G(T_s) = \left\{ \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} \mid a \in \mathbf{F}_p^{\times} \right\} \cup \left\{ \begin{pmatrix} 0 & a \\ -a^{-1} & 0 \end{pmatrix} \mid a \in \mathbf{F}_p^{\times} \right\}, \tag{6.27}$$

or to

$$N_G(T_{ns}) = \left\{ \begin{pmatrix} a & b \\ \varepsilon b & a \end{pmatrix} \mid a^2 - \varepsilon b^2 = 1 \right\} \cup \left\{ \begin{pmatrix} \alpha & \beta \\ -\varepsilon\beta & -\alpha \end{pmatrix} \mid -\alpha^2 + \varepsilon\beta^2 = 1 \right\} \tag{6.28}$$

and

$$2(p-1) \leqslant |N_G(T)| \leqslant 2(p+1). \tag{6.29}$$

Remark 6.6.11. A maximal torus in $SL_2(\mathbf{F}_p)$ is called *split* if it is conjugate to the diagonal subgroup (6.25), and *non-split* if it is conjugate to the group (6.26).

Proof. (1) By definition, there exists a maximal torus $\mathbf{T} \subset \mathbf{G}$ such that $T = \mathbf{T} \cap G$ that is invariant under σ. Let (e_1, e_2) be a basis of $\bar{\mathbf{F}}_p^2$ such that $\mathbf{T}$ is the group of elements of $\mathbf{G}$ diagonal in the basis (e_1, e_2) (Proposition 6.6.9). The image $\mathbf{T}^{\sigma}$ of $\mathbf{T}$ under the Frobenius automorphism is then, on the one hand, the group of elements diagonal in the basis $(e_1^{\sigma}, e_2^{\sigma})$, and on the other hand it is equal to $\mathbf{T}$ by assumption. We see that there are two cases: either e_i^{σ} is proportional to e_i for $i = 1, 2$, or e_i^{σ} is proportional to the other vector e_j. These correspond to the two types of maximal tori, as we will now see.

We first assume that there exist t_1 and t_2 in $\bar{\mathbf{F}}_p^{\times}$ such that $e_i^{\sigma} = t_i e_i$ for $i = 1, 2$. The idea is to search for s_1 and s_2 in $\bar{\mathbf{F}}_p^{\times}$ such that the vectors $f_i = s_i e_i$ form a new basis of $\bar{\mathbf{F}}_p^2$ that has coordinates in $\mathbf{F}_p$. This property holds if and only if $f_i^{\sigma} = f_i$, which translates to

$$s_i^{\sigma} t_i e_i = s_i e_i.$$

It will be satisfied provided $s_i^{\sigma} s_i^{-1} = t_i^{-1}$, and this equation is solvable because it states that s_i is a solution of the polynomial equation $X^{p-1} = t_i^{-1}$, of degree $p - 1 \geqslant 1$.

The group $\mathbf{T}$ is still the group of elements diagonal in the basis $(f_1, f_2) \in \mathbf{F}_p^4$. It is then elementary that $T = \mathbf{T} \cap G$ is the group of elements of G diagonal in this basis, and that it is conjugate to the split torus T_s, the conjugating element being the change of basis matrix from the canonical basis of $\mathbf{F}_p^2$ to the basis (f_1, f_2).

In the second case, there exist t_1 and t_2 in $\bar{\mathbf{F}}_p^{\times}$ such that $e_1^{\sigma} = t_1 e_2$ and $e_2^{\sigma} = t_2 e_1$. Applying σ twice, we obtain $e_1^{\sigma^2} = t_1^{\sigma} t_2 e_1$. By solving the equation $s^{p^2-1} = (t_1^{\sigma} t_2)^{-1}$ as in the first case, we find $f_1 = s e_1$ such that $f_1^{\sigma^2} = f_1$. This means now that the coordinates of f_1 belong to the quadratic extension $\mathbf{F}_{p^2}$ of $\mathbf{F}_p$. Moreover, f_1 does not belong to $\mathbf{F}_p^2$, and in fact it is not proportional to an element of $\mathbf{F}_p^2$: if that were the case, say $f_1 = yf$ with $f \in \mathbf{F}_p^2$ and $y \in \bar{\mathbf{F}}_p^{\times}$, we would get

$$s^{\sigma} e_1^{\sigma} = f_1^{\sigma} = y^{\sigma} f = y^{\sigma} y^{-1} f_1 = y^{\sigma} y^{-1} s e_1,$$

so e_1^σ would be proportional both to e_1 and to e_2, contradicting the fact that (e_1, e_2) is a basis.

We can therefore write $f_1 = g_1 + \sqrt{\varepsilon} g_2$ where $g_i \in \mathbf{F}_p^2$. Since f_1 is not proportional to an element of $\mathbf{F}_p^2$, the vectors (g_1, g_2) are linearly independent, and form a basis of $\mathbf{F}_p^2$.

Let now $x \in T$. Then f_1 is an eigenvector of x, and the eigenvalue α is also in $\mathbf{F}_{p^2}$ (since it is a root of a quadratic equation). We write $\alpha = a + b\sqrt{\varepsilon}$. The condition $xf_1 = \alpha f_1$ becomes

$$xg_1 + \sqrt{\varepsilon} x g_2 = (a + b\sqrt{\varepsilon})(g_1 + g_2\sqrt{\varepsilon}) = (ag_1 + b\varepsilon g_2) + \sqrt{\varepsilon}(bg_1 + ag_2),$$

and this implies that the matrix of x with respect to the basis (g_1, g_2) of $\mathbf{F}_p^2$ is

$$\begin{pmatrix} a & b \\ b\varepsilon & a \end{pmatrix} \in T_{ns}.$$

Hence we have shown that a conjugate of T (corresponding to the change of basis) is contained in T_{ns}. But in fact, it is also easy to see that any element of T_{ns} is diagonalizable with eigenvalues $a+b\sqrt{\varepsilon}$ and $a-b\sqrt{\varepsilon}$ in the basis (e_1, e_2). This gives the converse inclusion.

(2) This is straightforward using the definition and the corresponding property of maximal tori in $\mathbf{G}$.

(3) Using (1), we see by direct computations using (6.25) and (6.26) that the normalizer of T_s and T_{ns} are given by (6.27) or by (6.28), so by (1), the normalizer of a maximal torus is conjugate to one of these. We then see that $N_G(T)/T \simeq \mathbf{Z}/2\mathbf{Z}$ in all cases.

To compute the size of $N_G(T)$, it suffices to compute the size of T itself since $|N_G(T)| = [N_G(T) : T]|T| = 2|T|$, and we can assume that $T = T_s$ or that $T = T_{ns}$ by conjugation. We have clearly $|T_s| = p - 1$. For the case of T_{ns}, its size is the number of solutions $(a, b) \in \mathbf{F}_p^2$ of the equation $a^2 - \varepsilon b^2 = 1$. It is well-known that there are $p + 1$ solutions. Indeed, the norm homomorphism $N\colon \mathbf{F}_{p^2}^\times \to \mathbf{F}_p^\times$ given by $N(x) = x^\sigma x = x^{p+1}$ satisfies $N(a + \sqrt{\varepsilon} b) = a^2 - \varepsilon b^2$. The first form shows that its kernel has size at most $p + 1$, since it is the set of solutions of the polynomial equation $X^{p+1} = 1$. Hence the image of the norm has size at least $(p^2 - 1)/(p + 1) = p - 1$. This means that the norm is surjective, and therefore has kernel of size $(p^2 - 1)/(p - 1) = p + 1$. □

Exercise 6.6.12. Let p be a prime with $p \equiv 3 \bmod 4$. Show that we can take $\varepsilon = -1$ in the previous proposition.

Here is one last fact that we will use.

Lemma 6.6.13. *Let p be a prime number. Let $\mathbf{B} \subset \mathbf{G}$ be the subgroup of upper-triangular matrices, and let $x \in \mathbf{G}$. Then $x\mathbf{B}x^{-1} \cap G$ is either the center $\{\pm 1\}$ of G, or a maximal torus in G, or a G-conjugate of $G \cap \mathbf{B}$.*

Proof. The group $\mathbf{B}$ is the subgroup of elements g of $\mathbf{G}$ such that the first vector e_1 of the canonical basis is an eigenvector of g. Hence $\mathbf{B}_1 = x\mathbf{B}x^{-1}$ is the subgroup of elements such that $e = xe_1$ is an eigenvector.

If $g \in G \cap \mathbf{B}_1$, then e^σ is also an eigenvector of g since $g^\sigma = g$.

Assume first that e^σ is proportional to e. As in the proof of Proposition 6.6.10, we can then replace e by a non-zero multiple f such that $f^\sigma = f$, i.e., f belongs to $\mathbf{F}_p^2$. Since $G \cap \mathbf{B}_1$ is the subgroup of G of elements with f as eigenvector, completing f to a basis (f, f') of $\mathbf{F}_p^2$, it follows that $G \cap \mathbf{B}_1$ is conjugate to $G \cap \mathbf{B}$.

Assume now that e^σ is not proportional to e. Then (e, e^σ) is a basis of $\bar{\mathbf{F}}_p^2$, and $G \cap \mathbf{B}_1$ is contained in the subgroup $\mathbf{T}$ of matrices diagonal in the basis (e, e^σ), indeed $G \cap \mathbf{B}_1 = G \cap \mathbf{T}$. There are two cases: if $\mathbf{T}$ is not σ-invariant, then $G \cap \mathbf{T} \subset \mathbf{T} \cap \mathbf{T}^\sigma$, hence is reduced to $\{\pm 1\}$ by Proposition 6.6.9 (2). Otherwise, $G \cap \mathbf{T}$ is by definition a maximal torus in G. □

EXERCISE 6.6.14. Check that in Lemma 6.6.13, if $x\mathbf{B}x^{-1} \cap G$ is a maximal torus, then it is a non-split maximal torus.

6.7. Proof of the growth theorem

We are now ready to start the proof of Helfgott's Theorem. Variants of the following concept that will be crucial in the argument were introduced (under different names and disguises) by Helfgott, Pyber-Szabó, and Breuillard-Green-Tao. We chose the name from the last team.

DEFINITION 6.7.1 (A set involved with a torus). Let p be a prime number, $H \subset \mathrm{SL}_2(\mathbf{F}_p)$ a finite set and $\mathbf{T} \subset \mathrm{SL}_2(\bar{\mathbf{F}}_p)$ a maximal torus. Then H *is involved with* $\mathbf{T}$, or $\mathbf{T}$ with H, if and only if $\mathbf{T}$ is σ-invariant and H contains a regular semisimple element of $\mathbf{T}$ with non-zero trace, i.e., $H \cap \mathbf{T}_{sreg} \neq \emptyset$ where the superscript "sreg" restricts to regular semisimple elements with non-zero trace.

REMARK 6.7.2. There is a twist in this definition, compared with the one in [97] or [16], namely we insist on having non-zero trace. This will be helpful later on, as it will eliminate a whole subcase in the key estimate (the proof of Proposition 6.7.5), and lead to a shorter proof. However, this restriction is not really essential in the greater scheme of things.

The first step in the proof of Helfgott's Theorem is to ensure that we have some regular semisimple elements to play with, and in fact we will require one with non-zero trace, which means that *some* maximal torus will be involved with H. But it is not always the case that a generating set contains such elements. For instance, the generating set

$$H = \left\{ \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \right\}$$

contains no semisimple element, and a fortiori no regular ones!

However, this can be remedied by replacing H by a fixed k-fold product set. In fact the threefold one is enough for $p \geqslant 7$ (though using any $H^{(k)}$, with k fixed, would not compromise the proof, except for the values of the constants.)

LEMMA 6.7.3 (Helfgott). *Let $p \geqslant 7$ be a prime number and let $H \subset \mathrm{SL}_2(\mathbf{F}_p)$ be a symmetric generating set with $1 \in H$. Then $H^{(3)}_{sreg} \neq \emptyset$, i.e., the three-fold product set $H^{(3)}$ contains a regular semisimple element with non-zero trace.*

In fact, this is a very special case of so-called "escape from subvarieties" properties, which is one basic ingredient in all known proofs of classification of approximate subgroups. Since we only need this special case, we can do it by hand. The reader can safely skip the proof for the moment in order to see where the argument will go.

PROOF. This will be a bit fussy, but we hope that the simple idea will be clear. We assume that H does not contain elements which are regular semisimple, except possibly some with trace 0. We will then make products in various ways to show that $H^{(3)}$ does not share the same sad fate.

The basic point that allows us to give a quick proof is that the set $\mathbf{N} = \mathbf{G} - \mathbf{G}_{reg}$ of elements which are not regular semisimple is invariant under conjugation, and (as observed in Proposition 6.6.9) is the set of all matrices with trace equal to 2 or -2. It is precisely the union of the two central elements ± 1 and the four conjugacy classes of

$$u = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad v = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}, \quad u' = \begin{pmatrix} 1 & \varepsilon \\ 0 & 1 \end{pmatrix}, \quad v' = \begin{pmatrix} -1 & \varepsilon \\ 0 & -1 \end{pmatrix}$$

(where $\varepsilon \in \mathbf{F}_p^\times$ is a fixed non-square) while elements of trace 0 are the conjugates of

$$g_0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

(for this last standard fact, see Proposition B.2.4 in Appendix B).

We next note that, if the statement of the lemma fails for a given H, it also fails for every conjugate of H, and that this allows us to normalize at least one element to a specific representative of its conjugacy class. It is convenient to argue by contradiction, though this is somewhat cosmetic. So we assume that $H^{(3)}_{sreg}$ is empty and $p \geqslant 7$, and will derive a contradiction.

We distinguish two cases. In the first case, we assume that H contains one element of trace ± 2 which is not ± 1. The observation above shows that we can assume that one of u, v, u', v' is in H, and we deal with the case $u \in H$.

Since H is a symmetric generating set, it must contain some element

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

with $c \neq 0$, since otherwise, all elements of H would be upper-triangular, and H would not generate $\mathrm{SL}_2(\mathbf{F}_p)$. Then $H^{(3)}$ contains ug, $u^2 g$, $u^{-1} g$, $u^{-2} g$, which have traces, respectively, equal to $\mathrm{Tr}(g) + c$, $\mathrm{Tr}(g) + 2c$, $\mathrm{Tr}(g) - c$, $\mathrm{Tr}(g) - 2c$. Since $c \neq 0$, and p is not 2 or 3, we see that these traces are distinct, and since there are 4 of them, one at least is not in $\{-2, 0, 2\}$, which contradicts our assumption.

If $v \in H$, the argument is almost identical. If u' (or similarly v') is in H, the set of traces of $(u')^j g$ for $j \in \{-2, -1, 0, 1, 2\}$ is

$$\{\mathrm{Tr}(g) + 2c, -\mathrm{Tr}(g) - c, \mathrm{Tr}(g), -\mathrm{Tr}(g) + c, \mathrm{Tr}(g) - 2c\},$$

and one can check that for $p \geqslant 5$, one of these is not 0, -2 or 2, although some could coincide (for instance, if $\mathrm{Tr}(g) = 2$, the other traces are $\{2 + 2c, -2 - c, -2 + c, 2 - 2c\}$, and if $c - 2 = 2$, we get traces $\{2, -6, 10\}$, but $-6 \notin \{0, 2, -2\}$ for $p \geqslant 5$).

In the second case, all elements of H except ± 1 have trace 0. We split in two subcases, but depending on properties of $\mathbf{F}_p$.

The first one is when -1 is *not* a square in $\mathbf{F}_p$ (in other words, $p \equiv 3 \bmod 4$, see Exercise 6.6.12). Conjugating again, we can assume that $g_0 \in H$. Because H generates $\mathrm{SL}_2(\mathbf{F}_p)$, we claim that there must exist a matrix

$$g = \begin{pmatrix} a & b \\ c & -a \end{pmatrix}$$

in H with (i) $a \neq 0$; (ii) $b \neq c$. Indeed if all elements $\neq \pm 1$ of H are of the form

$$g = \begin{pmatrix} 0 & -c^{-1} \\ c & 0 \end{pmatrix},$$

we can find such an element with $c \neq \pm 1$ (i.e., $g \neq \pm g_0$), since otherwise H is not a generating set; then the trace of $g_0 g$ is $c + c^{-1}$, which is not in $\{-2, 0, 2\}$ (non-zero because -1 is not a square in our first subcase), so $H^{(2)}_{nreg} \neq \emptyset$, which we excluded. So all elements of H, except for ± 1 and $\pm g_0$ are of the type

$$g = \begin{pmatrix} a & b \\ c & -a \end{pmatrix},$$

with $a \neq 0$. Then, if all these satisfied $b = c$, it would follow that H is contained in the normalizer of the non-split maximal torus defined with $\varepsilon = -1$ (see (6.28)), again contradicting the assumption that H is a generating set.

Now we argue with g as above. We have

$$g_0 g = \begin{pmatrix} c & -a \\ -a & -b \end{pmatrix} \in H^{(2)},$$

with non-zero trace $t = c - b$. Moreover, if $t = 2$, i.e., $c = b + 2$, the condition $\det(g_0 g) = 1$ implies

$$-2b - b^2 - a^2 = 1$$

or $(b+1)^2 = -a^2$. Similarly, if $t = -2$, we get $(b-1)^2 = -a^2$. Since $a \neq 0$, it follows in both cases that -1 is a square in $\mathbf{F}_p$, which contradicts our assumption in the first subcase.

Now we come to the second subcase when $-1 = z^2$ is a square in $\mathbf{F}_p$. We can then diagonalize g_0 over $\mathbf{F}_p$, and conjugating again, this means we can assume that H contains

$$g_0' = \begin{pmatrix} z & 0 \\ 0 & -z \end{pmatrix}$$

as well as some other matrix

$$g' = \begin{pmatrix} a & b \\ c & -a \end{pmatrix}$$

(the values of a, b, c are not the same as before; we are still in the case when every element of H has trace 0 except for ± 1).

Now the trace of $g_0' g' \in H^{(2)}$ is $2za$. But we can find g' with $a \neq 0$, since otherwise H would again not be a generating set, being contained in the normalizer (6.27) of a split maximal torus and so this trace is non-zero.

The condition $2za = \pm 2$ would give $za = \pm 1$, which leads to $-a^2 = 1$. But since $1 = \det(g') = -a^2 - bc$, we then get $bc = 0$ for all elements of H. Finally, if all elements of H satisfy $b = 0$, the set H would be contained in the subgroup of upper-triangular matrices. So we can find a matrix in H with $b \neq 0$, hence $c = 0$. Similarly, we can find another

$$g'' = \begin{pmatrix} a & 0 \\ c & -a \end{pmatrix}$$

in H with $c \neq 0$. Taking into account that $z^2 = -1$, computing the traces of $g'g''$ and of $g_0 g' g''$ gives

$$bc - 2, \qquad bcz$$

respectively. If $bc = 2$, the third trace (of an element in $H^{(3)}$) is $2z \notin \{0, 2, -2\}$ since $p \neq 2$, and if $bc = 4$, it is $4z \notin \{0, 2, -2\}$ since $p \neq 5$. And of course if $bc \notin \{2, 4\}$, the first trace is already not in $\{-2, 0, 2\}$. So we are done... □

Example 6.7.4. If $p = 5$, one can check (e.g., with Magma [86]) that the set

$$H = \left\{1, \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}^{\pm 1}, \begin{pmatrix} 2 & 2 \\ 0 & -2 \end{pmatrix}^{\pm 1}, \begin{pmatrix} 2 & 0 \\ 2 & -2 \end{pmatrix}^{\pm 1}\right\} \subset \mathrm{SL}_2(\mathbf{F}_5)$$

is a generating set of $\mathrm{SL}_2(\mathbf{F}_5)$ such that $H^{(3)}$ is contained in the set of matrices of traces -2, 2 and 0. Hence the lemma is sharp, as far as the condition on p goes. (Though one can obtain a similar result, where $H^{(3)}$ is replaced by a higher product set, e.g., $H^{(4)}$ in this example.)

Here is now the Key Proposition that will be used for the proof of Helfgott's Theorem:

Proposition 6.7.5 (Involving dichotomy). (1) *For all prime number p, all subsets $H \subset \mathrm{SL}_2(\mathbf{F}_p)$ and all maximal tori $\mathbf{T} \subset \mathrm{SL}_2(\bar{\mathbf{F}}_p)$, if $\mathbf{T}$ and H are not involved, we have*

$$|H \cap \mathbf{T}| \leqslant 4.$$

(2) *There exists $\delta > 0$ such that if $p \geqslant 3$ and $H \subset \mathrm{SL}_2(\mathbf{F}_p) = G$ is a symmetric generating set containing 1, we have*

$$|\mathbf{T}_{reg} \cap H^{(2)}| \gg \alpha^{-4}|H|^{1/3} \tag{6.30}$$

for any maximal torus $\mathbf{T} \subset \mathrm{SL}_2(\bar{\mathbf{F}}_p)$ which is involved with H, where $\alpha = \mathrm{trp}(H)$, unless

$$\alpha \gg |H|^{\delta}, \tag{6.31}$$

where the implied constants are absolute.

