Expander graphs

E. Kowalski
... 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)
# Contents

Chapter 1. Introduction 1
   1.1. Motivation 1
   1.2. Prerequisites and notation 5

Chapter 2. Graphs 7
   2.1. Graphs 7
   2.2. Metric, diameter, and so on 15
   2.3. Cayley graphs, action graphs, Schreier graphs 18

Chapter 3. Expansion in graphs 28
   3.1. Expansion in graphs 28
   3.2. Random walks 39
   3.3. The discrete Laplace operator 61
   3.4. Expansion of Cayley graphs 66
   3.5. Existence of expanders 73

Chapter 4. Applications 76
   4.1. The Barzdin-Kolmogorov graph-embedding theorem 76
   4.2. Error reduction in probabilistic algorithms 77
   4.3. Sieve methods 84
   4.4. Arithmetic geometry 84

Chapter 5. Expanders from finite linear groups 85
   5.1. Introduction 85
   5.2. Preliminaries and strategy 85
   5.3. The Bourgain-Gamburd argument 90
   5.4. Implementing the Bourgain-Gamburd argument 101
   5.5. Quasi-random groups 106
   5.6. Growth of generating subsets of $\text{SL}_2(\mathbf{F}_p)$ 109
   5.7. A survey of generalizations 128

Appendix A. Explicit multiplicative combinatorics 129
   A.1. Introduction 129
   A.2. Diagrams 130
   A.3. Statements and “Proofs” 131

Appendix B. Some group theory 136
   B.1. Free groups 136
   B.2. Properties of $\text{SL}_2$ 139

Bibliography 142
CHAPTER 1

Introduction

1.1. Motivation

This section 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.

The goal of this course 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, and then give the earliest proof of the existence of expanders, based on probabilistic methods (this existence is by no means obvious from the definition!); (2) 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 would require too much space); (3) finally, we will prove that certain concrete graphs are expanders. This is important, because the earlier probabilistic proof of existence is sometimes not sufficient, for at least two possible reasons: (1) in “applied” applications, one might need to have a physical expander at hand to use (for numerical simulation purposes, etc); (2) in “pure” applications, 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.

There are by now quite a few methods to prove that certain graphs give expanders, and some have already been presented in textbooks. We have chosen to present the combination of the work of Helfgott and Bourgain-Gamburd, from which it follows that very general families of Cayley graphs of \(\text{SL}_2(\mathbb{Z}/\ell\mathbb{Z})\) are expanders, when \(\ell\) runs over primes. This choice has many advantages: it is a very recent result, which is quite deep, and answers some questions going back more than 20 years. Moreover, the argument – though rather devilishly clever and intricate – is fundamentally elementary (e.g., it does not involve automorphic forms or Kazhdan’s Property (T)). Finally, because it is rather new, we both avoid the feeling of repeating badly some well-known presentations of the material, and we can give an introduction to the ideas surrounding this area, which has been spectacularly successful (and important) in recent years.

The remainder of this section is 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 more about, and turn to the rest of the book...

To start with, graphs seem very intuitive mathematical object. 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, the figure below\(^1\) 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 axons as edges, the brain is – as a rough first approximation – also a graph; as we will see, this graph appears as one of the first motivations that led to expander graphs.

- **[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. Genealogical trees form another example of this type, though 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 $\Gamma_n$ 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 $\Gamma_n$ be bounded by some constant $v \geq 1$ 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

\(^1\) Author: mateusch, license Creative Commons Attribution-Share Alike 3.0 Unported.
physical edge, this assumption means that increasing the number of vertices (by taking a larger graph $\Gamma_n$ from our family) will increase the cost linearly with respect to the increase in the number of vertices, since the number of edges of $\Gamma_n$ 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. 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 $B$, one obtains a disconnected graph. 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. What is by no means obvious is that they exist at all: 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. Here are some illustrations:

- [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 [39] 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 [21]). However, as pointed out recently by L. Guth, an earlier work of Barzdin and Kolmogorov [1] turns out to contain 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 their 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 $\mathbb{R}^3$ (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 $\Gamma$ in $\mathbb{R}^3$ 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, any non-adjacent edges must also be separated at least by such a distance. Barzdin and Kolmogorov first show, constructively, 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 3-regular expander can not 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  

---

2 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...
best possible. We will prove a variant of this fact in Section 4.1; see also [1] for the original paper, and the recent work of Gromov and Guth [18, §1.1] for a modern re-interpretation and many generalizations.