Here (1) is obvious, since $|\mathbf{T} - \mathbf{T}_{reg}| \leqslant 2$ (part (4) in Proposition 6.6.9) and there are also at most two elements of trace 0 in $\mathbf{T}$ (as one can check quickly), but (2) is much more delicate, and is – in final analysis – the deus ex machina for this argument, since it shows essentially that whenever H contains a regular semisimple element of some (σ-invariant) maximal torus, it actually contains many of them, provided the ratio $|H^{(2)}|/|H|$ is "small", which translates to the tripling constant being small via Ruzsa's Lemma.

In the first reading, the conclusion (6.30) should be interpreted as

$$|\mathbf{T}_{reg} \cap H^{(k)}| \gg \alpha^{-m}|H|^{1/3}$$

for some $k \geqslant 1$ and $m \geqslant 0$ independent of p (large enough) – in other words, the specific values $k = 2$, $m = 3$ are irrelevant as long as the actual value of the exponent δ in Helfgott's Theorem is not crucial.

On the other hand, the exponent $1/3$ (in $|H|^{1/3}$) here is "the right one": this will be seen first by the way it fits with the steps of the coming arguments (though there is a little leeway), and also more clearly when we motivate the proof. This proof is deferred until we have seen how, remarkably, this key proposition implies Helfgott's Theorem.

Proof of Theorem 6.6.1. We can assume that $p \geqslant 7$, which means that Lemma 6.7.3 is applicable. We will show that for some $\delta > 0$, we have

$$\mathrm{trp}(H) \gg |H|^{\delta} \tag{6.32}$$

with an absolute implied constant, unless $H^{(3)} = \mathrm{SL}_2(\mathbf{F}_p)$. Then using Lemma 6.6.7, we absorb the small values of p as well as the implied constant in this inequality to derive the form of Helfgott's Theorem we claimed.

We define $\widetilde{H} = H^{(2)}$, so that (by Lemma 6.7.3) there exists at least one maximal torus $\mathbf{T}$ involved with $L = \widetilde{H}^{(2)} = H^{(4)}$.

If, among all maximal tori involved with L, none satisfies (6.30), we obtain directly from Proposition 6.7.5 (applied to $\widetilde{H}$ instead of H) the lower bound

$$\mathrm{trp}(\widetilde{H}) \gg |\widetilde{H}|^{\delta} \gg |H|^{\delta},$$

and since $\operatorname{trp}(\widetilde{H}) \leqslant \alpha^4$ by Ruzsa's Lemma, we get

$$\alpha \gg |H|^{\delta/4} \tag{6.33}$$

which gives (6.32) after renaming δ.

Otherwise, we distinguish two cases.

Case (1). There exists a maximal torus $\mathbf{T}$ involved with L such that, for any $g \in G$, the torus $g\mathbf{T}g^{-1}$ is involved with L.

As we can guess from (6.30) and (6.24), in that case, the set L will tend to be rather large, so $|L|$ is close to $|G|$, *unless* the tripling constant is itself large enough.

Precisely, writing $T = \mathbf{T} \cap G$, we note that the maximal tori

$$gTg^{-1} = (g\mathbf{T}g^{-1}) \cap G$$

are distinct for g taken among representatives of $G/N_G(T)$. Then we have the inequalities

$$|L^{(2)}| \geqslant \sum_{g \in G/N_G(T)} |L^{(2)} \cap g\mathbf{T}_{reg}g^{-1}| \gg \beta^{-4}|L|^{1/3}\frac{|G|}{|N_G(T)|}$$

where $\beta = \operatorname{trp}(L)$, since each $g\mathbf{T}g^{-1}$ is involved with L and distinct regular semisimple elements lie in distinct maximal tori (and we are in a case where (6.30) holds for all tori involved with L).

Now we unwind this inequality in terms of H and $\alpha = \operatorname{trp}(H)$. We have $L^{(2)} = H^{(8)}$, so (using (6.29))

$$|H| \geqslant \alpha^{-6}|L^{(2)}| \gg \alpha^{-6}\beta^{-4}(p-1)^2|L|^{1/3} \gg \alpha^{-6}\beta^{-4}(p-1)^2|H|^{1/3}$$

by Ruzsa's Lemma. Furthermore, we have

$$\beta = \operatorname{trp}(L) = \operatorname{trp}(H^{(4)}) \leqslant \alpha^{10}$$

by Ruzsa's Lemma again, and hence the inequality gives the bound

$$|H| \gg \alpha^{-69}(p-1)^3,$$

which implies that $|H| \gg \alpha^{-69}|G|$. But then either

$$|H| \geqslant 2|G|^{8/9}$$

or $\alpha \gg |G|^{\delta} \gg |H|^{\delta}$ for any $\delta < 1/(9 \cdot 69)$, which are versions of the two alternatives we are seeking.

Case (2). Since we know that *some* torus is involved with L, the complementary situation to Case (1) is that there exists a maximal torus $\mathbf{T}$ involved with $L = H^{(4)}$ and a conjugate $g\mathbf{T}g^{-1}$, for some $g \in G$, which is *not* involved with L. We are then going to get growth in a subgroup of G, which turns out to imply growth in all of G by a simple lemma (Lemma 6.7.6 below). There is a first clever observation (the idea of which goes back to work of Glibichuk and Konyagin [44] on the "sum-product phenomenon"): one can assume, possibly after changing $\mathbf{T}$ and g, that g is in H.

Indeed, to check this claim, we start with $\mathbf{T}$ and h as above. Since H is a generating set, we can write

$$g = h_1 \cdots h_m$$

for some $m \geqslant 1$ and some elements $h_i \in H$. Now let $i \leqslant m$ be the smallest index such that the maximal torus

$$\mathbf{T}' = (h_{i+1} \cdots h_m)\mathbf{T}(h_{i+1} \cdots h_m)^{-1}$$

is involved with L. Taking $i = m$ means that $\mathbf{T}$ is involved with L, which is the case, and therefore the index i exists. Moreover $i \neq 0$, again by definition. It follows that

$$(h_i h_{i+1} \cdots h_m)\mathbf{T}(h_i h_{i+1} \cdots h_m)^{-1}$$

is not involved with L. But this means that we can replace $(\mathbf{T}, g)$ with $(\mathbf{T}', h_i)$, and since $h_i \in H$, this gives us the claim.

We now write h for the conjugator g such that L and the torus $\mathbf{S} = g\mathbf{T}g^{-1} = h\mathbf{T}h^{-1}$ are not involved. We now use the following lemma, which will tell us that, in order to show that H grows, it is enough to show that $H \cap \mathbf{S}$ grows inside $\mathbf{S} \cap G$.

Lemma 6.7.6. *Let G be an arbitrary finite group, $K \subset G$ a subgroup, and $H \subset G$ an arbitrary symmetric subset. For any $n \geqslant 1$, we have*

$$\frac{|H^{(n+1)}|}{|H|} \geqslant \frac{|H^{(n)} \cap K|}{|H^{(2)} \cap K|}.$$

Proof. Let $X \subset G/K$ be the set of cosets of K intersecting H:

$$X = \{xK \in G/K \mid xK \cap H \neq \emptyset\}.$$

We can estimate the size of this set from below by splitting H into its intersections with cosets of K: we have

$$|H| = \sum_{xK \in X} |H \cap xK|.$$

But for any $xK \in X$, fixing some $g_0 \in xK \cap H$, we have $g^{-1}g_0 \in K \cap H^{(2)}$ if $g \in xK \cap H$, and hence

$$|xK \cap H| \leqslant |K \cap H^{(2)}|.$$

This gives the lower bound

$$|X| \geqslant \frac{|H|}{|K \cap H^{(2)}|}.$$

Now take once more some $xK \in X$, and fix an element $xk = h \in xK \cap H$. Then all the elements xkg are distinct for $g \in K$, and they are in $xK \cap H^{(n+1)}$ if $g \in K \cap H^{(n)}$, so that

$$|xK \cap H^{(n+1)}| \geqslant |K \cap H^{(n)}|$$

for any $xK \in X$, and (cosets being disjoint)

$$|H^{(n+1)}| \geqslant |X||K \cap H^{(n)}|,$$

which gives the result when combined with the lower bound for $|X|$. □

Apply Lemma 6.7.6 with $(H, K) = (\widetilde{H}, h\mathbf{T}h^{-1} \cap G)$ and $n = 5$. This gives

$$\frac{|\widetilde{H}^{(6)}|}{|\widetilde{H}|} \geqslant \frac{|\widetilde{H}^{(5)} \cap S|}{|\widetilde{H}^{(2)} \cap S|}.$$

But since $L = \widetilde{H}^{(2)}$ and $\mathbf{S}$ are not involved (by construction), we have $|\widetilde{H}^{(2)} \cap S| \leqslant 2$, by the easy part of the Key Proposition 6.7.5, and therefore

$$\frac{|\widetilde{H}^{(6)}|}{|\widetilde{H}|} \geqslant \frac{1}{2}|\widetilde{H}^{(5)} \cap \mathbf{S}|.$$

However, L and $\mathbf{T}$ *are* involved, and moreover

$$h(H^{(8)} \cap \mathbf{T})h^{-1} \subset H^{(10)} \cap \mathbf{S} = \widetilde{H}^{(5)} \cap \mathbf{S},$$

so that

$$|\widetilde{H}^{(5)} \cap \mathbf{S}| \geqslant |H^{(8)} \cap T| = |L^{(2)} \cap T| \gg \widetilde{\alpha}^{-4}|L|^{1/3}$$

where $\widetilde{\alpha} = \operatorname{trp}(L)$, by the Key Proposition 6.7.5 (again, because (6.30) holds for all tori involved with L).

Thus

$$\frac{|\widetilde{H}^{(6)}|}{|\widetilde{H}|} \gg \widetilde{\alpha}^{-4}|H|^{1/3},$$

which translates to

$$\alpha^{10}|H| \gg \alpha^{-36}|H|^{4/3},$$

by Ruzsa's Lemma. This gives a fairly strong bound for α, namely

$$\alpha = \operatorname{trp}(H) \gg |H|^{1/138}. \tag{6.34}$$

To summarize, we have obtained three possible lower bounds of the right kind for α, and one of them holds if $H^{(3)} \neq \mathrm{SL}_2(\mathbf{F}_p)$. All imply (6.32), and hence we are done. □

We now come to the proof of the Key Proposition. The first observation is that the difficult-looking task of finding a *lower bound* for $\mathbf{T}_{sreg} \cap H^{(k)}$ (for some fixed integer k) is in fact equivalent with a simpler-looking upper-bound involving, instead of $\mathbf{T}$, the conjugacy class of a regular semisimple element for which $\mathbf{T}$ is the centralizer. This is an "approximate" version of the orbit-stabilizer theorem in group theory.

PROPOSITION 6.7.7 (Helfgott). *Let G be a finite group acting on a non-empty finite set X. Fix some $x \in X$ and let $K \subset G$ be the stabilizer of x in G. For any non-empty symmetric subset $H \subset G$, we have*

$$|K \cap H^{(2)}| \geqslant \frac{|H|}{|H \cdot x|}$$

where $H \cdot x = \{h \cdot x \mid h \in H\}$.

(Note that since H is symmetric, we have $1 \in K \cap H^{(2)}$.)

PROOF. As in the classical proof of the orbit-stabilizer theorem, we consider the orbit map, but restricted to H

$$\phi : \begin{cases} H & \longrightarrow & X \\ h & \mapsto & h \cdot x \end{cases}$$

and we use it to count the number of elements in H: we have

$$|H| = \sum_{y \in \phi(H)} |\phi^{-1}(y)|,$$

and the point is that $\phi(H) = H \cdot x$ on the one hand, and

$$|\phi^{-1}(y)| \leqslant |K \cap H^{(2)}|$$

for all y, since if $y = \phi(h_0)$, with $h_0 \in H$, all elements $h \in H$ with $\phi(h) = y$ satisfy $hh_0^{-1} \in K \cap H^{(2)}$. Hence we get

$$|H| \leqslant |H \cdot x||K \cap H^{(2)}|,$$

as claimed. □

As a corollary, let $T = \mathbf{T} \cap \mathbf{G}$ be a maximal torus in G. Fixing any $g \in T_{reg}$, we have $T = C_G(g)$, the stabilizer of g in G for its conjugacy action on itself. We find therefore that

$$|\mathbf{T} \cap H^{(2)}| \geqslant \frac{|H|}{|\{hgh^{-1} \mid h \in H\}|} \tag{6.35}$$

for any symmetric subset H. If we now assume that $\mathbf{T}$ *and* H *are involved*, we can select g in $T_{sreg} \cap H$ in this inequality, and the denominator on the right becomes

$$|\{hgh^{-1} \mid h \in H\}| \leqslant |H^{(3)} \cap \mathrm{Cl}(g)| \leqslant |H^{(3)} \cap \mathbf{Cl}(g)|$$

where $\mathrm{Cl}(g)$ (resp. $\mathbf{Cl}(g)$) is the conjugacy class of g in G (resp. in $\mathbf{G}$). Hence, the Key Proposition follows from an *upper-bound* on the number of elements of $H^{(3)}$ in a given regular semisimple conjugacy class with non-zero trace. We are therefore led to prove the following theorem, which is a special case of what are now called the (generalized) Larsen-Pink inequalities:

Theorem 6.7.8 (Larsen-Pink non-concentration inequality). *Let $p \geqslant 3$ be a prime number and let $g \in \mathrm{SL}_2(\mathbf{F}_p) = G$. There exists $\delta > 0$ such that, if $H \subset G$ is a symmetric generating set and g is regular semisimple with non-zero trace, we have*

$$|\,\mathbf{Cl}(g) \cap H| \ll \alpha^{2/3}|H|^{2/3} \tag{6.36}$$

where $\alpha = \mathrm{trp}(H)$ is the tripling constant of H, unless

$$\alpha \gg |H|^{\delta}. \tag{6.37}$$

The implied constants are absolute.

Applying this result to $H^{(3)}$, with tripling constant bounded by α^6 (by Ruzsa's Lemma), we obtain by (6.35) the lower bound

$$|\mathbf{T} \cap H^{(2)}| \geqslant \frac{|H|}{|H^{(3)} \cap \mathbf{Cl}(g)|} \gg \alpha^{-4}|H|^{1/3},$$

unless $\alpha = \mathrm{trp}(H) \gg |H|^{\delta/6}$. Since there are at most two elements of $\mathbf{T} \cap H^{(2)}$ which are not regular, the first bound gives

$$|\mathbf{T}_{reg} \cap H^{(2)}| \gg \alpha^{-14/3}|H|^{1/3},$$

unless $\alpha^{-4}|H|^{1/3} \ll 1$. This means that Proposition 6.7.5 is proved.

Remark 6.7.9. The original Larsen-Pink inequalities proved in [75, §4] concern the intersection of a set like $\mathbf{Cl}(g)$, defined by algebraic equations in an algebraic group (like $\mathrm{SL}_n(k)$ where k is an algebraically closed field), with a *finite subgroup* of $\mathrm{SL}_n(k)$. It was observed first by Hrushovski [55] that these inequalities extend in a natural way to approximate groups.

The general case of the Larsen-Pink inequality is rather tricky to prove. In particular, it is rather hard to keep track of the constants which appear, although they are, in principle, effective. Our argument is a concrete version of the general arguments involved, with shortcuts that are available in this very specific situation.

The basic idea – which also "explains" the exponent $2/3$ here – is described by Larsen and Pink at the beginning of [75, §4]. We wish to consider a map like

$$\phi \begin{cases} \mathbf{Cl}(g) \times \mathbf{Cl}(g) \times \mathbf{Cl}(g) & \longrightarrow \quad \mathbf{G} \times \mathbf{G} \\ (x_1, x_2, x_3) & \mapsto \quad (x_1 g_1 x_2, x_1 g_2 x_3) \end{cases}$$

(where (g_1, g_2) are parameters), and we hope to ensure – for suitable choice of the auxiliary parameters (g_1, g_2) – that (i) for $(x_1, x_2, x_3) \in (\mathbf{Cl}(g) \cap H)^3$, we have $\phi(x_1, x_2, x_3) \in H^{(k)}$ for some constant k independent of (x_1, x_2, x_3); (ii) the fibers $\phi^{-1}(y_1, y_2)$ of ϕ are all finite with size bounded independently of $(y_1, y_2) \in \mathbf{G} \times \mathbf{G}$, say of size at most $c_1 \geqslant 1$. The hope behind (ii) is that $\mathbf{Cl}(g)^3$ and $\mathbf{G}^2$ have the same dimension, [4] and hence unless something special happens, we would expect the fibers to have dimension 0, which corresponds to having fibers of bounded size since everything is defined using polynomial equations.

If this construction succeeds, we can count $|\mathbf{Cl}(g) \cap H|$ by summing according to the values of ϕ: denoting $Z = (\mathbf{Cl}(g) \cap H)^3$ and $W = \phi(Z) = \phi((\mathbf{Cl}(g) \cap H)^3)$, we have

$$|\mathbf{Cl}(g) \cap H|^3 = |Z| = \sum_{(y_1,y_2) \in W} |\phi^{-1}(y_1, y_2) \cap Z|$$

which – under our optimistic assumption – leads to the estimate

$$|\mathbf{Cl}(g) \cap H|^3 \leqslant c_1 |W| \leqslant c_1 |H^{(k)}|^2 \leqslant c_1 \alpha^{2(k-2)} |H|,$$

which has the form we want.

As it turns out, the assumption on ϕ are too optimistic, but we can notice an encouraging point in any case: even here in this argument, we obtain the desired result despite using overcounting steps which might look dangerous (e.g., bounding the size of $\phi^{-1}(y_1, y_2) \cap Z$ by that of the whole fiber.)

Our first step to attempt a rigorous implementation of the idea is to realize that the parameters g_1, g_2 do not, in fact, play any role (the idea of introducing them is that they give the impression that the products $x_1 g_1 x_2$, $x_1 g_2 x_3$ are " more unrelated", hence more likely to be distinct), and in fact taking $g_1 = g_2 = 1$ has the immediate advantage that it becomes immediately clear that (i) holds for the map ϕ in that case, namely

$$\phi(x_1, x_2, x_3) = (x_1 x_2, x_1 x_3) \in (H^{(2)})^2$$

if $(x_1, x_2, x_3) \in (\mathbf{Cl}(g) \cap H)^3$.

We are therefore led to analyze ruthlessly the fibers of the map ϕ.

Lemma 6.7.10. *Let k be any field, and let $G = \mathrm{SL}_2(k)$. Let $C \subset G$ be a conjugacy class, and define*

$$\phi \begin{cases} C^3 & \longrightarrow & G^2 \\ (x_1, x_2, x_3) & \mapsto & (x_1 x_2, x_1 x_3) \end{cases}.$$

Then for any $(y_1, y_2) \in G \times G$, we have a bijection

$$\begin{cases} C \cap y_1 C^{-1} \cap y_2 C^{-1} & \longrightarrow & \phi^{-1}(y_1, y_2) \\ x_1 & \mapsto & (x_1, x_2, x_3) \end{cases}.$$

In particular, if $k = \bar{\mathbf{F}}_p$ and C is a regular semisimple conjugacy class,[5] *we have a bijection*

$$\phi^{-1}(y_1, y_2) \longrightarrow C \cap y_1 C \cap y_2 C.$$

Proof. Taking x_1 as a parameter, any (x_1, x_2, x_3) with $\phi(x_1, x_2, x_3) = (y_1, y_2)$ can certainly be written $(x_1, x_1^{-1} y_1, x_1^{-1} y_2)$. Conversely, such an element in $\mathrm{SL}_2(k)^3$ really belongs to C^3 (hence to the fiber) if and only if $x_1 \in C$, $x_1^{-1} y_1 \in C$, $x_1^{-1} y_2 \in C$, i.e., if and only if $x_1 \in C \cap y_1 C^{-1} \cap y_2 C^{-1}$, which proves the first part.

[4] Here dimension can be understood intuitively for readers who do not know the rigorous definition in algebraic geometry.

[5] By which we mean the conjugacy class of a regular semisimple element.

For the second part, we need only notice that if C is a regular semisimple conjugacy class, say that of g, then $C = C^{-1}$ because g^{-1} has the same characteristic polynomial as g, hence is conjugate to g. □

We are now led to determine when an intersection of the form $C \cap y_1 C \cap y_2 C$ can be infinite. The answer is as follows:

Lemma 6.7.11 (Pink). *Let $p \geqslant 3$ be a prime and $k = \bar{\mathbf{F}}_p$. Let $g \in \mathrm{SL}_2(k)$ be a regular semisimple element, and denote by C the conjugacy class of g. For y_1, $y_2 \in G$, the intersection $X = C \cap y_1^{-1} C \cap y_2^{-1} C$ is finite, containing at most two elements, unless one of the following cases holds:*

(1) *We have $y_1 = 1$, or $y_2 = 1$ or $y_1 = y_2$.*

(2) *There exists a conjugate $\mathbf{B} = x\mathbf{B}_0 x^{-1}$ of the subgroup*

$$\mathbf{B}_0 = \left\{ \begin{pmatrix} a & b \\ 0 & a^{-1} \end{pmatrix} \right\} \subset \mathrm{SL}_2(k)$$

and an element $t \in \mathbf{B} \cap C$ such that

$$y_1, y_2 \in \mathbf{U} \cup t^2 \mathbf{U} \tag{6.38}$$

where

$$\mathbf{U} = x\mathbf{U}_0 x^{-1}, \qquad \mathbf{U}_0 = \left\{ \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \right\} \subset \mathbf{B}_0.$$

In that case, we have $X \subset C \cap \mathbf{B}$.

(3) *The trace of g is 0.*

Remark 6.7.12. This result is the one place where it is really useful for us to work with $\mathrm{SL}_2(\bar{\mathbf{F}})$. As we will see in the proof, this simplifies the computations.

Note that case (1) is unavoidable in view of the bijection

$$\phi^{-1}(y_1, y_2) \simeq C \cap y_1 C \cap y_2 C,$$

where the triple intersection becomes a double one, which we expect to be of dimension 1, if the elements of the set $\{1, y_1, y_2\}$ are not really distinct. Case (2) is more "exceptional", and we need a bit more care to handle it. As for case (3), it can be detailed much more precisely, but because it is a restriction on the conjugacy class, we can avoid it entirely, by working with "superregular" conjugacy classes, which are those regular conjugacy classes with non-zero trace. (Note that this corresponds to the condition that $-C \neq C$, where $-C$ is the conjugacy class of $(-1) \cdot g$.) We defer the proof of this lemma to the end of this section.

We come back to the case of interest in Theorem 6.7.8, assuming $\mathrm{Tr}(g) \neq 0$. We construct the map

$$\phi \begin{cases} \mathbf{Cl}(g) \times \mathbf{Cl}(g) \times \mathbf{Cl}(g) & \longrightarrow & \mathbf{G} \times \mathbf{G} \\ (x_1, x_2, x_3) & \mapsto & (x_1 x_2, x_1 x_3) \end{cases},$$

and write (as before)

$$Z = (\mathbf{Cl}(g) \cap H)^3, \quad W = \phi(Z) = \phi((\mathbf{Cl}(g) \cap H)^3)$$

to obtain

$$|\,\mathbf{Cl}(g) \cap H|^3 = \sum_{(y_1, y_2) \in W} |\phi^{-1}(y_1, y_2) \cap Z| = S_0 + S_1 + S_2, \tag{6.39}$$

where S_i denotes the sum restricted to a subset $W_i \subset W$, W_0 being the subset where the fiber has order at most 2, while W_1, W_2 correspond to those (y_1, y_2) where cases (1) and

(2) of Lemma 6.7.11 hold. Precisely, we do not put into W_2 the (y_1, y_2) for which both cases (1) and (2) are valid, e.g., $y_1 = 1$, and we *add* to W_1 the cases where $y_1 = -1$, which may otherwise appear in Case (2). We will prove:

PROPOSITION 6.7.13. *With notation as above, we have*

$$S_0 \ll |H^{(2)}|^2 \ll \alpha^2|H|^2, \qquad S_1 \ll |H^{(3)}|^2 \ll \alpha^2|H|^2,$$
$$S_2 \ll \alpha^{34/3}|H|^{5/3},$$

where the implied constants are absolute.

Assuming this, we get immediately

$$|\,\mathbf{Cl}(g) \cap H| \ll \alpha^{2/3}|H|^{2/3} + \alpha^{34/9}|H|^{5/9}$$

from (6.39). Now either the second term is smaller than the first, and we get (6.36) or

$$\alpha^{34/9}|H|^{5/9} \gg \alpha^{2/3}|H|^{2/3}$$

which gives

$$\alpha \gg |H|^{1/28},$$

the second alternative (6.37) of Theorem 6.7.8, which is therefore proved. And with it Helfgott's Theorem...

PROOF. The case of S_0 follows by the fact that the fibers over W_0 have at most two elements, hence also their intersection with Z, and that $|W_0| \leqslant |W| \leqslant |H^{(2)}|^2$.

The case of S_1 splits into four almost identical subcases, corresponding to $y_1 = 1$, $y_1 = -1$ (remember that we added this, borrowing it from Case (2)...), $y_2 = 1$ or $y_1 = y_2$. We deal only with the first, say $S_{1,1}$: we have

$$S_{1,1} \leqslant \sum_{y_2 \in H^{(2)}} |\phi^{-1}(1, y_2) \cap Z|.$$

But using Lemma 6.7.10, we have

$$|\phi^{-1}(1, y_2) \cap Z| = |\{(x_1, x_1^{-1}, x_1^{-1}y_2) \in (\mathbf{Cl}(g) \cap H)^3\}| \leqslant |H^{(3)}|$$

for any given $y_2 \in H^{(2)}$, since x_1 determines the triple $(x_1, x_1^{-1}, x_1^{-1}y_2)$ and $x_1^{-1} = x_1^{-1}y_2y_2^{-1} \in H^{(3)}$ for any such triple if $y_2 \in H^{(2)}$. Therefore

$$S_{1,1} \leqslant |H^{(2)}||H^{(3)}| \leqslant |H^{(3)}|^2,$$

and similarly for the other two cases.

Now for S_2. Here also we sum over y_1 first, which is $\neq \pm 1$ (by our definition of W_2). The crucial point is then that an element $y_1 \neq \pm 1$ is included in at most two conjugates of $\mathbf{B}_0$. Hence, up to a factor 2, the choice of y_1 fixes that of the relevant conjugate $\mathbf{B}$ for which Case (2) applies. Next we observe that $C_{\mathbf{B}} = \mathbf{Cl}(g) \cap \mathbf{B}$ is a conjugate of the union

$$C_\alpha \cup C_{\alpha^{-1}},$$

where

$$C_\alpha = \left\{\begin{pmatrix} \alpha & t \\ 0 & \alpha^{-1} \end{pmatrix}\right\},$$

and α is such that $\alpha + \alpha^{-1} = \mathrm{Tr}(g)$. Given $y_1 \in H^{(2)}$ and $\mathbf{B}$ containing y_1, we have by (6.38)

$$y_2 \in (H^{(2)} \cap \mathbf{U}) \cup (H^{(2)} \cap t^2\mathbf{U})$$

for some $t \in C_{\mathbf{B}}$. We note that $t^2\mathbf{U}$ is itself conjugate to C_{α^2} or $C_{\alpha^{-2}}$.

Then the size of the fiber $\phi^{-1}(y_1, y_2) \cap Z$ is determined by the number of possibilities for x_1. As the latter satisfies

$$x_1 \in C_{\mathbf{B}} \cap H,$$

we see that we should attempt to bound from above

$$H^{(k)} \cap C_\gamma$$

for a fixed k and a fixed $\gamma \in \mathbf{F}_p^\times$, as this will lead us to estimates for the number of possibilities for y_2 as well as x_1. Using Lemma 6.7.14 below, we get indeed

$$|\{y_2 \mid (y_1, y_2) \in W_2\}| \leqslant 8\mathrm{trp}(H^{(2)})^2 |H^{(2)}|^{1/3} \ll \alpha^{25/3}|H|^{1/3},$$

(the factor 8 accounts for the two possible choices of $\mathbf{B}$ and the two "components" for y_2, and the factor 2 in the lemma) and

$$|\phi^{-1}(y_1, y_2) \cap Z| \ll \alpha^2 |H|^{1/3}.$$

This gives

$$S_2 \ll \alpha^{31/3}|H|^{2/3}|H^{(2)}| \ll \alpha^{34/3}|H|^{5/3},$$

as claimed. □

Lemma 6.7.14. *For any prime $p \geqslant 3$, any $\gamma \in \mathbf{F}_p^\times$, any $x \in \mathrm{SL}_2(\bar{\mathbf{F}}_p)$ and any symmetric generating set H of $\mathrm{SL}_2(\mathbf{F}_p)$ containing* 1*, we have*

$$|H \cap xC_\gamma x^{-1}| = \Bigl|H \cap x\Bigl\{\begin{pmatrix}\gamma & t\\ 0 & \gamma^{-1}\end{pmatrix} \mid t \in \mathbf{F}_p\Bigr\}x^{-1}\Bigr| \leqslant 2\alpha^2|H|^{1/3}$$

where $\alpha = \mathrm{trp}(H)$.

Proof. We first need to deal with the fact that x and γ are not necessarily in $\mathrm{SL}_2(\mathbf{F}_p)$ (since we used Lemma 6.7.11, which refers to algebraically closed field – we will see in the proof that it brings helpful simplifications). We have $xC_\gamma x^{-1} \cap \mathrm{SL}_2(\mathbf{F}_p) \subset x\mathbf{B}_0x^{-1} \cap \mathrm{SL}_2(\mathbf{F}_p)$, and there are three possibilities for the latter: either $x\mathbf{B}_0x^{-1} \cap \mathrm{SL}_2(\mathbf{F}_p) = \{\pm 1\}$, or $x\mathbf{B}_0x^{-1} \cap \mathrm{SL}_2(\mathbf{F}_p) = T$ is a maximal torus of $\mathrm{SL}_2(\mathbf{F}_p)$, or $x\mathbf{B}_0x^{-1} \cap \mathrm{SL}_2(\mathbf{F}_p) = B$ is an $\mathrm{SL}_2(\mathbf{F}_p)$-conjugate of the group $B_0 = \mathbf{B}_0 \cap \mathrm{SL}_2(\mathbf{F}_p)$ of upper-triangular matrices (see Lemma 6.6.13). In this last case, we can assume that $x \in \mathrm{SL}_2(\mathbf{F}_p)$ and $\gamma \in \mathbf{F}_p$. In the first, of course, there is nothing to do. And as for the second, note that γ and γ^{-1} are the eigenvalues of any element in $\mathrm{SL}_2(\mathbf{F}_p) \cap xC_\gamma x^{-1}$, and there are at most two elements in a maximal torus with given eigenvalues. A fortiori, we have $|H \cap xC_\gamma x^{-1}| \leqslant 2 \leqslant 2\alpha^2|H|^{1/3}$ in that case.

Thus we are left with the situation where $x \in \mathrm{SL}_2(\mathbf{F}_p)$. Using $\mathrm{SL}_2(\mathbf{F}_p)$-conjugation, it is enough to deal with the case $x = 1$. Then either the intersection is empty (and the result true) or we can fix

$$g_0 = \begin{pmatrix}\gamma & t_0\\ 0 & \gamma^{-1}\end{pmatrix} \in H \cap C_\gamma,$$

and observe that for any $g \in H \cap C_\gamma$, we have

$$g_0^{-1}g \in H^{(2)} \cap C_1,$$

hence

$$|H \cap C_\gamma| \leqslant |H^{(2)} \cap C_1| = |H^{(2)} \cap \mathbf{U}_0|,$$

which reduces further to the case $\gamma = 1$.

In that case we have another case of the Larsen-Pink inequalities, and in fact precisely the situation we considered in our first motivating example of growth, Proposition 6.6.3. Applying the latter to $H^{(2)}$, we get

$$|H^{(2)} \cap \mathbf{U}_0| \leqslant 2|H^{(8)}|^{1/3} \leqslant 2\alpha^2|H|^{1/3}$$

by Ruzsa's Lemma (the statement of the proposition suggests $H^{(10)}$ instead, but looking at the proof, we see that we can define (6.22) using an auxiliary element $h \in H$ instead of $H^{(2)}$, so that the image is in $H^{(8)}$.)

□

We must still prove Lemma 6.7.11.

Proof of Lemma 6.7.11. This computation is based on a list of simple checks. We can assume that the regular semisimple element g is

$$g = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha^{-1} \end{pmatrix}$$

where $\alpha^4 \neq 1$, because $\alpha = \pm 1$ implies that g is not regular semisimple, and α a fourth root of unity implies that $\mathrm{Tr}(g) = 0$, which is the third case of the lemma. (Note that here we use the fact that k is assumed to be algebraically closed!) Thus the conjugacy class is the set of matrices of trace equal to $t = \alpha + \alpha^{-1}$.

The only trick involved is that, for any $y_1 \in \mathrm{SL}_2(k)$ and $x \in \mathrm{SL}_2(k)$, we have

$$C \cap (xy_1x^{-1})^{-1}C = x(x^{-1}C \cap y_1^{-1}x^{-1}C) = x(C \cap y_1^{-1}C)x^{-1}$$

since $x^{-1}C = Cx^{-1}$, by definition of conjugacy classes. This means we can compute $C \cap y_1^{-1}C$, up to conjugation, by looking at $C \cap (y_1')^{-1}C$ for any y_1' in the conjugacy class of y_1. In particular, of course, determining whether $C \cap y_1^{-1}C$ is infinite or not only depends on the conjugacy class of y_1...

The conjugacy classes in $\mathrm{SL}_2(k)$ are well-known (see Proposition B.2.4 in Appendix B). We will run through representatives of these classes in order, and determine the corresponding intersection $C \cap y_1^{-1}C$. Then, to compute $C \cap y_1^{-1}C \cap y_2^{-1}C$, we take an element x in $C \cap y_1^{-1}C$, compute y_2x, and $C \cap y_1^{-1}C \cap y_2^{-1}C$ corresponds to those x for which the trace of y_2x is also equal to t...

We assume $y_1 \neq \pm 1$. Then we distinguish four cases:

$$\begin{gathered} y_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad y_1 = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}, \\ y_1 = \begin{pmatrix} \beta & 0 \\ 0 & \beta^{-1} \end{pmatrix}, \quad \beta \neq \pm 1, \beta \neq \alpha^{\pm 2} \\ y_1 = \begin{pmatrix} \alpha^2 & 0 \\ 0 & \alpha^{-2} \end{pmatrix}. \end{gathered} \tag{6.40}$$

We claim that $D = C \cap y_1^{-1}C$ is then given, respectively, by the sets containing all matrices of the following forms, parameterized by an element $a \in k$ (with $a \neq 0$ in the third case):

$$\begin{gathered} \begin{pmatrix} \alpha & a \\ 0 & \alpha^{-1} \end{pmatrix} \text{ or } \begin{pmatrix} \alpha^{-1} & a \\ 0 & \alpha \end{pmatrix}, \\ \begin{pmatrix} a & (-a^2 + at - 1)/(2t) \\ 2t & t - a \end{pmatrix}, \end{gathered} \tag{6.41}$$

$$\frac{1}{\beta+1}\begin{pmatrix} t & (\beta-\alpha^2)a \\ -(\beta-\alpha^{-2})a^{-1} & t\beta \end{pmatrix}, \tag{6.42}$$

$$\begin{pmatrix} \alpha^{-1} & a \\ 0 & \alpha \end{pmatrix} \text{ or } \begin{pmatrix} \alpha^{-1} & 0 \\ a & \alpha \end{pmatrix}. \tag{6.43}$$

Let us check, for instance, the third and fourth cases (cases (1) and (2) are left as exercise...), which we can do simultaneously, taking y_1 as in (6.40) but without assuming $\beta \neq \alpha^{\pm 2}$. For

$$x = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in C,$$

we compute

$$y_1 x = \begin{pmatrix} \beta a & \beta b \\ \beta^{-1}c & \beta^{-1}d. \end{pmatrix}$$

This matrix belongs to C if and only if $\beta a + \beta^{-1}d = t = a + d$. This means that (a, d) is a solution of the linear system

$$\begin{cases} a + d = t \\ \beta a + \beta^{-1}d = t, \end{cases}$$

of determinant $\beta^{-1} - \beta \neq 0$, so that we have

$$a = \frac{t}{\beta+1}, \qquad d = \frac{\beta t}{\beta+1}.$$

Write $c = c'/(\beta+1)$, $d = d'/(\beta+1)$; then the condition on c' and d' to have $\det(x) = 1$ can be expressed as

$$-c'd' = (\beta-\alpha^2)(\beta-\alpha^{-2}).$$

This means that either β is not one of α^2, α^{-2} (the third case), and then c and d are non-zero, and we can parameterize the solutions as in (6.42), or else (the fourth case) c or d must be zero, and then we get upper or lower-triangular matrices, as described in (6.43).

Now we intersect D (in the general case again) with $y_2^{-1}C$. We write

$$y_2 = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix}.$$

We consider the first of our four possibilities now, so that $x \in D$ is upper-triangular with diagonal coefficients α, α^{-1} (as a set), see (6.41). We compute the trace of $y_2 x$, and find that is

$$ax_3 + x_1\alpha + x_4\alpha^{-1}, \text{ or } ax_3 + x_1\alpha^{-1} + x_4\alpha.$$

Thus, if $x_3 \neq 0$, there is at most one value of a for which the trace is t, i.e., $D \cap y_2^{-1}C$ has at most two elements (one for each form of the diagonal). If $x_3 = 0$, we find that x_1 is a solution of

$$\alpha x_1 + \alpha^{-1}x^{-1} = t,$$

or

$$\alpha x_1^{-1} + \alpha^{-1}x_1 = t,$$

for which the solutions are among 1, α^2 and α^{-2}, so that y_2 is upper-triangular with diagonal coefficients $(1, 1)$, (α^2, α^{-2}) or (α^{-2}, α^2), and this is one of the instances of Case (2) of Lemma 6.7.11.

Let us now consider the second of our four cases, leaving this time the third and fourth to the reader. Thus we take x as in (6.42), and compute the trace of y_2x as a function of a, which gives

$$\mathrm{Tr}(y_2x) = -\frac{x_3}{2t}a^2 + \Bigl(x_1 - x_4 + \frac{x_3}{2}\Bigr)a + (x_4 + 2x_2)t.$$

The equation $\mathrm{Tr}(y_2x) = t$ has therefore at most two solutions, unless $x_3 = 0$ and $x_4 = x_1$. In that case we have $x_4 = \pm 1$, and the constant term is equal to t if and only if $x_4 = 1$ and $x_2 = 0$ (so $y_2 = 1$) or $x_4 = \pm 1$ and $x_2 = 1$ (and then $y_2 = y_1$). Each of these possibilities corresponds to the exceptional situation of Case (1) of Lemma 6.7.11.

All in all, going through the remaining situations, we finish the proof... □

APPENDIX A

Explicit multiplicative combinatorics

A.1. Introduction

In this appendix, we will state and prove some basic results about product sets and approximate subgroups of finite groups, where the main emphasis is to obtain *explicit* forms of the estimates. The basic structure of the results is found for instance in Tao's paper [108], which we follow closely (keeping track of the constants). For simplicity, we work only with symmetric sets, usually containing the identity.

Below all sets are subsets of a fixed finite group G, and are all non-empty. The notation AB, for subsets A, $B \subset G$, refer to the product sets

$$AB = \{ab \mid a \in A,\ b \in B\},$$

and we use again the notation $A^{(n)}$ for the n-fold product set $A \cdot A \cdots A$. We extend it to negative values of n, defining $A^{(-n)} = A^{-1(n)}$, where $A^{-1} = \{a \mid a^{-1} \in A\}$. We also recall the notation $E(A, B)$ and $e(A, B)$ for the multiplicative energy (Definition 6.3.3), and we now define the Ruzsa distance:

Definition A.1.1 (Ruzsa distance). Let G be a finite group and A, B non-empty subsets of G. The *Ruzsa distance* $d(A, B)$ is defined by

$$d(A, B) = \log\Bigl(\frac{|A \cdot B^{-1}|}{\sqrt{|A||B|}}\Bigr).$$

The crucial property of the Ruzsa distance is the triangle inequality:

Lemma A.1.2 (Ruzsa triangle inequality). *Let G be a finite group, A, B and C non-empty subsets of G. We have*

$$d(A, B) \leqslant d(A, C) + d(C, B).$$

Proof. Spelling out the meaning of this inequality, we see that it is equivalent with

$$|A \cdot B^{-1}| \leqslant \frac{|A \cdot C^{-1}||C \cdot B^{-1}|}{|C|},$$

which one proves by constructing an injective map

$$i : C \times (A \cdot B^{-1}) \longrightarrow (A \cdot C^{-1}) \times (C \cdot B^{-1}),$$

as follows: for any element $x \in A \cdot B^{-1}$, fix (once and for all!) some elements $a(x) \in A$ and $b(x) \in B$ such that

$$x = a(x)b(x)^{-1}.$$

Then we define

$$i(c, x) = (a(x)c^{-1}, cb(x)^{-1})$$

for all $(c, x) \in A \times (A \cdot B^{-1})$. To show that i is injective we observe that $i(c, x) = i(d, y)$ means that

$$\begin{cases} a(x)c^{-1} = a(y)d^{-1} \\ cb(x)^{-1} = db(y)^{-1} \end{cases}$$

and if we take the product of these two equalities, we derive

$$x = a(x)b(x)^{-1} = a(y)b(y)^{-1} = y,$$

and then furthermore $c = db(y)^{-1}b(x) = d$, so that i is indeed injective. □

We will also use the following further elementary properties of the multiplicative energy.