- **[Gromov–Guth; Knot distorsion]** A remarkable application of expander graphs to the construction of “complicated” knots was uncovered by Gromov and Guth [18, §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 \( \gamma : [0, 1] \rightarrow \mathbb{R}^3 \), which is injective on \([0, 1[\) and satisfies \( \gamma(0) = \gamma(1) \). Gromov introduced an invariant, called the **distorsion**, to measure the complexity of a knot: it compares the distance between points of \( \gamma \) when computed by following the knot itself, and by seeing the points as being in \( \mathbb{R}^3 \), and is given formally by

\[
\text{dist}(\gamma) = \sup_{0 \leq s \neq t \leq 1} \frac{d_{\gamma}(\gamma(s), \gamma(t))}{\|\gamma(s) - \gamma(t)\|_1},
\]

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

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

\[
\text{idist}(\gamma) = \inf_{\Phi} \text{dist}(\phi \circ \gamma),
\]

where \( \Phi \) runs over all diffeomorphisms of \( \mathbb{R}^3 \) (so that \( \phi \circ \gamma \) remains a knot, which is intuitively equivalent to \( \gamma \), as it can be obtained from it by smooth deformation). As Gromov and Guth explains it is very hard to construct knots with \( \text{idist}(\gamma) \) arbitrarily large. The first examples were found by Pardon [36], and they are explicit but very special. Gromov and Guth show that sequences of knots \( \gamma_n \) constructed using special manifolds which are related to certain types
of expander graphs always have a large distortion, and in fact that if $\gamma_n$ comes from a family $(\Gamma_n)$ of graphs, we have
\[ \text{idist}(\gamma_n) \geq c|\Gamma_n|, \]
for some constant $c > 0$, so that the distortion grows linearly with the number of vertices of the graphs, which is, as they show, as fast as the distortion of these knots can grow.

It is important to notice in this example, in contrast with the previous 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 5.

- [Bourgain–Gamburd–Sarnak; Sieve] See [5].
- [Specializations of arithmetic properties in families] See [13].

A more detailed discussion of these and other applications can be found in A. Lubotzky’s survey [32]; we will discuss briefly some details of the last two in Chapter 4.

1.2. Prerequisites and notation

We only require basic linear algebra (including finite-dimensional Hilbert spaces), calculus and probability, though a number of concepts are presented with terminology influenced by functional analysis ($L^2$-spaces in particular).

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. 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!). The subgroup $[G, G]$ is normal in $G$, and the quotient group $G/[G, G]$ is abelian; it is called the abelianization of $G$.

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

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
\[ \langle \alpha v, w \rangle = \alpha \langle v, w \rangle, \quad \langle v, \alpha w \rangle = \bar{\alpha} \langle v, w \rangle, \]
for two vectors $v, w$ and a scalar $\alpha \in \mathbb{C}$.

Acknowledgments. The first version of this book was prepared for a course “Expander graphs” that I taught at ETH Zürich during the Fall Semester 2011. Thanks are obviously due to 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 5.
The text was finalized (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.
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 intuitive 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.

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

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

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 $\Gamma$, the elements of $\text{ep}(\alpha)$ are called extremities of $\alpha$. If $\alpha \neq \beta$ are distinct edges of $\Gamma$, they are called adjacent at a vertex $x \in V$ if $x \in \text{ep}(\alpha) \cap \text{ep}(\beta)$ is a common extremity.

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

A graph is finite when both $V$ and $E$ are finite.