Lemma A.1.3. *The following properties hold:*
(1) *We have*

$$E(A, B) = \sum_{x\in G} |A \cap xB^{-1}|^2.$$

(2) *We have* $E(A, B) \geqslant |A|^2|B|^2/|AB|$.
(3) *We have* $E(A, A^{-1}) = E(A^{-1}, A)$.

Proof. (1) follows from the definition by writing

$$E(A, B) = \sum_{\substack{(a,a',b,b')\in A^2\times B^2\\ ab=a'b'}} \quad = \sum_{x\in G}\Big(\sum_{\substack{a,b\in A\times B\\ ab=x}} 1\Big)^2.$$

Then we obtain (2) using the Cauchy-Schwarz inequality and the fact that $A \cap xB^{-1}$ is empty unless $x \in AB$:

$$|A||B| = \sum_{x\in G} |A \cap xB^{-1}| = \sum_{x\in AB} |A \cap xB^{-1}| \leqslant \Big(\sum_{x} |A \cap xB^{-1}|^2\Big)^{1/2} |AB|^{1/2}.$$

Finally, to prove (3) we observe that for $(a_1, a_2, a_3, a_4) \in A^4$, the equations

$$a_1a_2^{-1} = a_3a_4^{-1}, \quad \text{and} \quad a_2^{-1}a_4 = a_1^{-1}a_3$$

are equivalent. The number of solutions of the first is $E(A, A^{-1})$, and of the second is $E(A^{-1}, A)$. □

A.2. Diagrams

We will use the following diagrammatic conventions, which allow us to keep track of constants:

(1) If A and B are sets with $d(A, B) \leqslant \log\alpha$, we write

$$\begin{tikzcd} A \arrow[r, {Circle}-{Circle}, "\alpha"'] & B \end{tikzcd},$$

(2) If A and B are sets with $|B| \leqslant \alpha|A|$, we write

$$\begin{tikzcd} B \arrow[r, {Circle}->, "\alpha"] & A \end{tikzcd},$$

and in particular if $|X| \leqslant \alpha$, we write

$$\begin{tikzcd} X \arrow[r, {Circle}->, "\alpha"] & 1 \end{tikzcd},$$

(3) If A and B are sets with $e(A, B) \geqslant 1/\alpha$, we write

$$\begin{tikzcd} A \arrow[r, {Circle}-{Circle}, squiggly, "\alpha"] & B \end{tikzcd},$$

(4) If $A \subset B$, we write

$$\begin{tikzcd} A \arrow[r, tail] & B \end{tikzcd}.$$

The following rules are easy to check (in addition to some more obvious ones which we do not spell out):

(1) From

$$\begin{tikzcd} A \arrow[r, *-*, "\alpha"'] & B \end{tikzcd}$$

we can get

$$\begin{tikzcd} A \arrow[r, *->, "\alpha^2"] & B \end{tikzcd}\ , \qquad \begin{tikzcd} B \arrow[r, *->, "\alpha^2"] & A \end{tikzcd}\ .$$

(2) (Ruzsa's triangle inequality) From

$$\begin{tikzcd} A \arrow[r, *-*, "\alpha_1"'] & B \arrow[r, *-*, "\alpha_2"'] & C \end{tikzcd}$$

we get

$$\begin{tikzcd} A \arrow[r, *-*, "\alpha_1\alpha_2"'] & C \end{tikzcd}\ .$$

(3) From

$$\begin{tikzcd} C \arrow[r, *->, "\alpha_1"] & B \arrow[r, *->, "\alpha_2"] & A \end{tikzcd}$$

we get

$$\begin{tikzcd} C \arrow[r, *->, "\alpha_1\alpha_2"] & A \end{tikzcd}\ .$$

(4) ("Unfolding edges") From

$$\begin{tikzcd} B \arrow[r, *->, "\alpha"] \arrow[r, *-*, bend right, "\beta"'] & A \end{tikzcd}$$

we get

$$\begin{tikzcd} AB^{-1} \arrow[r, *->, "\sqrt{\alpha}\beta"] & A \end{tikzcd}$$

(note that by the second point in this list, we only need to have

$$\begin{tikzcd} B \arrow[r, *-*, "\beta"'] & A \end{tikzcd}$$

to obtain the full statement with $\alpha = \beta^2$, which is usually qualitatively equivalent.)

(5) ("Folding") From

$$\begin{tikzcd} AB^{-1} \arrow[r, *->, "\alpha"] & A \arrow[r, *->, "\beta"] & B \end{tikzcd}$$

we get

$$\begin{tikzcd} A \arrow[r, *-*, "\alpha\beta^{1/2}"'] & B \end{tikzcd}\ .$$

Note that the relation $A \overset{\alpha}{\bullet\!\!\longrightarrow} B$ is purely a matter of the size of A and B, while the other arrow types depend on structural relations involving the sets (for $A \rightarrowtail B$) and product sets (for $A \underset{\alpha}{\bullet\!\!-\!\!\!-\!\!\bullet} B$ or $A \overset{\alpha}{\bullet\!\!\leadsto\!\!\bullet} B$).

A.3. Statements and proofs

The main result that we use in Chapter 6 is Theorem A.3.7 below. We present the arguments leading to the proof, following Tao's arguments [108, §3,4,5].

THEOREM A.3.1 (Ruzsa covering lemma). *If*

$$\begin{tikzcd} AB \arrow[r, *->, "\alpha"] & A \end{tikzcd}\ ,$$

there exists a set X which satisfies

$$\begin{tikzcd} X \arrow[r, tail] & B \end{tikzcd}\ , \qquad \begin{tikzcd} X \arrow[r, *->, "\alpha"] & 1 \end{tikzcd}\ , \qquad \begin{tikzcd} B \arrow[r, tail] & A^{-1}AX \end{tikzcd}\ ,$$

and symmetrically, if

$$BA \bullet\!\!\xrightarrow{\alpha} A \ ,$$

there exists Y with

$$Y \rightarrowtail B \ , \quad Y \bullet\!\!\xrightarrow{\alpha} 1 \ , \quad B \rightarrowtail XAA^{-1} \ .$$

Proof. Let $X \subset B$ be a subset that is maximal under inclusion and such that the "cosets" $A \cdot x \subset G$ are disjoint. We have $X \bullet\!\!\xrightarrow{\alpha} 1$, because $AB \bullet\!\!\xrightarrow{\alpha} A$. For any $b \in B$, the set $A \cdot b$ cannot be disjoint from $A \cdot X$, by maximality. This means that $b \in A^{-1} \cdot A \cdot X$. The other case is obtained similarly. □

Next is another result which is essentially due to Ruzsa: the tripling constant of a symmetric set controls all other n-fold product sets. This was stated and proved as Proposition 6.6.5 in Chapter 6, but we state it again in our diagrammatic language.

Theorem A.3.2 (Ruzsa's Lemma). *If A is symmetric and*

$$A^{(3)} \bullet\!\!\xrightarrow{\alpha} A$$

then we have

$$A^{(n)} \bullet\!\!\xrightarrow{\quad\alpha^{n-2}\quad} A$$

for all $n \geqslant 3$. In particular, we get

$$A^{(7)} \bullet\!\!\xrightarrow{\alpha^5} A \ .$$

Petridis [94, Th. 1.6] and Tao [108, Lemma 3.4] (for instance) show that there are also versions of this result with A^n replaced by any n-fold product of factors equal to A or A^{-1}. But we will only use symmetric subsets, in which case the above has better constants.

Theorem A.3.3. *Let $A = A^{-1}$ with $1 \in A$ and*

$$A^{(3)} \bullet\!\!\xrightarrow{\alpha} A \ .$$

Then $H = A^{(3)}$ is a $(2\alpha^{44})$-approximate subgroup containing A.

Proof. We have first

$$H \bullet\!\!\xrightarrow{\alpha} A \ , \quad A \rightarrowtail H \ .$$

Then by Ruzsa's tripling inequality, we get

$$AH^{(2)} = A^{(7)} \bullet\!\!\xrightarrow{\alpha^5} A \ ,$$

and by the Ruzsa covering lemma there exists X with

$$X \rightarrowtail H^{(2)} \ , \quad X \bullet\!\!\xrightarrow{\alpha^5} 1 \ ,$$

such that

$$H^{(2)} \rightarrowtail A^{(2)}X \rightarrowtail A^{(3)}X = HX \ .$$

Taking $X_1 = X \cup X^{-1}$, we get

$$X_1 \rightarrowtail H^{(2)} \ , \quad X_1 \bullet\!\!\xrightarrow{2\alpha^5} 1 \ ,$$

and

$$H^{(2)} \rightarrowtail HX \ , \quad H^{(2)} \rightarrowtail XH \ ,$$

which are the properties defining a $(2\alpha^5)$-approximate subgroup. □

Theorem A.3.4. *Let A and B with*

$$A \bullet\!\!\underset{\alpha}{\rule[0.5ex]{2.5em}{0.4pt}}\!\!\bullet B^{-1}$$

Then there exists a γ-approximate subgroup H and a set X with

$$X \bullet\!\!\xrightarrow{\gamma_1} 1\ , \qquad A \rightarrowtail XH\ , \qquad B \rightarrowtail HX\ , \qquad H \bullet\!\!\xrightarrow{\gamma_2} A\ ,$$

where

$$\gamma \leqslant 2^{21}\alpha^{80}, \qquad \gamma_1 \leqslant 2^{28}\alpha^{104}, \qquad \gamma_2 \leqslant 8\alpha^{14}.$$

Furthermore, one can ensure that

$$H^{(3)} \bullet\!\!\xrightarrow{2^{10}\alpha^{40}} H\ . \tag{A.1}$$

This will require first a lemma.

Lemma A.3.5. *Let A be such that*

$$AA^{-1} \bullet\!\!\xrightarrow{\alpha} A\ .$$

The set

$$S = \{x \in G \mid\ A \bullet\!\!\xrightarrow{2\alpha} A \cap Ax\ \}.$$

is symmetric, it contains 1, and satisfies

$$A \bullet\!\!\xrightarrow{2\alpha} S\ , \qquad AS^{(n)}A^{-1} \bullet\!\!\xrightarrow{2^n\alpha^{2n+1}} A$$

for all $n \geqslant 1$.

Proof. It is elementary that S contains 1 and is symmetric. To prove the lower bound for the size of S, we consider $I_x = |A \cap Ax|$ for $x \in G$. We will determine the average of I_x and get a lower bound for the second moment to deduce that I_x is often enough large enough. First

$$\sum_x I_x = \sum_x \sum_{\substack{a \in A \\ ax^{-1} \in A}} 1 = \sum_{a \in A} \sum_{\substack{x \in G \\ ax^{-1} \in A}} = |A|^2,$$

and next by Lemma A.1.3 (2), (3) and (1), we obtain

$$\sum_x I_x^2 = E(A^{-1}, A) = E(A, A^{-1}) \geqslant \frac{|A|^4}{|A \cdot A^{-1}|} \geqslant \frac{|A|^3}{\alpha}.$$

Since

$$\sum_{x \notin S} I_x^2 \leqslant \Big(\max_{x \notin S} I_x\Big) \sum_x I_x \leqslant \frac{|A|^3}{2\alpha},$$

it follows that

$$\frac{|A|^3}{2\alpha} \leqslant \sum_{x \in S} I_x^2 \leqslant |A|^2|S|,$$

which gives the desired lower bound $A \bullet\!\!\xrightarrow{2\alpha} S$.

Fix now an integer $n \geqslant 1$, and let $B = A \cdot S^{(n)} \cdot A^{-1}$. Let N_n be the number of $(n+1)$-tuples $(x_0, \ldots, x_n)$ of elements of $A \cdot A^{-1}$ such that the product $x_0 \cdots x_n$ belongs to B. We can write

$$N_n = \sum_{y \in B} M_y$$

where M_y is the number of $(x_0, \ldots, x_{n-1}) \in (A \cdot A^{-1})^n$ such that $(x_0 \cdots x_{n-1})^{-1} y \in A \cdot A^{-1}$.

For $y \in B$, we can write $y = a_0 s_1 \cdots s_n b_{n+1}^{-1}$ where $a_0 \in A$, the s_i belong to S and $b_{n+1} \in A$. Thus

$$(x_0 \cdots x_{n-1})^{-1} y = x_{n-1}^{-1} \cdots x_0^{-1} a_0 s_1 \cdots s_n b_{n+1}^{-1}.$$

We introduce elements b_1, …, b_n, such that

$$x_0^{-1} a_0 = b_1, \quad x_1^{-1} b_1 s_1 = b_2, \quad \dots, \quad x_{n-1}^{-1} b_{n-1} s_{n-1} = b_n,$$

noting that $(x_0, \dots, x_{n-1}) \mapsto (b_1, \dots, b_n)$ is bijective. Then $(x_0 \cdots x_{n-1})^{-1} y = b_n s_n b_{n+1}^{-1}$, and M_y is therefore the number of tuples $(b_1, \dots, b_n) \in G^n$ such that $a_0 b_1^{-1} \in A \cdot A^{-1}$ and $b_i s_i b_{i+1}^{-1} \in A \cdot A^{-1}$ for $1 \leqslant i \leqslant n$. This number is at least $I_{s_1} \cdots I_{s_n}$, since the tuples $(b_1, \dots, b_n)$ with $b_i \in A \cap A s_i$ satisfy the conditions. This means that for $y \in B$, we have $M_y \geqslant (2\alpha)^{-n} |A|^n$, and $N_n \geqslant (2\alpha)^{-n} |A|^n |B|$. Comparing with the elementary upper bound $N_n \leqslant |A \cdot A^{-1}|^{n+1} \leqslant \alpha^{n+1} |A|^{n+1}$, we obtain

$$|B| \leqslant (2\alpha)^n \alpha^{n+1} |A|.$$

□

Proof of Theorem A.3.4. From

$$\begin{tikzcd} A \arrow[r, *->, "1"] \arrow[r, *-*, bend right, "\alpha^2"'] & A \end{tikzcd} ,$$

we get first

$$\begin{tikzcd} AA^{-1} \arrow[r, *->, "\alpha^2"] & A \end{tikzcd} .$$

By the lemma above, there exists a symmetric set S with $1 \in S$ such that

(A.2) $$\begin{tikzcd} A \arrow[r, *->, "2\alpha^2"] & S \end{tikzcd} , \qquad \begin{tikzcd} AS^{(n)}A^{-1} \arrow[r, *->, "2^n\alpha^{4n+2}"] & A \end{tikzcd}$$

for all $n \geqslant 1$.

Taking $n = 1$, we get first

$$AS^{-1} = \begin{tikzcd} AS \arrow[r, *->, "2\alpha^6"] & A \end{tikzcd}$$

and thus

$$\begin{tikzcd} AS^{-1} \arrow[r, *->, "2\alpha^6"] & A \arrow[r, *->, "2\alpha^2"] & S \end{tikzcd} ,$$

which gives

$$\begin{tikzcd} A \arrow[r, *-*, "\beta"'] & S \end{tikzcd}$$

by folding, with $\beta = 2\sqrt{2}\alpha^7$.

In addition, we have

$$\begin{tikzcd} S^{(3)} \arrow[r, *->, "8\alpha^{14}"] & A \arrow[r, *->, "2\alpha^2"] & S \end{tikzcd} ,$$

and Theorem A.3.3 says that $H = S^{(3)}$ is a γ-approximate subgroup containing S, with $\gamma = 2(16\alpha^{16})^5 = 2^{21}\alpha^{80}$, and (as we see)

$$\begin{tikzcd} H \arrow[r, *->, "8\alpha^{14}"] & A \end{tikzcd} .$$

Moreover, once more from (A.2) with $n = 3$, we deduce

$$H^{(3)} = \begin{tikzcd} S^{(9)} \arrow[r, tail] & AS^{(9)}A^{-1} \arrow[r, *->, "2^9\alpha^{38}"] & A \arrow[r, *->, "2\alpha^2"] & S \end{tikzcd} ,$$

which gives (A.1).

Now from

$$AH = AS^{(3)} \bullet\!\xrightarrow{\quad 8\alpha^{14} \quad} A \bullet\!\xrightarrow{\quad 2\alpha^{2} \quad} S \bullet\!\xrightarrow{1} H \ ,$$

we see by the Ruzsa covering lemma that there exists Y with

$$Y \rightarrowtail A \ , \quad Y \bullet\!\xrightarrow{\quad 16\alpha^{16} \quad} 1 \ , \quad A \rightarrowtail YHH \ .$$

By definition of an approximate subgroup, there exists Z with

$$Z \bullet\!\xrightarrow{\gamma} 1 \ , \quad HH \rightarrowtail ZH \ ,$$

and hence

$$A \rightarrowtail (YZ)H \ .$$

Now we go towards B. First we have

$$AH^{-1} = AS^{(3)} \bullet\!\xrightarrow{\quad 8\alpha^{14} \quad} A \bullet\!\xrightarrow{\alpha^{2}} H$$

which, again by folding, gives

$$A \underset{\alpha_1}{\bullet\!\!-\!\!\!-\!\!\!-\!\!\bullet} H$$

with $\alpha_1 = 8\sqrt{2}\alpha^{15}$. Hence we can write

$$H \underset{\alpha_1}{\bullet\!\!-\!\!\!-\!\!\!-\!\!\bullet} A \underset{\alpha}{\bullet\!\!-\!\!\!-\!\!\!-\!\!\bullet} B^{-1} \ ,$$

and so

$$H \underset{\alpha\alpha_1}{\bullet\!\!-\!\!\!-\!\!\!-\!\!\!-\!\!\!-\!\!\!-\!\!\bullet} B^{-1} \ .$$

In addition, from (A.2) with $n = 3$, we deduce

$$H \bullet\!\xrightarrow{\quad 8\alpha^{14} \quad} A \bullet\!\xrightarrow{\alpha^{2}} B^{-1} \ ,$$

and therefore we get

$$\begin{tikzcd} H \arrow[r, "8\alpha^{16}"] \arrow[r, dash, bend right, "\alpha\alpha_1"'] & B^{-1}, \end{tikzcd}$$

from which it follows by unfolding that

$$B^{-1}H^{-1} = B^{-1}H \bullet\!\xrightarrow{\quad 32\alpha^{20} \quad} B^{-1} \bullet\!\xrightarrow{\alpha^{2}} A \bullet\!\xrightarrow{2\alpha^{2}} H \ .$$

Once more by the Ruzsa covering lemma, we find Y_1 with

$$Y_1 \rightarrowtail B^{-1} \ , \quad Y_1 \bullet\!\xrightarrow{\quad 32\alpha^{20} \quad} 1 \ , \quad B^{-1} \rightarrowtail Y_1HH \rightarrowtail (Y_1Z)H \ .$$

Now we need only take $X = (Y_1Z \cup YZ)$, so that

$$X \bullet\!\xrightarrow{\gamma_1} 1$$

with $\gamma_1 = \gamma(64\alpha^{24} + 16\alpha^{16})$, in order to conclude. Since

$$\gamma_1 \leqslant 2^{28}\alpha^{104},$$

we are done. □

The next result is a version of the Balog-Gowers-Szemerédi Theorem.

Theorem A.3.6 (Balog-Gowers-Szemerédi). *Let A and B with*

$$\begin{tikzcd} A \bullet \arrow[r, squiggly, no head, "\alpha"] & \bullet B \end{tikzcd} .$$

Then there exist A_1, B_1 with

$$\begin{tikzcd} A_1 \arrow[r, tail] & A , & B_1 \arrow[r, tail] & B , \end{tikzcd}$$

as well as

$$\begin{tikzcd} A \bullet \arrow[r, "8\sqrt{2}\alpha"] & A_1 , & B \bullet \arrow[r, "8\alpha"] & B_1 , \end{tikzcd}$$

and

$$\begin{tikzcd} A_1 \bullet \arrow[r, no head, "\alpha_1"'] & \bullet B_1^{-1} \end{tikzcd}$$

where $\alpha_1 = 2^{20}\alpha^9$.

The proof will use Proposition 2.2.17.

Proof. For $x \in G$, define $I_x = |A \cap xB^{-1}|$. Note that $I_x = 0$ unless $x \in A \cdot B$, and that

$$I_x \leqslant \sqrt{|A||xB^{-1}|} = \sqrt{|A||B|} \tag{A.3}$$

by the Cauchy-Schwarz inequality.

Let S be the set

$$S = \{x \in G \mid I_x \geqslant (2\alpha)^{-1}|A|^{1/2}|B|^{1/2}\} \subset AB.$$

Since

$$|S|\frac{|A|^{1/2}|B|^{1/2}}{2\alpha} \leqslant \sum_x I_x \leqslant |A||B|,$$

we have $|S| \leqslant 2\alpha|A|^{1/2}|B|^{1/2}$. We then let

$$T = \{(a,b) \in A \times B \mid ab \in S\}.$$

We will show that

$$|T| \geqslant \frac{|A||B|}{2\alpha}. \tag{A.4}$$

Indeed, we have

$$|T| = \sum_{x\in S} |\{(a,b) \in A \times B \mid ab \in S\}| = \sum_{x \in S} I_x \leqslant \frac{1}{|A|^{1/2}|B|^{1/2}} \sum_{x\in S} I_x^2$$

by (A.3). But using Lemma A.1.3 (1) and the assumption on the multiplicative energy of A and B, we have

$$\sum_{x\in S} I_x^2 \geqslant \sum_x I_x^2 - |A||B|\frac{|A|^{1/2}|B|^{1/2}}{2\alpha} = E(A,B) - \frac{|A|^{3/2}|B|^{3/2}}{2\alpha} \geqslant \frac{|A|^{3/2}|B|^{3/2}}{2\alpha},$$

which combined with the previous inequality gives (A.4).

We now define a simple bipartite graph Γ with vertex set the disjoint union of A and B and edges the pairs $(a,b) \in T \subset A \times B$ (with endpoints $\{a,b\}$). By Proposition 2.2.17, there exist $A_1 \subset A$ and $B_1 \subset B$, with

$$\begin{tikzcd} A \bullet \arrow[r, "8\sqrt{2}\alpha"] & A_1 , & B \bullet \arrow[r, "8\alpha"] & B_1 , \end{tikzcd} \tag{A.5}$$

such that for any pair $(a,b)\in A_1\times B_1$, there are at least $2^{-12}\alpha^{-4}|A||B|$ paths of length 3 in Γ from a to b. Let $b'\in B$ and $a'\in A$ be intermediate vertices in such a path

$$a\sim b'\sim a'\sim b.$$

By definition, the elements ab', $a'b'$, $a'b$ belong to S. The identity $ab=(ab')(a'b')^{-1}a'b$ shows then that there are $\geqslant 2^{-12}\alpha^{-4}|A||B|$ triples (x,y,z) in S such that $xy^{-1}z=ab$. This implies that

$$|A_1B_1|\leqslant \frac{|S|^3}{2^{-12}\alpha^{-4}|A||B|}\leqslant 2^{15}\alpha^7|A|^{1/2}|B|^{1/2},$$

and combining with (A.5), we obtain the bound for the Ruzsa distance. □

Now comes the main result of this section.

Theorem A.3.7. *Let A and B with*

$$\begin{tikzcd} A \arrow[r, squiggly, "\alpha", Circle-Circle] & B \end{tikzcd}.$$

Then there exist a β-approximate subgroup H and x, $y\in G$, such that

$$\begin{tikzcd} H \arrow[r, "\beta_2", Circle->] & A \end{tikzcd},\qquad \begin{tikzcd} A \arrow[r, "\beta_1", Circle->] & A\cap xH \end{tikzcd},\qquad \begin{tikzcd} B \arrow[r, "\beta_1", Circle->] & B\cap Hy \end{tikzcd},$$

where

$$\beta\leqslant 2^{1621}\alpha^{720},\qquad \beta_1\leqslant 2^{2112}\alpha^{937},\qquad \beta_2\leqslant 2^{283}\alpha^{126}.$$

Moreover, one can ensure that

$$\begin{tikzcd} H^{(3)} \arrow[r, "\beta_3", Circle->] & H \end{tikzcd}$$

where $\beta_3=2^{810}\alpha^{360}$.

Proof. By the Balog-Gowers-Szemerédi Theorem, we get A_1, B_1 with

$$\begin{tikzcd} A_1 \arrow[r, tail] & A \end{tikzcd},\qquad \begin{tikzcd} B_1 \arrow[r, tail] & B \end{tikzcd},$$

as well as

$$\begin{tikzcd} A \arrow[r, "8\sqrt{2}\alpha", Circle->] & A_1 \end{tikzcd},\qquad \begin{tikzcd} B \arrow[r, "8\alpha", Circle->] & B_1 \end{tikzcd},$$

and

$$\begin{tikzcd} A_1 \arrow[r, "\alpha_1"', Circle-Circle] & B_1^{-1} \end{tikzcd}$$

where $\alpha_1=2^{20}\alpha^9$. Applying Theorem A.3.4 to A_1 and B_1, we get a β-approximate subgroup H and a set X with

$$\begin{tikzcd} H \arrow[r, "2\alpha_1^{14}", Circle->] & A_1 \arrow[r, "1", Circle->] & A \end{tikzcd}$$

and

$$\begin{tikzcd} X \arrow[r, "\gamma", Circle->] & 1 \end{tikzcd},\qquad \begin{tikzcd} A_1 \arrow[r, tail] & XH \end{tikzcd},\qquad \begin{tikzcd} B_1 \arrow[r, tail] & HX \end{tikzcd},$$

where

$$\beta=2^{21}\alpha_1^{80}=2^{1621}\alpha^{720},\qquad \gamma=2^{28}\alpha_1^{104}=2^{2108}\alpha^{936},$$

and moreover

$$\begin{tikzcd} H^{(3)} \arrow[r, "\beta_3", Circle->] & H \end{tikzcd}$$

where $\beta_3=2^{10}\alpha_1^{40}=2^{810}\alpha^{360}$.

Applying the pigeonhole principle, we find x such that

$$\begin{tikzcd} A \arrow[r, "8\sqrt{2}\alpha", Circle->] & A_1 \arrow[r, "\gamma", Circle->] & A_1\cap xH \arrow[r, tail] & A\cap xH \end{tikzcd}$$

and y with

$$\begin{tikzcd} B \arrow[r, Circle->, "8\alpha"] & B_1 \arrow[r, Circle->, "\gamma"] & B_1 \cap Hy \arrow[r, tail] & B \cap Hy \ . \end{tikzcd}$$

This gives what we want with

$$\beta_1 \leqslant 8\sqrt{2}\alpha\gamma \leqslant 2^{2112}\alpha^{937}, \qquad \beta_2 = 8\alpha_1^{14} = 2^{283}\alpha^{126}.$$

□

APPENDIX B

Some group theory

B.1. Free groups

In Chapter 6, we use a few basic facts about free groups, which we quickly summarize here; we give a few proofs when it's possible in a few lines, simply to minimize the prerequisites. For a more detailed survey, we already mentioned the excellent book [50, Ch. II].

Proposition B.1.1. *Let G be a free group on a set S of $k \geqslant 1$ generators.*

(1) *Any subgroup of G is also a free group.*

(2) *Any $x \neq 1$ is of infinite order.*

(3) *If $x \neq 1$, then the centralizer of x is infinite cyclic.*

(4) *Fix a set $S = \{a_1^{\pm 1}, \ldots, a_k^{\pm 1}\}$ of free generators of G and their inverses and let $\Gamma = \mathcal{C}(G, S)$. If $x \neq 1$, then $d_\Gamma(1, x^n) \geqslant |n|$ for all $n \in \mathbf{Z}$.*

Proof. (1) This is the Nielsen-Schreier theorem, proofs of which can be found in many books (using either topological or algebraic means to reach the goal), e.g., [61, Cor. 2.9] or [10, p. 417, ex. 2].

(2) The subgroup generated by $x \in G$ is free by (1), and this implies that it is either trivial or infinite; if $x \neq 1$, the first possibility is of course excluded.

(3) The centralizer $H = C_G(x)$ is a free group by (1), and is infinite by (2) as it contains x^n for all $n \in \mathbf{Z}$. By definition, we have in H the relation $xyx^{-1}y^{-1} = 1$ for all $y \in H$, with $x \neq 1$, which implies that H can not be of rank $\geqslant 2$. So H is free of rank 1, i.e., it is infinite cyclic.

(4) Write $x = s_1 \cdots s_m$ as a reduced word in the generators S (see, e.g., [50, II.A]). Let $r \geqslant 0$ be the largest integer such that $y = s_1 \cdots s_r$ is the inverse of $s_{m-r+1} \cdots s_m$, so that we have

$$x = yx_1y^{-1}$$

where $x_1 = s_{r+1} \cdots s_{m-r}$. Note that $x_1 \neq 1$ since $x \neq 1$, and that $s_{r+1} \neq s_{m-r}^{-1}$ since otherwise r could be increased by 1. We have

$$x^n = yx_1^n y^{-1}$$

for $n \in \mathbf{Z}$. Assume $n \geqslant 0$; if we look at the corresponding word

$$(s_1 \cdots s_r)(s_{r+1} \cdots s_{m-r}) \cdots (s_{r+1} \cdots s_{m-r})(s_{m-r+1} \cdots s_m)$$

we see that no cancellation can occur, i.e., we have

$$d_\Gamma(1, x^n) = 2r + nd_\Gamma(1, x_1) \geqslant n.$$

For $n \leqslant 0$, we can argue similarly for x^{-1} to finish the proof. □

Remark B.1.2. Parts (2) and (3) can also be proved directly and elementarily, without using the Nielsen-Schreier Theorem (see, e.g., [61, §1.4, Cor. 1.2.2 and Prob. 2]).

We use the following example in Chapter 6.

PROPOSITION B.1.3 (The Lubotzky group is free but thin). (1) *For $k \geqslant 2$, let*

$$u_k = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}, \ v_k = \begin{pmatrix} 1 & 0 \\ k & 1 \end{pmatrix} \subset \mathrm{SL}_2(\mathbf{Z}).$$

Then u_k, v_k generate a free subgroup L_k of rank 2 in $\mathrm{SL}_2(\mathbf{Z})$. Moreover, for all $p \nmid k$, the image of $\{u_k, v_k\}$ modulo p generate $\mathrm{SL}_2(\mathbf{F}_p)$.

In particular the Lubotzky group $L = L_3$ of Example 6.4.7 is free of rank 2 and the Cayley graphs $\mathcal{C}(\mathrm{SL}_2(\mathbf{F}_p), \{u_3^{\pm 1}, v_3^{\pm 1}\})$ are connected for all primes $p \neq 3$.

(3) *For $k \geqslant 3$, the subgroup L_k has infinite index in $\mathrm{SL}_2(\mathbf{Z})$. For $k = 2$, the subgroup L_2 is of finite index in $\mathrm{SL}_2(\mathbf{Z})$.*

PROOF. (1) This is one of the simplest examples of the so-called "ping-pong" argument that is often used to produce free groups (see, e.g., [50, §II.B] for a more general discussion).

We consider the action of L_k on $\mathbf{R}^2$ by left multiplication, and introduce the two subsets

$$X_1 = \{(x, y) \in \mathbf{R}^2 \mid |x| > |y|\}, \quad X_2 = \{(x, y) \in \mathbf{R}^2 \mid |y| > |x|\}$$

of $\mathbf{R}^2$. If $(x, y) \in X_2$ and $m \in \mathbf{Z}$ we have

$$u_k^m(x, y) = (x + kmy, y)$$

which is in X_1 if $m \neq 0$, since

$$|x + kmy| \geqslant |kmy| - |x| = k|m||y| - |x| \geqslant (k-1)|m||y| \geqslant |y|$$

(note that $k \geqslant 2$ is important here). Thus we have $u_k^m(X_2) \subset X_1$ for all $m \neq 0$, and one checks similarly that $v_k^m(X_1) \subset X_2$ for $m \neq 0$.

Now we can start playing ping-pong to show that L_k is freely generated by u_k and v_k. First, let $\ell \geqslant 1$ be odd and let $n_1, \ldots, n_\ell$ be non-zero integers; consider the element

$$g = u_k^{n_1} v_k^{n_2} u_k^{n_3} \cdots v_k^{n_{\ell-1}} u_k^{n_\ell} \in L_k.$$

We must show that $g \neq 1$ (since it is a non-trivial reduced word in the generators); but indeed g is non-trivial since

$$g(X_2) = (u_k^{n_1} v_k^{n_2} u_k^{n_3} \cdots v_k^{n_{\ell-1}} u_k^{n_\ell})(X_2) \subset (u_k^{n_1} v_k^{n_2} u_k^{n_3} \cdots v_k^{n_{\ell-1}})(X_1) \subset \cdots \subset X_1,$$

and X_1 and X_2 are disjoint! Similarly, if a word begins and ends with a power of v_k, it is non-trivial in L_k, and if

$$g = u_k^{n_1} v_k^{n_2} u_k^{n_3} \cdots u_k^{n_{\ell-1}} v_k^{n_\ell},$$

the conjugate element

$$v_k^{-n_\ell} g v_k^{n_\ell} = v_k^{-n_\ell} u_k^{n_1} v_k^{n_2} u_k^{n_3} \cdots u_k^{n_{\ell-1}} v_k^{2n_\ell}$$

is non-trivial by the previous case, which means that g itself is, and the same for words beginning with v_k and ending with u_k.

This proves the first part of (1). The second part, concerning the group generated by $u_k \pmod p$ and $v_k \pmod p$, follows from Proposition B.2.1, (2), below: $\mathrm{SL}_2(\mathbf{F}_p)$ is generated by

$$u_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad v_1 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},$$

and for $p \nmid k$ we have $u_1 = u_k^n$, $v_1 = v_k^n$ in $\mathrm{SL}_2(\mathbf{F}_p)$, where n is the inverse of k in $\mathbf{Z}/p\mathbf{Z}$. Thus u_k and v_k also generate $\mathrm{SL}_2(\mathbf{F}_p)$ in that case.

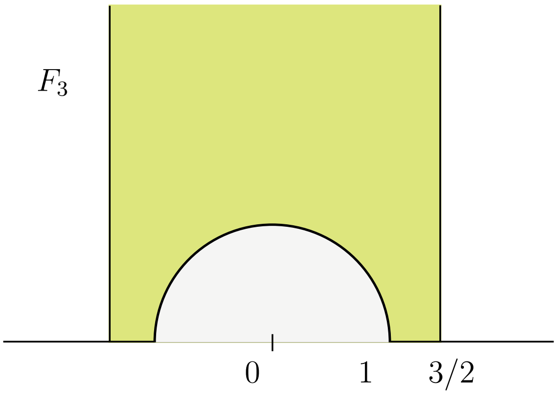


FIGURE B.1. The domain F_3

As for (3), we will only prove the first part (see the remark below concerning the case $k = 2$). We note that

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} u_k \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = -v_k^{-1}.$$

We will show that the group H_k generated by u_k, by $-\mathrm{Id}$ and by

$$w = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

is of infinite index in $\mathrm{SL}_2(\mathbf{Z})$ (this result essentially goes back to Hecke [51, §3]); since the relation above shows that H_k contains L_k, this does imply (3). We now argue following Evans [36], using a modicum of hyperbolic geometry. We consider the image of H_k in $\mathrm{PSL}_2(\mathbf{Z}) = \mathrm{SL}_2(\mathbf{Z})/\{\pm 1\}$, which acts on the upper half-plane $\mathbf{H}$ by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z = \frac{az + b}{cz + d}$$

(this is also the context of Example 5.4.1). It suffices to show that for $k \geqslant 3$ the image $\bar{H}_k$ of H_k in $\mathrm{PSL}_2(\mathbf{Z})$ is of infinite index in $\mathrm{PSL}_2(\mathbf{Z})$. To do this, we will show that the subset

$$F_k = \{z = x + iy \in \mathbf{H} \mid |\mathrm{Re}(z)| < \tfrac{1}{2}k \text{ and } |\mathrm{Im}(z)| > 1\}$$

has the property that no two distinct elements of F_k are in the same $\bar{H}_k$-orbit: if $z_1 \neq z_2$ are elements of F and $g \cdot z_1 = z_2$ with $g \in \bar{H}_k$, then $g = 1$ (see Figure B.1 for the drawing when $k = 3$).

This claim allows us to conclude. Indeed, it is a fundamental fact that, on the other hand, any $z \in \mathbf{H}$ is in the $\mathrm{PSL}_2(\mathbf{Z})$-orbit of some element in F_1 (see, e.g., [102, §VII.1.2]). Since the hyperbolic measure of F_1 (i.e., the measure according to the invariant measure $\mu_{\mathbf{H}} = y^{-2}dxdy$, see (5.11)) is finite (as the reader can quickly check) whereas the hyperbolic measure of F_k is infinite, namely

$$\int_{F_k} \frac{dxdy}{y^2} \geqslant \int_1^{k/2} \int_0^{+\infty} \frac{dy}{y^2} dx = +\infty$$

it follows that the index of $\bar{H}_k$ in $\mathrm{PSL}_2(\mathbf{Z})$ must be infinite (otherwise, F_k would be contained in the union of finitely many translates γF_1 with $\gamma \in \mathrm{SL}_2(\mathbf{R})$, and its measure would be finite, since the hyperbolic measure $y^{-2}dxdy$ is invariant under $\mathrm{SL}_2(\mathbf{R})$.)

We will now check the claim. We denote $u = u_k$ for simplicity. The advantage of using the group $\bar{H}_k$ is that since $w^2 = 1$, any element $g \neq 1$ in $\bar{H}_k$ can be written

$$g = u^{n_j}\ w\ u^{n_{j-1}}\ w\ \cdots\ w\ u^{n_2}\ w\ u^{n_1} \tag{B.1}$$

for some integer $j \geqslant 1$, where $n_i \in \mathbf{Z}$ for all i, and $n_i \neq 0$ when $2 \leqslant i \leqslant j-1$ (where $j = 1$ means that $g = u_k^{n_1}$, while $g = w$ corresponds to $j = 2$ with $n_1 = n_2 = 0$).

Let D_- (resp. D_+) be the subset of $\mathbf{H}$ defined by the condition $|z| < 1$ (resp. $|z| > 1$). The action of w is $z \mapsto -1/z$, hence it exchanges D_+ and D_-. For $n \in \mathbf{Z}$ and $z = x + iy$ in $\mathbf{H}$, we have

$$|u^n \cdot z|^2 = (x + nk)^2 + y^2.$$

Now we observe:

- If $z \in F_k$, then $u^n \cdot z \in D_+$ if $n \in \mathbf{Z}$, so $wu^n \cdot z \in D_-$ (indeed, if $n = 0$, then $z \in D_+$ by definition of F_k, and otherwise we have $|x + nk| \geqslant |n|k - k/2 \geqslant k/2$, so $|u^n \cdot z|^2 \geqslant k^2/4 + y^2 > k^2/4 > 1$).
- If $\in D_-$, then $u^n \cdot z \in D_+$ if $n \neq 0$, so $wu^n \cdot z \in D_-$ (indeed, $|x + nk| \geqslant |n|k - k/2 \geqslant k/2$, so $|u^n \cdot z|^2 \geqslant k^2/4 + y^2 > k^2/4 > 1$ again).
- If $z \in D_-$, then $u^n \cdot z \notin F_k$ if $n \in \mathbf{Z}$ (indeed, this holds by definition if $n = 0$, and otherwise $|\operatorname{Re}(u^n \cdot z)| = |x + nk| \geqslant |n|k - k/2 \geqslant k/2$.)

Let $z \in F_k$ and $g \neq 1$ in $\bar{H}_k$, given by (B.1). By the first observation, we have $wu^{n_1} \cdot z \in D_-$. Applying then repeatedly the second observation, we get

$$wu^{n-2}wu^{n-1} \cdot z \in D_-, \quad \ldots, \quad wu^{n_j-1}w \cdots u^{n_2}wu^{n_1} \cdot z \in D_-,$$

so that $g \cdot z \notin F_k$ by the third observation.

□

Exercise B.1.4. Rephrase the argument for (3) as a "proper" ping-pong argument.

Remark B.1.5. For $k = 1$, as just recalled, the group L_1 generated by

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},$$

is simply $\mathrm{SL}_2(\mathbf{Z})$. This is not a free group: e.g., we have

$$w = u_k^{-1} v_k u_k^{-1} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

which satisfies $w^4 = 1$.

For $k = 2$, one can check that L_2 (which is a free group) is of finite index in $\mathrm{SL}_2(\mathbf{Z})$, more precisely that the group generated by L_2 and the element -1 (which doesn't belong to L_2 since it is of finite order) is the finite index subgroup

$$\ker(\mathrm{SL}_2(\mathbf{Z}) \longrightarrow \mathrm{SL}_2(\mathbf{Z}/2\mathbf{Z})) = \Bigl\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \pmod 2 \Bigr\}.$$

Since L_2 is of index 2 in this subgroup, it is also of finite index.

B.2. Properties of SL_2

We gather here some properties of $\mathrm{SL}_2(k)$, for k algebraically closed, and of the finite groups $\mathrm{SL}_2(\mathbf{F}_p)$ which are used in Chapter 6. All of these are very standard, and go back to the earliest investigations of finite linear groups by Dickson.

Proposition B.2.1. (1) *For p prime, we have*

$$|\operatorname{SL}_2(\mathbf{F}_p)| = p(p^2-1).$$

(2) *For p prime, the group $\operatorname{SL}_2(\mathbf{F}_p)$ is generated by the elements*

$$u_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad v_1 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

(3) *For $p \geqslant 3$ prime, the invariant $d(\operatorname{SL}_2(\mathbf{F}_p))$, i.e., the minimal dimension of a non-trivial irreducible unitary representation of $\operatorname{SL}_2(\mathbf{F}_p)$, is equal to $(p-1)/2$. In particular, $\operatorname{SL}_2(\mathbf{F}_p)$ is perfect if $p \geqslant 5$.*

Proof. (1) is left as an exercise; for (2), see for instance [74, Lemma 8.1]. For (3), we sketch a proof of the lower bound $d(\operatorname{SL}_2(\mathbf{F}_p)) \geqslant (p-1)/2$, since this is the direction that is relevant for our application. In particular, since it shows that there is no one-dimensional representation of $\operatorname{SL}_2(\mathbf{F}_p)$ for $p \geqslant 5$, it implies that the group is perfect in that case.

Let $\varrho \neq \mathbf{1}$ be an irreducible non-trivial representation of $\operatorname{SL}_2(\mathbf{F}_p)$, with $p \geqslant 3$ a prime. Since u_1 and v_1 generate $\operatorname{SL}_2(\mathbf{F}_p)$, one of $\varrho(u_1)$ or $\varrho(v_1)$ is non-trivial. We assume that it is $\varrho(u_1)$ (the other case being similar). Since $\varrho(u_1)$ is a unitary matrix of order p, it has therefore an eigenvalue ξ that is a non-trivial root of unity of order p, hence a primitive p-th root of unity. Let $x \neq 0$ be a ξ-eigenvector. Now we use the conjugation relation

$$\begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} u_1 \begin{pmatrix} a^{-1} & 0 \\ 0 & a \end{pmatrix} = u_1^{a^2},$$

for any integer $a \geqslant 1$ coprime to p, to deduce that

$$x_a = \varrho\Bigl(\begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}\Bigr)x$$

is an eigenvector of $\varrho(u_1)$ for the eigenvalue ξ^{a^2}. Since there are $(p-1)/2$ distinct squares in $\mathbf{F}_p^\times$, this gives $(p-1)/2$ eigenspaces of $\varrho(u_1)$, and in particular the dimension of the space on which it acts must be at least that large! □

Next, recall that a *maximal subgroup* of a finite group G is a maximal *proper* subgroup.

We use the knowledge of maximal subgroups of $\operatorname{SL}_2(\mathbf{F}_p)$, which we recall. The following terminology is defined and used in Section 6.6 (see Proposition 6.6.10). A *split maximal torus* T in $\operatorname{SL}_2(\mathbf{F}_p)$ is a conjugate of the subgroup

$$T_s = \Bigl\{\begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} \mid a \in \mathbf{F}_p^\times\Bigr\},$$

of diagonal matrices (equivalently, T is the subgroup of all elements which are diagonal with respect to a fixed basis of $\mathbf{F}_p^2$). A *non-split maximal torus* in $\operatorname{SL}_2(\mathbf{F}_p)$ is a conjugate of the subgroup

$$T_{ns} = \Bigl\{\begin{pmatrix} a & b \\ b\varepsilon & a \end{pmatrix} \mid a^2 - \varepsilon b^2 = 1\Bigr\},$$

where $\varepsilon \in \mathbf{F}_p^\times$ is a fixed element which is not a square.

Theorem B.2.2 (Dickson). *Let $p \geqslant 5$ be a prime number and let $H \subset \operatorname{SL}_2(\mathbf{F}_p)$ be a maximal subgroup. Then one of the following possibilities holds:*

(1) *The group $H/\{\pm 1\}$ is isomorphic to one of the groups A_4, $\mathfrak{S}_4$, A_5.*
(2) *The group H is the normalizer of a split or non-split maximal torus.*
(3) *The group H is conjugate to the subgroup B of upper-triangular matrices.*

Proof. Many book treatments (see, e.g., [107, Exercise 7, §3.6]) prove a similar statement for the simple group $\mathrm{PSL}_2(\mathbf{F}_p) = \mathrm{SL}_2(\mathbf{F}_p)/\{\pm 1\}$, but it is easy to reduce to that case. Indeed, a maximal subgroup H of $\mathrm{SL}_2(\mathbf{F}_p)$ must contain $\{\pm 1\}$, since $H \subset H\{\pm 1\}$. Then, denoting by π the projection $\mathrm{SL}_2(\mathbf{F}_p) \to \mathrm{PSL}_2(\mathbf{F}_p)$, we have $H \subset \pi^{-1}(H')$ and by maximality, either $\pi^{-1}(H') = \mathrm{SL}_2(\mathbf{F}_p)$, or else $H = \pi^{-1}(H')$. But the first case can not occur, since it would mean that π restricted to H is an isomorphism $\varphi : H \simeq \mathrm{PSL}_2(\mathbf{F}_p)$, and then H would be normal in $\mathrm{SL}_2(\mathbf{F}_p)$ (being of index 2) and one would obtain a surjective homomorphism

$$\mathrm{SL}_2(\mathbf{F}_p) \longrightarrow \mathrm{SL}_2(\mathbf{F}_p)/H \simeq \mathbf{Z}/2\mathbf{Z},$$

contradicting the fact that $\mathrm{SL}_2(\mathbf{F}_p)$ is perfect.

We sketch one possible argument, which is not quite as precise, but is enough to derive a variant of Corollary B.2.3 below that would suffice for the applications in Chapter 6. We distinguish two cases.

Case 1. If $p \mid |H|$, then up to conjugacy, we may assume that H contains the matrix $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, hence the subgroup U it generates. If H is contained in the subgroup B of upper-triangular matrices, then we are done. Otherwise, we apply the elementary non-concentration inequality of Proposition 6.6.3 to the subset H (which is symmetric!). We deduce that

$$p = |U \cap H| \leqslant 2|H^{(5)}|^{1/3} = 2|H|^{1/3},$$

or $|H| \geqslant (p/2)^3$. It is easy to see that this is not possible if $p \geqslant 5$, so this possibility does not arise.

Case 2. If $p \nmid |H|$, then we only need the classification of subgroups of $\mathrm{SL}_2(\mathbf{F}_p)$ of order coprime to p, and this is very classical and leads to the conclusion (adapt, for instance, the recent treatment of Beauville [6] for $\mathrm{PGL}_2(\mathbf{F}_p)$). □

We use in Chapter 6 the following immediate corollary:

Corollary B.2.3. *Let $p \geqslant 5$ be a prime number and let $H \subset \mathrm{SL}_2(\mathbf{F}_p)$ be a maximal subgroup. Then either $|H| \leqslant 120$ or we have*

$$[[x_1, x_2], [x_3, x_4]] = 1$$

for any $x_i \in H$.

In our proof of certain non-concentration inequalities, we also use the classification of conjugacy classes of $\mathrm{SL}_2(\mathbf{F}_p)$ and $\mathrm{SL}_2(\bar{\mathbf{F}}_p)$.

Proposition B.2.4. (1) *Let $p \geqslant 3$ be a prime, and let $\varepsilon \in \mathbf{F}_p^\times$ be a fixed element which is not a square. There are $p+4$ conjugacy classes in $\mathrm{SL}_2(\mathbf{F}_p)$, which have representatives of the following forms:*

$$1, \quad -1, \quad \begin{pmatrix} x & 0 \\ 0 & x^{-1} \end{pmatrix}, \ x \in \mathbf{F}_p^\times - \{\pm 1\},$$

$$\begin{pmatrix} a & b \\ \varepsilon b & a \end{pmatrix}, \ b \neq 0, \ a^2 - \varepsilon^2 = 1,$$

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix},$$

$$\begin{pmatrix} 1 & \varepsilon \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} -1 & \varepsilon \\ 0 & -1 \end{pmatrix}.$$

(2) *Let k be an algebraically closed field of characteristic $p \geqslant 3$. Representatives of the conjugacy classes in* $\mathrm{SL}_2(k)$ *are given by the elements*

$$1, \quad -1, \quad \begin{pmatrix} x & 0 \\ 0 & x^{-1} \end{pmatrix}, \; x \in k^\times - \{\pm 1\},$$

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}.$$

Proof. See [41, §5.2] for the first part (or adapt the argument for $\mathrm{GL}_2(\mathbf{F}_p)$ in [72, §4.6.4]); the second is easier and is proved along the same lines. □

The following last result also concerns SL_m for $m \geqslant 3$, and is used in Section 5.3.

Lemma B.2.5. *Let $m \geqslant 1$. The elementary matrices $s_{i,j} = \mathrm{Id} + E_{i,j}$ with $1 \leqslant i \neq j \leqslant m$ generate* $\mathrm{SL}_m(\mathbf{Z})$.

Proof. This follows of course from the more precise Carter-Keller Theorem, see [7, Th. 4.1.3]), but this is a classical statement that we can sketch. Recall that multiplying by $s_{i,j}$ and the left and right amounts to replacing a matrix $g \in \mathrm{SL}_m(\mathbf{Z})$ by the matrix where the i-th row is replaced by the sum of the i-th and j-th rows of g (resp. the j-th column is replaced by the sum of the i-th and j-th columns). Let $g \in \mathrm{SL}_m(\mathbf{Z})$. Fix a column j where $g_{1,j} \neq 0$ and has smallest absolute value. Using column operations (for instance) and euclidean division, one obtains a new matrix where $g'_{1,j}$ has the largest absolute value of the non-zero coefficients in the first row. Iterating, one obtains a matrix with a single non-zero coefficient in the first row, which must be ± 1 to have determinant 1. Continuing with each row in turn, one ends up with a permutation matrix in $\mathrm{SL}_m(\mathbf{Z})$. It is then an elementary manipulation to reduce this to the identity using further row and column operations. □

Proposition B.2.6. *Let $m \geqslant 3$ be an integer and let $q \geqslant 1$ be a positive squarefree integer. The homomorphism $\mathrm{SL}_m(\mathbf{Z}) \to \mathrm{SL}_m(\mathbf{Z}/q\mathbf{Z})$ of reduction modulo q is surjective.*

(In fact, the statement holds for all $q \geqslant 1$, but we do not need this).

Proof. For any prime number p, the elementary matrices $s_{i,j} = \mathrm{Id} + E_{i,j}$ with $1 \leqslant i \neq j \leqslant m$ reduce modulo p to the corresponding elementary matrices in $\mathrm{SL}_m(\mathbf{F}_p)$. These generate $\mathrm{SL}_m(\mathbf{F}_p)$ (see, e.g., [74, Lemma 8.1, Prop. 9.1], taking into account the fact that $(\mathrm{Id} + E_{i,j})^n = \mathrm{Id} + nE_{i,j}$ for all $n \in \mathbf{Z}$), so we obtain the statement when $q = p$ is a prime.

In the general case, observe that for any squarefree integer $q \geqslant 1$, the Chinese Remainder Theorem gives an isomorphism

$$\mathrm{SL}_m(\mathbf{Z}/q\mathbf{Z}) \to \prod_{p \mid q} \mathrm{SL}_m(\mathbf{F}_p).$$

Taking $1 \leqslant i \neq j \leqslant m$, and a prime $p \mid q$, we can find an integer k such that $k \equiv 1 \bmod p$ and $k \equiv 0 \bmod \ell$ for $\ell \mid q$ different from p. Then $\mathrm{Id} + kE_{i,j}$ has image modulo q that has p-component equal to the elementary matrix $s_{i,j}$, and ℓ-components the identity for $\ell \neq p$. Varying p, i and j, these matrices generate the direct product

$$\prod_{p \mid q} \mathrm{SL}_m(\mathbf{F}_p),$$

hence the result. □

APPENDIX C

Varia

We discuss in this section, with references when needed, a few standard facts of algebraic number theory and (finite-dimensional) normed vector spaces that are needed for some parts of the book. We also discuss some less well-known properties of polynomials with only real zeros or their multi-variable generalizations, which are used in the construction of Ramanujan graphs in Section 4.2.

C.1. The norm of linear maps

We use in some parts of the book the elementary properties of the norm on matrices (or linear maps).

Definition C.1.1. Let $n \geqslant 1$ and let $g \in \mathrm{GL}_n(\mathbf{C})$. The norm of g is

$$\|g\| = \max_{v,w\neq 0} \frac{|\langle gv, w\rangle|}{\|v\|\|w\|}$$

where $\langle\cdot,\cdot\rangle$ is the standard inner product on $\mathbf{C}^n$.

Lemma C.1.2. *Let $n \geqslant 1$.*

(1) *For g_1 and g_2 in $\mathrm{GL}_n(\mathbf{C})$, we have*

$$\|g_1 g_2\| \leqslant \|g_1\|\|g_2\| \tag{C.1}$$

(2) *For $g \in \mathrm{GL}_n(\mathbf{C})$, we have*

$$\max_{i,j} |g_{i,j}| \leqslant \|g\| \text{ for } g = (g_{i,j}), \tag{C.2}$$

(3) *For any $n \geqslant 0$, we have*

$$\|g^n\| \geqslant \max_{\lambda} |\lambda|$$

where λ runs over the eigenvalues of g.

Proof. (1) is elementary.

(2) This is because $g_{i,j} = \langle ge_i, e_j\rangle$ in terms of the canonical basis $(e_1, \ldots, e_n)$ of $\mathbf{C}^n$.

(3) If $gv = \lambda v$, then $\|g\| \geqslant |\langle gv, v\rangle|/\|v\|^2 = |\lambda|$. Moreover, since λ^n is an eigenvalue of g^n for any $n \geqslant 0$, we have also $\|g^n\| \geqslant |\lambda|^n$. □

C.2. Finite-dimensional unitary representations of abelian groups

Let A be a discrete abelian group and

$$\varrho\colon A \to \mathrm{U}(E)$$

a unitary representation of A on a finite-dimensional Hilbert space E. Let $\widehat{A}$ be the dual group

$$A = \{\chi\colon A \to \mathbf{S}^1\}$$

of characters of A. For each character χ, let

$$E_\chi = \{v \in E \mid \varrho(x)v = \chi(x)v \text{ for all } x \in A\}.$$

Then there exists a finite set $X \subset \widehat{A}$ such that we have the orthogonal decomposition

$$E = \bigoplus_{\chi \in X} E_\chi \tag{C.3}$$

and $E_\chi \neq 0$ for $\chi \in X$. This is the *isotypic decomposition* of E as a representation of A.

To see this, we use the spectral theorem for families of commuting unitary transformations of E:

THEOREM C.2.1. *Let E be a finite-dimensional Hilbert space, and let $M \subset \mathrm{U}(E)$ a set of unitary transformations of E. If all elements of M commute, then there exists a basis B of E such that any element $v \in B$ is an eigenvector of all operators $f \in M$.*

We apply this theorem to $M = \{\varrho(x) \mid x \in A\}$. If v is an element of a basis B of common eigenvectors, if follows that there are complex numbers $\lambda_v(x) \in \mathbf{C}$ such that

$$\varrho(x)v = \lambda_v(x)$$

for all $x \in A$. Moreover, $\lambda_v(x) \in \mathbf{S}^1$, as an eigenvalue of a unitary operator. Since we have $\varrho(x+y) = \varrho(x)\varrho(y)$ and $v \neq 0$, we obtain $\lambda_v(x+y) = \lambda_v(x)\lambda_v(y)$, or in other words, the map $x \mapsto \lambda_v(x)$ is an element χ of $\widehat{A}$. Hence $v \in E_\chi$. This means that the decomposition (C.3) holds, with X the set of characters χ arising from the vectors $v \in B$.

C.3. Algebraic integers

LEMMA C.3.1. *Let $d \geqslant 1$ be an integer and let $\eta \geqslant 0$ be a real number. Let $P_{d,\eta}$ to be the set of all integral monic polynomials f of degree d with $f(0) \neq 0$, such that all roots α of f have $|\alpha| \leqslant 1 + \eta$. Then $P_{d,\eta}$ is* finite.

PROOF. Let $f \in P_{d,\eta}$ be given. We can write

$$f = X^d + a_{d-1}X^{d-1} + \cdots + a_1 X + a_0 = \prod_{i=1}^{d} (X - \alpha_i),$$

where a_i are integers and $|\alpha_i| \leqslant 1 + \eta$. For $0 \leqslant j \leqslant d-1$, we have

$$a_j = (-1)^{d-j} \sum_{\substack{J \subset \{1,\ldots,d\} \\ |J| = j}} \prod_{\alpha \in J} \alpha,$$

and therefore

$$|a_j| \leqslant \binom{d}{j} (1+\eta)^j.$$

Counting the possibilities for each a_i, it follows that

$$|P_{d,\eta}| \leqslant \prod_{j=0}^{d-1} \Bigl(1 + 2\binom{d}{j}(1+\eta)^j\Bigr)$$

and so $P_{d,\eta}$ is finite. □

LEMMA C.3.2. *Let $f \in \mathbf{Z}[X]$ be a monic polynomial. If all roots of f have modulus 1, then they are roots of unity.*

PROOF. We use some basic facts of algebraic number theory in this proof. For an integer $d \geqslant 1$, let R_d be the set of complex numbers which are roots of a monic integral polynomial of degree at most d, with all roots of modulus 1. This set is finite, because the elements of R_d are roots of the finite set $P_{1,0} \cup \cdots \cup P_{d,0}$ of integral monic polynomials of degree $\leqslant d$, all of whose roots have modulus at most 1 (Lemma C.3.1).

Consider then a monic polynomial $f \in \mathbf{Z}[X]$, with all roots of modulus 1, and let d be its degree. Let α be a root of f. By algebraic number theory, for any integer $k \geqslant 1$, α^k is the root of some irreducible monic integral polynomial g of degree at most d, and all roots of g are of the form β^k for some root β of k. Hence α^k belongs to R_d for all $k \geqslant 1$. Since R_d is finite, there exist two distinct integers k_1 and k_2 with $\alpha^{k_1} = \alpha^{k_2}$, which means that α is a root of unity. □

Lemma C.3.3. *Let $d \geqslant 1$ be an integer. There exists a real number $\eta_d > 0$ such that if $f \in \mathbf{Z}[X]$ is a monic polynomial of degree d with $f(0) \neq 0$, then either all roots of f in $\mathbf{C}$ are roots of unity, or there exists a root $\alpha \in \mathbf{C}$ of f such that $|\alpha| \geqslant 1 + \eta_d$.*

Proof. We first observe that a monic integral polynomial f of degree d with $f(0) \neq 0$ has at least one root of modulus $\geqslant 1$: indeed, the product of the roots is (up to sign) equal to $f(0)$, which is a non-zero integer, hence of modulus $\geqslant 1$.

Since the set $P_{d,1}$ of Lemma C.3.1 is finite, there exists a smallest real number $\eta_d > 0$ such that if $f \in P_{d,1}$ has a root α of modulus > 1, then $|\alpha| \geqslant 1 + \eta_d$ (note that $(X-2)^d$ belongs to $P_{d,1}$, so some polynomial at least has a root with modulus > 1). By the observation at the beginning of the proof, a polynomial $f \in \mathbf{Z}[X]$ which does not have all roots of modulus 1 either has a root with modulus $\geqslant 2$, or belongs to $P_{d,1}$ and has a root with modulus $\geqslant 1 + \eta_d$. We conclude using the previous lemma, which shows that if all roots of f have modulus 1, then they are roots of unity. □

Corollary C.3.4. *Let $d \geqslant 1$ be an integer. There exists a real number $\eta_d > 0$ such that if $A \in \mathrm{SL}_d(\mathbf{Z})$ then either all eigenvalues of A are roots of unity, or there exists an eigenvalue $\alpha \in \mathbf{C}$ of A such that $|\alpha| \geqslant 1 + \eta_d$.*

Proof. The eigenvalues of A are the roots of the characteristic polynomial f_A of A, which is a monic integral polynomial of degree d with $f_A(0) \neq 0$. Hence the result follows from the previous lemma. □

C.4. Real stable polynomials

We collect in this section some properties of polynomials with real roots (or their generalizations to many variables, the "real stable" polynomials). For more details and further applications, the paper of Marcus, Spielman and Srivastava [83] as well as A. Valette's survey [114, Th. 1.8] are very readable.

We recall that $\mathbf{H}$ denotes the upper half-plane in $\mathbf{C}$, namely the set of $z \in \mathbf{C}$ such that $\mathrm{Im}(z) > 0$.

Definition C.4.1. Let $d \geqslant 1$ be an integer. A polynomial $p \in \mathbf{R}[X_1, \ldots, X_d]$ is called *real stable* if $p(\boldsymbol{z}) \neq 0$ for all $\boldsymbol{z} \in \mathbf{H}^d$.

If $d = 1$, a polynomial $p \in \mathbf{R}[X]$ is real stable if and only if it has only real roots, since if z is a root of p, then so is $\bar{z}$, and one of these belongs to $\mathbf{H}$ if $z \notin \mathbf{R}$.

Proposition C.4.2. *Let $d \geqslant 1$ be an integer and let H be a finite-dimensional Hilbert space.*

(1) *If $(u_i)_{1 \leqslant i \leqslant d}$ are positive endomorphisms of H, and u_1 is an isomorphism, then the polynomial*

$$\det(X_1 u_1 + \cdots + X_n u_d)$$

is real stable.

(2) *Let $p \in \mathbf{R}[X_1, \ldots, X_d]$ be real stable and $1 \leqslant i \leqslant d$. The polynomial $(1+\partial_i)p = p + \frac{\partial p}{\partial X_i}$ is real stable.*

(3) *Let $p \in \mathbf{R}[X_1, \ldots, X_d]$ be a real stable polynomial and $t \in \mathbf{R}$. Let*

$$p_t = p(X_1, \ldots, X_{d-1}, t) \in \mathbf{R}[X_1, \ldots, X_{d-1}].$$

Either p_t is zero, or p_t is real stable.

Proof. (1) Let $(z_1, \ldots, z_d) \in \mathbf{H}^d$ and let $x \in H$ be such that

$$z_1 u_1(x) + \cdots + z_d u_d(x) = 0.$$

By computing the inner product with x and taking the imaginary part, we get

$$\mathrm{Im}(z_1)\langle u_1(x), x\rangle + \cdots + \mathrm{Im}(z_d)\langle u_d(x), x\rangle = 0,$$

since each $\langle u_i(x), x\rangle$ is real. In fact, since $u_i \geqslant 0$, each term is non-negative, hence each of them must vanish, i.e., $\langle u_i(x), x\rangle = 0$ for all i. For $i = 1$, this implies that $x = 0$. Hence $z_1 u_1 + \cdots + z_d u_d$ is injective, and the determinant is non-zero at $(z_1, \ldots, z_d)$.

(2) We may assume that $i = 1$ and that $t \neq 0$. Fix $(z_2, \ldots, z_d) \in \mathbf{H}^{d-1}$ and let $q = p(X, z_2, \ldots, z_d) \in \mathbf{R}[X]$. This polynomial only has real roots, and we need to show that the polynomial $q + tq'$ has no root in $\mathbf{H}$. Let $z \in \mathbf{H}$. We have $q(z) \neq 0$ and

$$q(z) + tq'(z) = q(z)\Big(1 + t\frac{q'}{q}(z)\Big) = q(z)\Big(1 + t\sum_i \frac{1}{z - \alpha_i}\Big),$$

where (α_i) are the roots of q. These are real, hence

$$\mathrm{Im}\Big(1 + t\sum_i \frac{1}{z - \alpha_i}\Big) = t\sum_i \frac{\mathrm{Im}(z)}{|z - \alpha_i|^2} > 0,$$

so that z is not a root of $q + tq'$.

(3) For $(z_1, \ldots, z_{d-1})$ in a compact subset K of $\mathbf{H}^{d-1}$, we can write

$$p_t = \lim_{n \to +\infty} p(z_1, \ldots, z_{d-1}, t + i/n)$$

uniformly on K. By a theorem of Hurwitz (see lemma C.4.3 below, applied for each of the $d-1$ variables when the others are fixed) we have either $p_t = 0$ or p_t has no zero in $\mathbf{H}^{d-1}$. □

Lemma C.4.3. *Let D be a non-empty open disc in $\mathbf{C}$. Let (f_n) be a sequence of continuous functions on $\bar{D}$ which are holomorphic in D. Assume that (f_n) converges to f on D, so that f is also continuous in $\bar{D}$ and holomorphic in D. If $f_n(z) \neq 0$ all $n \geqslant 1$ and $z \in D$, then either $f = 0$ or f does not vanish on D.*

Proof. We assume that f is not zero, and show that it has no zero in D. Let $z_0 \in D$ be fixed, and let C be a circle of radius $r > 0$ centered at z_0 and such that $C \subset D$. Since f is non-zero, it is non-zero in the disc with boundary C, and by the maximum modulus principle, it is non-zero on C. In particular, we have $\delta = \inf_{z \in C} |f(z)| > 0$. For n large enough, we get

$$\sup_{z \in C} |f(z) - f_n(z)| < \delta,$$

and then the relation $f = f - f_n + f_n$ combined with Rouché's Theorem (see, e.g., [111, 3.42]) shows that f has the same number of zeros as f_n in the disc bounded by C. This means that f has no zeros there, and in particular that $f(z_0) \neq 0$. □

Finally, we have a convexity property of zeros of polynomials with only real roots. Recall that for $p \in \mathbf{R}[X]$, we denote by $\varrho^+(p)$ the largest real zero of p (if it exists).

Proposition C.4.4. *Let $(p_0,\dots,p_k)$ be monic polynomials in $\mathbf{R}[X]$. Assume that for every non-negative real numbers t_i with $t_0+\cdots+t_k=1$, the polynomial $q=t_0p_0+\cdots+t_kp_k$ has only real roots.*

Then $\varrho^+(q)$ belongs to the convex hull of the real numbers $\varrho^+(p_i)$ for all such (t_i).

We emphasize that, although one might be interested in the conclusion for a single choice of (t_i), the *assumption* must be verified for *all* choices.

Proof. Using induction on k, it is enough to show that if p_0 and p_1 are monic polynomials such that $(1-t)p_0+tp_1$ has only real roots for all $t\in[0,1]$, and $\varrho^+(p_0)\leqslant\varrho^+(p_1)$, then for any $t\in[0,1]$, we have $\varrho^+((1-t)p_0+tp_1)\in[\varrho^+(p_0),\varrho^+(p_1)]$.

Let $t\in[0,1]$ and put $q=(1-t)p_0+tp_1$.

If $x\in\mathbf{R}$ is $>\varrho^+(p_1)$, then it is also $>\varrho^+(p_0)$, and therefore we have $p_1(x)>0$ and $p_0(x)>0$ since both polynomials are monic. Hence $(1-t)p_0(x)+tp_1(x)>0$, which shows that $q=(1-t)p_0+tp_1$ has no zero $>\varrho^+(p_1)$. Consequently, we have $\varrho^+(q)\leqslant\varrho^+(p_1)$.

To prove that $\varrho^+(q)\geqslant\varrho^+(p_0)$, it is enough to show that $p_1(\varrho^+(p_0))\leqslant 0$, since we then have $q(\varrho^+(p_0))\leqslant 0$, hence by continuity the polynomial q has as zero $\geqslant p_0$.

We argue by contradiction. The inequality $p_1(\varrho^+(p_0))>0$ would imply that there exists $\delta>0$ such that p_1 is non-zero in the interval $[\varrho^+(p_0),\varrho^+(p_0)+\delta]$, and we may assume that $\varrho^+(p_0)<\varrho^+(p_1)$. Let $n(u)$ be the number of zeros of $(1-u)p_0+up_1$ in the interval $[\varrho^+(p_0)+\delta,\varrho^+(p_1)]$, counted with multiplicity. This is a continuous function on $[0,1]$, hence it is constant. Indeed, note that for any $u\in[0,1]$ we have

$$((1-u)p_0+up_1)(\varrho^+(p_0)+\delta)>0$$

while for any $x>\varrho^+(p_1)$, it holds $((1-u)p_0+up_1)(x)>0$, so that $n(u)$ is also the number of zeros in the disc D with diameter $[\varrho^+(p_0)+\delta,\varrho^+(p_1)+\delta]$ in $\mathbf{C}$ (there being no zeros that are not real), which can be expressed by the integral

$$n(u)=\frac{1}{2i\pi}\int_{\partial D}\frac{-up_0'(z)+up_1'(z)}{(1-u)p_0(z)+up_1(z)}dz,$$

which is a continuous function of u.

However, $n(0)=0$ since p_0 has no zero $>\varrho^+(p_0)$, while $n(1)\geqslant 2$, since $p_1(\varrho^+(p_0)+\delta)>0$ but p_1 vanishes at $\varrho^+(p_1)$. This is the desired contradiction. ☐

C.5. Mixed characteristic polynomials

Definition C.5.1. Let H be a finite dimensional $\mathbf{C}$-vector space. Let $d\geqslant 1$ be an integer and let $\boldsymbol{u}=(u_1,\dots,u_d)$ be a d-tuple of endomorphisms of H. The polynomial

$$\Big(1-\frac{\partial}{\partial X_1}\Big)\cdots\Big(1-\frac{\partial}{\partial X_d}\Big)\det(X+X_1u_1+\cdots+X_nu_n)|_{X_1=\cdots=X_d=0}\in\mathbf{C}[X]$$

is called the *mixed characteristic polynomial* of the family $\boldsymbol{u}$. It is denoted $\boldsymbol{\mu}(\boldsymbol{u})$.

It is elementary that $\boldsymbol{\mu}(\boldsymbol{u})$ is a monic polynomial of degree the dimension of H. Moreover, it is symmetric in the sense that $\boldsymbol{\mu}(\boldsymbol{v})=\boldsymbol{\mu}(\boldsymbol{u})$ if $\boldsymbol{v}$ is a permutation of $\boldsymbol{u}$.

Example C.5.2. Take $H=\mathbf{C}^2$ and $d=2$, and identify endomorphisms of H with square matrices of size 2. Then we compute that

$$\det\Big(X+X_1\begin{pmatrix}a&b\\c&d\end{pmatrix}+X_2\begin{pmatrix}\alpha&\beta\\\gamma&\delta\end{pmatrix}\Big)=X^2+X(X_1d+X_2\delta+X_1a+X_2\alpha)$$
$$+X_1^2(ad-bc)+X_2^2(\alpha\delta-\beta\gamma)+X_1X_2(a\delta+\alpha d-c\beta-d\gamma).$$

The differential operator to apply is

$$1-\frac{\partial}{\partial X_1}-\frac{\partial}{\partial X_2}+\frac{\partial^2}{\partial X_1\partial X_2},$$

so many terms vanish after applying it and putting $X_1=X_2=0$. We get

$$\boldsymbol{\mu}\Bigl(\begin{pmatrix}a & b\\ c & d\end{pmatrix},\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}\Bigr)=X^2-(a+d+\alpha+\delta)X+(a\delta-b\gamma+\alpha d-c\beta).$$

This can be compared with the characteristic polynomial of $\begin{pmatrix}a & b\\ c & d\end{pmatrix}$, which is

$$X^2-(a+d)X+(ad-bc).$$

We see that the formula for $\boldsymbol{\mu}$ “mixes” the two matrices.

Many of the properties below for mixed characteristic polynomials can be directly checked in this particular case.

Lemma C.5.3. *Let H be a finite-dimensional Hilbert space and $\boldsymbol{u}$ a finite family of positive endomorphisms of H. Then $\boldsymbol{\mu}(\boldsymbol{u})$ has only real roots.*

Proof. This follows from Proposition C.4.2: the first part shows that the polynomial $\det(X+X_1u_1+\cdots+X_nu_n)$ is real stable (as a polynomial in $(X,X_1,\ldots,X_d)$; then the second part, applied successively for each variable, shows that

$$\Bigl(1-\frac{\partial}{\partial X_1}\Bigr)\cdots\Bigl(1-\frac{\partial}{\partial X_d}\Bigr)\det(X+X_1u_1+\cdots+X_nu_n)$$

is real stable; since it is monic of degree $\dim(H)$ as a polynomial in X, it remains non-zero after any specialization of $(X_1,\ldots,X_d)$, and the third part of the lemma, applied again repeatedly, gives the result. □

Let $(V_i)_{i\in I}$ be a finite family of vector spaces over $\mathbf{C}$, and let V be a $\mathbf{C}$-vector space (e.g., $R=\mathbf{C}[X]$). A map

$$f\colon \bigoplus_{i\in I}V_i\to V$$

is said to be *affine-linear* if, for any $i_0\in I$ and any tuple $(x_i)_{i\neq i_0}$ with $x_i\in V_i$, the map from V_i to V given by

$$x\mapsto f(x_1,\ldots,x_{i_0-1},x,x_{i_0+1},\ldots,x_n)$$

is of the form $x\mapsto a_0+g(x)$ for some $a_0\in V$ and some linear map $g\colon V_i\to V$. From this definition, any linear combination of affine-linear maps is still affine-linear.

A consequence of the affine-linear property is that f preserves convex combinations in each argument: if $k\geqslant 1$ is an integer and $(t_i)_{1\leqslant i\leqslant k}$ are complex numbers such that $t_1+\cdots+t_k=1$, then for any $(y_1,\ldots,y_k)\in V_{i_0}$ and $(x_i)_{i\neq i_0}$, we have

$$f\Bigl(x_1,\ldots,x_{i_0-1},\sum_j t_jy_j,x_{i_0+1},\ldots,x_n\Bigr)=\sum_j t_jf(x_1,\ldots,x_{i_0-1},y_j,x_{i_0+1},\ldots,x_n).$$

Indeed, by assumption on the sum of the t_j’s, the left-hand side is

$$a_0+g\Bigl(\sum_j t_jy_j\Bigr)=a_0+\sum_j t_jg(y_j)=\sum_j t_j(a_0+g(y_j)),$$

hence the result.

LEMMA C.5.4. *Let H be a finite-dimensional $\mathbf{C}$-vector space and $\boldsymbol{u} = (u_1, \dots, u_d)$ a tuple of endomorphisms of H of rank $\leqslant 1$. Let v be an endomorphism of H. The polynomial*

$$\det(v + X_1u_1 + \cdots + X_du_d) \in \mathbf{C}[X_1, \dots, X_d]$$

is of degree $\leqslant 1$ in each variable.

PROOF. It suffices to prove that whenever $(x_2, \dots, x_d)$ are fixed in $\mathbf{C}$, the function $x \mapsto \det(v + xu_1 + x_2u_2 + \cdots + x_du_d)$ is a polynomial in x of degree $\leqslant 1$. If $u_1 = 0$, then this is obvious. Otherwise, let $w = v + x_2u_2 + \cdots + x_du_d$. Let e_1 be a vector generating the image of u_1 and $\boldsymbol{e} = (e_1, \dots, e_{\dim(H)})$ an ordered basis of H with first vector equal to e. The matrix representing $w + xu_1$ in the basis $\boldsymbol{e}$ has column vectors

$$w_i + \alpha_i x e_1, \qquad w_i = w(e_i),$$

for some $\alpha_i \in \mathbf{C}$, not all zero. By subtracting suitable multiples of a column where $\alpha_i \neq 0$ from the others, we obtain a matrix where x appears only in a single column. Expanding the determinant along this column, the result follows. □

PROPOSITION C.5.5. *Let H be a finite-dimensional $\mathbf{C}$-vector space of dimension n and $d \geqslant 1$. The map $\boldsymbol{u} \mapsto \boldsymbol{\mu}(\boldsymbol{u})$ from* $\mathrm{End}(H)^d$ *to* $\mathbf{C}[X]$ *is affine-linear in each argument u_i. In particular, if $k \geqslant 1$ and if t_1, ..., t_k are complex numbers with sum 1, we have*

$$\boldsymbol{\mu}\Big(\sum_j t_j u_{j,1}, u_2, \dots, u_d\Big) = \sum_j t_j \boldsymbol{\mu}(u_{j,1}, u_2, \dots, u_d).$$

PROOF. We identity $\mathrm{End}(H)$ with matrices of size n using a fixed basis of H. We will view the coefficients of the matrices representing u_1, ..., u_d as indeterminates $u_i = (u_{j,k}^{(i)})$. If we expand the determinant $\det(X + X_1u_1 + \cdots + X_du_d)$, we obtain a linear combination of terms of the form

$$\prod_{i=1}^{n} (\delta_i X + X_1 a_{i,1} + \cdots + X_d a_{i,d})$$

where each δ_i is either 0 or 1 and each $a_{i,j}$ is one of the coefficients of u_j. Expanding further this product leads to a sum of terms of the form

$$T = X^m \prod_{j=1}^{d} X_j^{n_j} \prod_{k=1}^{n_j} a_{j,k}, \tag{C.4}$$

for some integers $m \geqslant 0$ and $n_j \geqslant 0$, where $a_{j,k}$ is one of the coefficients of u_j.

Now we apply the differential operator to such an expression. The differential operator expands as

$$\Big(1 - \frac{\partial}{\partial X_1}\Big) \cdots \Big(1 - \frac{\partial}{\partial X_d}\Big) = \sum_{J \subset \{1,\dots,d\}} (-1)^{|J|} \partial_J,$$

where

$$\partial_J = \prod_{j \in J} \frac{\partial}{\partial X_j}.$$

For any d-tuple $\alpha = (\alpha_1, \dots, \alpha_d)$ of non-negative integers, we denote (as usual)

$$X^\alpha = \prod_{j=1}^{d} X_j^{\alpha_j}.$$

The key point is the following "orthogonality" relation: for any J and any α, we have

$$(\partial_J X^\alpha)|_{X_1=\cdots=X_d=0} = 0$$

unless $j \mapsto \alpha_j$ is the characteristic function of J, in which case

$$(\partial_J X^\alpha)|_{X_1=\cdots=X_d=0}$$

is a non-zero constant (depending on J). Applied to a term of the form (C.4), this fact implies that

$$\sum_{J\subset\{1,\ldots,d\}} (-1)^{|J|}\partial_J T$$

is a linear combination (with integral coefficients) of terms

$$X^m \prod_{j\in J} b_j,$$

where $m \geqslant 0$, $J \subset \{1, \ldots, d\}$ and b_j is a coefficient of u_j. Because, for each argument u_j, at most *one* coefficient appears in each term of the sum, this expression (and consequently the whole mixed characteristic polynomial) is affine-linear as a function of u_j, as desired.

The last assertion follows from the general remark about affine-linear maps above. □

Remark C.5.6. The reader can follow this argument step by step when $H = \mathbf{C}^2$ and $d = 2$ in Example C.5.2.

Corollary C.5.7. *Let H be a finite-dimensional $\mathbf{C}$-vector space and $\boldsymbol{u} = (u_1, \ldots, u_d)$ a tuple of endomorphisms of H of rank $\leqslant 1$. We have*

$$\boldsymbol{\mu}(\boldsymbol{u}) = \det(X - (u_1 + \cdots + u_d)).$$

Proof. By Lemma C.5.4, $\det(v + X_1u_1 + \cdots + X_du_d)$ is a polynomial of degree $\leqslant 1$, hence is equal to its Taylor expansion to order 1 at 0, so the value at $(t_1, \ldots, t_d)$ is equal to

$$\Bigl(1 + t_1\frac{\partial}{\partial X_1}\Bigr)\cdots\Bigl(1 + t_d\frac{\partial}{\partial X_d}\Bigr)\det(v + X_1u_1 + \cdots + X_du_d)|_{X_1=\cdots=X_d=0}$$

We specialize to $v = z\mathrm{Id}$ and $t_i = -1$ to obtain

$$\begin{aligned}\det(z - u_1 - \cdots - u_d) &= \Bigl(1 - \frac{\partial}{\partial X_1}\Bigr)\cdots\Bigl(1 - \frac{\partial}{\partial X_d}\Bigr)\det\Bigl(v + \sum_i X_iu_i\Bigr)|_{X_1=\cdots=X_d=0}\\ &= \boldsymbol{\mu}(\boldsymbol{u}).\end{aligned}$$

□

Corollary C.5.8. *Let H be a finite-dimensional $\mathbf{C}$-vector space and $\boldsymbol{u} = (u_1, \ldots, u_d)$ independent random variables with values in the endomorphisms of H of rank $\leqslant 1$. We have*

$$\mathbf{E}(\det(X - (u_1 + \cdots + u_d))) = \boldsymbol{\mu}((\mathbf{E}(u_i))_{1\leqslant i\leqslant d}).$$

Proof. Let $v = u_1 + \cdots + u_d$, a random variable with values in endomorphisms of H. The previous corollary shows that $\det(X - v) = \boldsymbol{\mu}(\boldsymbol{u})$. By Proposition C.5.5, the random variable $\boldsymbol{\mu}(\boldsymbol{u})$ is affine-linear in the coefficients of the matrices in $\boldsymbol{u}$. Since these are independent, and expectation is convex combination, it follows that

$$\mathbf{E}(\boldsymbol{\mu}(\boldsymbol{u})) = \boldsymbol{\mu}((\mathbf{E}(u_i))_{1\leqslant i\leqslant d}),$$

hence

$$\mathbf{E}(\det(X - v)) = \boldsymbol{\mu}((\mathbf{E}(u_i))_{1\leqslant i\leqslant d}).$$

□

Corollary C.5.9. *Let H be a finite-dimensional Hilbert space and $\boldsymbol{u} = (u_2, \ldots, u_d)$ a tuple of positive endomorphisms of H. For any $k \geqslant 1$ and any $(t_1, \ldots, t_k)$ non-negative real numbers with sum 1, and for any positive endomorphisms $(v_1, \ldots, v_k)$, the polynomial*

$$\sum_{i=1}^{k} t_i \boldsymbol{\mu}(v_i, u_2, \ldots, u_d)$$

is real stable.

Proof. Indeed, Proposition C.5.5 shows that

$$\sum_{i=1}^{k} t_i \boldsymbol{\mu}(v_i, u_2, \ldots, u_d) = \boldsymbol{\mu}\Big(\sum_{i=1}^{k} t_i v_i, u_2, \ldots, u_d\Big),$$

and since the space of positive endomorphisms is a cone in $\mathrm{End}(H)$, this mixed characteristic polynomial is real stable by Lemma C.5.3. □

Bibliography

[1] D. Abramovich: *A linear lower bound on the gonality of modular curves*, International Math. Res. Notices 20 (1996), 1005–1011. 146

[2] D. Abramovich and F. Voloch: *Lang's conjectures, fibered powers, and uniformity*, New York J. of Math. 2 (1996), 20–34. 145

[3] M. Agrawal, N. Kayal and N. Saxena: *PRIMES is in P*, Annals of Math. 160 (2004), 781–793. 119

[4] M. Aschenbrenner, S. Friedl and H. Wilton: 3-*manifold groups*, EMS Lectures in Math., E.M.S Publ. House, 2015. 140

[5] Ya.M. Barzdin and A.N. Kolmogorov: *On the realization of networks in three-dimensional space*, in Selected Works of Kolmogorov, Volume 3, Kluwer Academic Publishers, Dordrecht, 1993. 5, 51, 117, 118

[6] A. Beauville: *The finite subgroups of* $\mathrm{PGL}_2(K)$, in "Vector bundles and complex geometry", 23–29, Contemp. Math., 522, Amer. Math. Soc., 2010. 213

[7] B. Bekka, P. de la Harpe and A. Valette: *Kazhdan's Property* (T), New Math. Monographs 11, Cambridge Univ. Press (2008). 4, 108, 109, 214

[8] N. Bergeron: *Variétés en expansion*, Séminaire Bourbaki, exposé 1132 (June 2017). 117, 118, 138

[9] J.A. Bondy and U.S.R. Murty: *Graph theory*, Graduate Texts Math. 244, Springer, 2008. 11, 12

[10] N. Bourbaki: *Topologie algébrique*, Springer (2016). 13, 145, 208

[11] N. Bourbaki: *Espaces vectoriels topologiques*, Springer. 114

[12] J. Bourgain and A. Gamburd: *Uniform expansion bounds for Cayley graphs of* $\mathrm{SL}_2(\mathbf{F}_p)$, Ann. of Math. 167 (2008), 625–642. 104, 150, 168

[13] J. Bourgain, A. Gamburd and P. Sarnak: *The affine linear sieve*, Invent. math. 179 (2010), 559–644. 7, 106

[14] J. Bourgain and P. Varjú: *Expansion in* $\mathrm{SL}_d(\mathbf{Z}/q\mathbf{Z})$*, q arbitrary*, arXiv:1006.3365. 106

[15] E. Breuillard and H. Oh (editors): *Thin groups and super-strong-approximation*, MSRI Publications 61, Cambridge (2014). 105

[16] E. Breuillard, B. Green and T. Tao: *Approximate subgroups of linear groups*, GAFA 21 (2011), 774–819; arXiv:1005.1881. 105, 106, 174, 183

[17] E. Breuillard, B. Green and T. Tao: *The structure of approximate groups*, Publ. Math. IHÉS 116 (2012), 115–221. 161

[18] R. Brooks: *The spectral geometry of a tower of coverings*, J. Diff. Geometry 23 (1986), 97–107. 137

[19] M. Burger: *Petites valeurs propres du Laplacien et topologie de Fell*, doctoral thesis (1986), Econom Druck AG (Basel). 137, 141, 142, 144

[20] M. Burger: *Kazhdan constants for* $SL(3,\mathbf{Z})$, J. reine angew. Math. 413 (1991), 36–67. 111, 115

[21] P. Buser: *A note on the isoperimetric constant*, Ann. Sci. École Norm. Sup. 15 (1982), 213–230. 74

[22] P. Buser: *Geometry and spectra of compact Riemann surfaces*, Modern Birkhäuser Classics, 2010. 134, 135, 143

[23] J. Button and C. Roney-Dougal: *An explicit upper bound for the Helfgott delta in* $\mathrm{SL}(2,p)$, J. Algebra 421 (2015), 493–511. 174

[24] D. Carter and G. Keller: *Bounded elementary generation of* $\mathrm{SL}_n(\mathcal{O})$, Amer. J. Math., 105 (1983), 673–687. 109

[25] T. Ceccherini-Silberstein and M. Coornaert: *Cellular automata and groups*, Springer Monographs in Math., 2010. 71

[26] T. Ceccherini-Silberstein, F. Scarabotti and F. Tolli: *Discrete harmonic analysis: representations, number theory, expanders and the Fourier transform*, Cambridge Studies Adv. Math. 172, Cambridge Univ. Press, 2018. 11, 94

[27] I. Chavel: *Eigenvalues in Riemannian geometry*, Academic Press, 1984. 137

[28] J. Cheeger: *A lower bound for the smallest eigenvalue of the Laplacian*, in “Problems in analysis (Papers dedicated to Salomon Bochner, 1969)”, Princeton Univ. Press. 1970, 195–199. 40, 74, 141

[29] L. Clozel: *Démonstration de la conjecture τ*, Invent. math. 151 (2003), 297–328. 105

[30] D. Cox: *Primes of the form $x^2 + my^2$*, Wiley 1989. 148

[31] G. Davidoff, P. Sarnak, and A. Valette: *Elementary number theory, group theory, and Ramanujan graphs*, LMS Student Text 55, Cambridge University Press 2003. 4, 94

[32] P. Diaconis and L. Saloff-Coste: *Comparison techniques for random walk on finite groups*, Annals of Prob. 21 (1993), 2131–2156. 14, 34, 88

[33] R. Diestel: *Graph theory*, Graduate Texts Math. 173, Springer, 2017. 12

[34] J. Ellenberg, C. Hall and E. Kowalski: *Expander graphs, gonality and variation of Galois representations*, Duke Math. Journal 161 (2012), 1233–1275. 8, 134, 136, 138, 141, 144, 147, 149

[35] J. Ellenberg: *Superstrong approximation for monodromy groups*, in “Thin groups and superstrong-approximation”, MSRI Publications 61, Cambridge (2014), edited by E. Breuillard and H. Oh. 144

[36] R. Evans: *A fundamental region for Hecke’s modular group*, Journal Number Theory 5 (1973), 108–115. 210

[37] G. Faltings: *Diophantine approximation on abelian varieties*, Annals of Math. 133 (1991), 549–576. 145

[38] G. Farkas: *Brill–Noether loci and the gonality stratification of $\mathcal{M}_g$*, J. reine angew. Math. 539 (2001), 185–200. 136

[39] H. Farkas and I. Kra: *Riemann surfaces*, 2nd edition, Grad. Texts in Math. 71, Springer 1992. 134, 135, 136

[40] G. Frey: *Curves with infinitely many points of fixed degree*, Israel J. Math. 85 (1994), 79–83. 145

[41] W. Fulton and J. Harris: *Representation theory, a first course*, Universitext, Springer (1991). 214

[42] P.X. Gallagher: *The large sieve and probabilistic Galois theory*, in Proc. Sympos. Pure Math., Vol. XXIV, Amer. Math. Soc. (1973), 91–101. 24

[43] A. Gamburd: *On the spectral gap for infinite index “congruence” subgroups of* $\mathrm{SL}_2(\mathbf{Z})$, Israel J. Math. 127 (2002), 157-–200. 153

[44] A.A. Glibichuk and S.V. Konyagin: *Additive properties of product sets in fields of prime order*, in “Additive Combinatorics”, C.R.M. Proc. and Lecture Notes 43, A.M.S (2006), 279–286. 187

[45] C. Godsil: *Matchings and walks in graphs*, Journal of Graph Theory 5 (1981), 285–297. 91

[46] W.T. Gowers: *Quasirandom groups*, Comb. Probab. Comp. 17 (2008), 363–387. 171

[47] M. Gromov: *Filling Riemannian manifolds*, J. Differential Geom. 18 (1983), 1–147. 138

[48] M. Gromov and L. Guth: *Generalizations of the Kolmogorov-Barzdin embedding estimates*, Duke Math. J. 161 (2012), 2549–2603; arXiv:1103.3423. 5, 6, 117, 134, 138

[49] C. Hall, D. Puder and W. Sawin: *Ramanujan coverings of graphs*, preprint arxiv:1506.02335; abridged version in STOC 2016, 48th annual ACM SIGACT Symposium in the Theory of Computing, 533–541, 2016. 98

[50] P. de la Harpe: *Topics in Geometric Group Theory*, Chicago Lectures in Math., Univ. of Chicago Press (2000). 30, 31, 208, 209

[51] E. Hecke: *Über die Bestimmung Dirichletscher Reihen durch ihre Funktionalgleichung*, Math. Annalen 112 (1936), 664–699. 210

[52] H. Helfgott: *Growth and generation in* $\mathrm{SL}_2(\mathbf{Z}/p\mathbf{Z})$, Ann. of Math. 167 (2008), 601–623. 105, 174

[53] H.M. Hilden: *Three-fold branched coverings of* $\mathbf{S}^3$, Amer. J. Math. 98 (1976), 989–997. 139

[54] S. Hoory, N. Linial and A. Wigderson: *Expander graphs and their applications*, Bull. Amer. Math. Soc. 43 (2006), 439–561. 1, 5, 6, 94, 98, 117, 119, 124

[55] E. Hrushovski: *Stable group theory and approximate subgroups*, Journal of the A.M.S 25 (2012), 189–243. 190

[56] M. Huxley: *Exceptional eigenvalues and congruence subgroups*, in "The Selberg trace formula and related topics", p. 341– 349; edited by D. Hejhal, P. Sarnak and A. Terras, Contemporary Math. 53, A.M.S, 1986. 153

[57] K. Ireland and M. Rosen: *A Classical Introduction to Modern Number Theory*, 2nd Edition, GTM 84, Springer-Verlag (1990). 121, 122

[58] H. Iwaniec and E. Kowalski: *Analytic number theory*, Colloq. Publ. 53, Amer. Math. Soc. (2004). 121, 133, 149

[59] M. Kaluba, P. Nowak and N. Ozawa: $\mathrm{Aut}(\mathbb{F}_5)$ *has property* (T), preprint (2017), arXiv:1712.07167. 115

[60] M. Kaluba, D. Kielak and P. Nowak: *On Property* (T) *for* $\mathrm{Aut}(\mathbb{F}_n)$ *and* $\mathrm{SL}_n(\mathbf{Z})$, preprint (2018), arXiv:1812.03456. 115

[61] A. Karrass, W. Magnus and D. Solitar: *Combinatorial group theory*, 2nd Edition, Dover (1976). 208

[62] M. Kassabov: *Kazdhan constants for* $SL_n(\mathbf{Z})$, Int. J. Algebra Comput. 15 (2005), 971–995. 115

[63] M. Kassabov: *Symmetric groups and expander graphs*, Inventiones math. 170 (2007), 327–354. 106

[64] D. Kazhdan: *Connection of the dual space of a group with the structure of its closed subgroups*, Funct. Anal. Appl. 1 (1967), 63–65. 106

[65] H. Kesten: *Symmetric random walks on groups*, Trans. Amer. Math. Soc. 92 (1959), 336–354. 68

[66] A. Kontorovich: *From Apollonius to Zaremba: Local-Global phenomena in thin orbits*, Bull. AMS 50, (2013), 187–228. 7

[67] E. Kowalski: *Elliptic curves, rank in families and random matrices*, in *Ranks of Elliptic Curves and Random Matrix Theory*, edited by J. B. Conrey, D. W. Farmer, F. Mezzadri, and N. C. Snaith, LMS Lecture Note 341, (Cambridge University Press 2007). 144

[68] E. Kowalski: *Crible en expansion*, Séminaire Bourbaki, exposé 1028, November 2010, Astérisque 348, Soc. Math. France (2012), 17–64. 125

[69] E. Kowalski: *The large sieve and its applications*, Cambridge Tracts in Math., vol 175, C.U.P (2008). 125

[70] E. Kowalski: *Explicit growth and expansion for* SL_2, International Math. Res. Notices. 2012; doi:10.1093/imrn/rns214 115, 150, 161, 174

[71] E. Kowalski: *Sieve in discrete groups, especially sparse*, in "Thin groups and super-strong-approximation", MSRI Publications 61, Cambridge (2014), edited by E. Breuillard and H. Oh. 125

[72] E. Kowalski: *An introduction to the representation theory of groups*, Graduate Studies in Mathematics 155, AMS, 2014. 9, 214

[73] M. Lackenby: *Heegaard splittings, the virtually Haken conjecture and property* (τ), Invent. math. 164 (2006), 317–359. 134

[74] S. Lang: *Algebra*, 2nd edition, Addison Wesley 1984. 212, 214

[75] M. Larsen and R. Pink: *Finite subgroups of algebraic groups*, Journal of the A.M.S 24 (2011), 1105–1158. 190

[76] D. Levin, Y. Peres and E. Wilmer: *Markov chains and mixing times*, A.M.S 2009. 52, 53, 55, 82

[77] P. Li and S.T. Yau: *A new conformal invariant and its applications to the Willmore conjecture and the first eigenvalue of compact surfaces*, Invent. math. 69 (1982), 269–291. 136, 137

[78] A. Lubotzky: *Discrete groups, expanding graphs and invariant measures*, Progress in Math. 125, Birkaüser (1994). 1, 4, 71, 80, 94, 105, 117

[79] A. Lubotzky: *Expander graphs in pure and applied mathematics*, Bulletin AMS 49 (2012), 113–162. 1, 8, 117

[80] A. Lubotzky: *Ramanujan complexes and high dimensional expanders.* Japan. J. Math. 9 (2014), 137–169. 8

[81] A. Lubotzky and C. Meiri: *Sieve methods in group theory, I: powers in linear groups*, Journal A.M.S 25 (2012), 1119–1148. 7, 125

[82] A. Lubotzky, R. Phillips and P. Sarnak: *Ramanuajan graphs*, Combinatorica 8 (1988), 261–277. 94, 98

[83] A. Marcus, D.A. Spielman and N. Srivastava: *Interlacing families, I: bipartite Ramanuajan graphs of all degrees*, Ann. of Math. (2) 182 (2015), 307–325. 98, 217

[84] G. Margulis: *Explicit group theoretic constructions of combinatorial schemes and their applications for the construction of expanders and con- centrators*, J. of Problems of Information Transmission, 24 (1988), 39–46. 94, 98

[85] J. McCarthy: *Recursive functions of symbolic expressions and their computations by machines, I*, Communications of the A.C.M 3 (1960), 184–195. 119

[86] W. Bosma, J. Cannon and C. Playoust: *The Magma algebra system, I. The user language*, J. Symbolic Comput. 24 (1997), 235–265; also http://magma.maths.usyd.edu.au/magma/ 185

[87] R. Miranda: *Riemann surfaces*, Grad. Studies in Math. 5 (A.M.S), 1995. 134, 135, 136

[88] J.M. Montesinos: *Three-manifolds as 3-fold branched covers of* $\mathbf{S}^3$, Quart. J. Math. Oxford Ser. (2) 27 (1976), 85–94. 139

[89] T. Netzer and A. Thom: *Kazhdan's property* (T) *via semidefinite optimization*, Exp. Math. 24 (2015), 371–374. 115

[90] N. Nikolov and L. Pyber: *Product decompositions of quasirandom groups and a Jordan-type theorem*, J. European Math. Soc. 13 (2011), 1063–1077. 171

[91] M. Orr: *Unlikely intersections with Hecke translates of a special subvariety*, preprint (2017), arXiv:1710.04092. 146

[92] N. Ozawa: *Noncommutative real algebraic geometry of Kazhdan's property* (T), J. Inst. Math. Jussieu 15 (2016), 85–90. 115

[93] J. Pardon: *On the distorsion of knots on embedded surfaces*, Annals of Math. 174 (2011), 637–646. 6, 139

[94] G. Petridis: *New proofs of Plünnecke-type estimates for product sets in groups*, Combinatorica 32 (2012), 721–733. 201

[95] M. Pinsker: *On the complexity of a concentrator*, in "7th International Telegrafic Conference", pages 318/1–318/4, 1973. 5, 51, 94

[96] B. Poonen: *Gonality of modular curves in characteristic* p, Math. Res. Lett. 14, no. 4 (2007), 691–701. 146

[97] L. Pyber and E. Szabó: *Growth in finite simple groups of Lie type of bounded rank*, Journal A.M.S 29 (2016), 95–146. 105, 106, 174, 176, 183

[98] M. Rudnev and I Shkredov: *On growth rate in* $\mathrm{SL}_2(\mathbf{F}_p)$*, the affine group and sum-product type implications*, preprint (2018), arXiv:1812.01671. 174

[99] A. Salehi Golsefidy and P. Varjú: *Expansion in perfect groups*, G.A.F.A 22 (2012), 1832–1891, 106, 149

[100] P. Sarnak and X. Xue: *Bounds for multiplicities of automorphic representations*, Duke Math. J. 64, (1991), 207–227. 153

[101] P. Sarnak: *Some applications of modular forms*, Cambridge Tracts in Math. 99, Cambridge Univ. Press 1990. 1, 4, 94, 98, 117

[102] J-P. Serre: *A course in arithmetic*, Grad. Texts in Math. 7, Springer 1973. 121, 210

[103] J-P. Serre: *Trees*, Springer Monographs in Math., Springer (2003). 12, 13

[104] Y. Shalom: *Bounded generation and Kazhdan's property (T)*, Publ. Math I.H.É.S 90 (1999), 145–168. 108, 116

[105] J. Silverman: *The arithmetic of elliptic curves*, Grad. Texts in Math 106, Springer Verlag (1986). 144, 146

[106] R. Solovay and V. Strassen: *A fast Monte-Carlo test for primality*, SIAM J. Comput. 6 (1977), 84–85. 120

[107] M. Suzuki: *Group Theory, I*, Grundlehren math. Wiss. 247, Springer (1982). 213

[108] T. Tao: *Product set estimates for non-commutative groups*, Combinatorica 28 (2008), 547–594. 161, 198, 200, 201

[109] T. Tao: *Expansion in finite simple groups of Lie type*, Grad. Studies Math. 164, AMS (2015). 1, 4, 150

[110] T. Tao and V. Vu: *Additive combinatorics*, Cambridge Studies Adv. Math. 105, Cambridge Univ. Press (2006). 161

[111] E.C. Titchmarsh: *The theory of functions*, 2nd edition, Oxford Univ. Press, 1939. 218

[112] L. Trevisan: *Graph partitioning and expander*, 2011 lectures notes available at theory.stanford.edu/~trevisan/cs359g/index.html. 75, 76, 84

[113] L. Trevisan: *The Spectral Partitioning algorithm*, blog post available at lucatrevisan.wordpress.com/2008/05/11/the-spectral-partitioning-algorithm/ 76

[114] A. Valette: *Le problème de Kadison-Singer*, Séminaire Bourbaki, exposé 1088; Astérisque 367–368, S.M.F 2015. 102, 217

[115] L. Valiant: *What must a global theory of cortex explain?*, Current Opinion in Neurobiology 25 (2014), 15–19, dx.doi.org/10.1016/j.conb.2013.10.006 5

[116] P. Varjú: *Expansion in* $\mathrm{SL}_d(\mathcal{O}_K/I)$*, I square-free*, J. Eur. Math. Soc. (JEMS) 14 (2012), 273–305. 104

[117] L.R. Varshney, B.L. Chen, E. Paniagua, D.H. Hall and D.B. Chklovskii: *Structural Properties of the Caenorhabditis elegans Neuronal Network*, PLoS Comput Biol 7(2) (2011): e1001066; doi.org/10.1371/journal.pcbi.1001066. 2, 84

[118] J. G. White, E. Southgate, J. N. Thomson and S. Brenner: *The Structure of the Nervous System of the Nematode Caenorhabditis elegans*, Phil. Trans. R. Soc. Lond. B 314 (1986), 1–340; DOI:10.1098/rstb.1986.0056. 2

[119] W. Woess: *Random walks on infinite graphs and groups*, Cambridge Tracts in Math. 138, Cambridge Univ. Press 2000. 52, 68, 69

[120] WormBook, ed. The *C. elegans* Research Community, WormBook, www.wormbook.org. 2

[121] P. Zograf: *Small eigenvalues of automorphic Laplacians in spaces of cusp forms*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov (LOMI) 134 (1984), 157–168, English translation in Journal of Math. Sciences 36, Number 1, 106–114, DOI:10.1007/BF01104976. 146

[122] A. Żuk: *Property* (T) *and Kazhdan constants for discrete groups*, G.A.F.A 13 (2003), 643–670. 115

Index of notation

Index