**Remark 2.1.2.** (1) The intuition should be clear, as the terminology indicates: to bring a graph (say, one drawn on paper) to this form, one takes as set of edges the “physical” ones, and one defines $\text{ep}(\alpha)$ to be the set of extremities of such an edge. This allows loops, which are edges where $\text{ep}(\alpha) = \{x\}$ 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 $\text{ep}(\alpha_1) = \text{ep}(\alpha_2) = \{x, y\}$.

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

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

$$\text{ep}(1) = \{a, b\}, \quad \text{ep}(2) = \{b, c\}, \quad \text{ep}(3) = \{c, d\}, \quad \text{ep}(4) = \{a, d\},$$

$$\text{ep}(5) = \{a, c\}, \quad \text{ep}(6) = \{b, d\}, \quad \text{ep}(7) = \{c, d\},$$

and check that it can be represented as
As in this figure, it is not always possible to draw the edges without some overlap. However, for any finite graph, it is possible to “draw” it in \( \mathbb{R}^3 \) without overlap. This should be fairly clear intuitively, and the reader should attempt to see what is involved in a rigorous proof. (Basically, \( \mathbb{R}^3 \) minus a finite number of continuous, compact, arcs, seen as images of maps \( \gamma : [0, 1] \rightarrow \mathbb{R}^3 \), is path-connected.)

(2) We will often write \( \text{ep}(\alpha) = \{x, y\} \), 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. To give a more geometric flavor to this formula, we might write \( x \xleftarrow{\alpha} y \).

(3) If \( \Gamma \) has no loops (which means every set of endpoints \( \text{ep}(\alpha) \) contains two elements) and no multiple edges (so that \( \text{ep} \) 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 \( (y, x) \in R \), and such that \( (x, x) \notin R \) for all \( x \in V \). This is a more common way of “coding” simple graphs.

(4) By convention, for a graph \( \Gamma \), we write \( |\Gamma| = |V| \): the “size” of \( \Gamma \) 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 \( V_\Gamma \) for the set of vertices (resp. \( E_\Gamma \) for the set of edges) of a graph \( \Gamma \).

**Exercise 2.1.3** (Number of edges vs. number of vertices). Show that if \( \Gamma = (V, E, \text{ep}) \) is a finite graph, we have
\[
\sum_{x \in V} \text{val}(x) = 2|E_2| + |E_1|
\]
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 more efficiently a finite graph, one can use its **adjacency matrix**:

**Definition 2.1.4** (Adjacency matrix). Let \( \Gamma \) be a finite graph. The **adjacency matrix** \( A_\Gamma = (a(x, y)) \) is the matrix with rows and columns indexed by \( V_\Gamma \) and with \( a(x, y) \) equal to the number of edges with extremities \( (x, y) \), formally
\[
a(x, y) = |\{ \alpha \in E_\Gamma \mid \text{ep}(\alpha) = \{x, y\} \}|.
\]

Note that the adjacency matrix is always symmetric, 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 with 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 (for instance!)
\[
E = \{(x, y, i) \in V \times V \times \mathbb{Z} \mid a_{x,y} \neq 0 \text{ and } 1 \leq i \leq a_{x,y}\}
\]
with
\[
\text{ep}(x, y, i) = \{x, y\}
\]
for all \((x, y, i) \in E\).

**Exercise 2.1.5.** Devise at least one other way of formalizing the same class of graphs.

**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 definitions.

1. **[Cycle]** Let \(m \geq 1\) be an integer. The \(m\)-cycle \(C_m\) is the graph with vertices \(V_m = \mathbb{Z}/m\mathbb{Z}\), edges \(E_m = \mathbb{Z}/m\mathbb{Z}\), and endpoint map given by
   \[\text{ep}(i) = \{i, i + 1\}\]
   for \(i \in \mathbb{Z}/m\mathbb{Z}\). 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 \(m = 1\), \(m = 5\) and \(m = 2\):

   ![Graphs for m = 1, 5, 2]

2. **[Path]** Let \(m \geq 0\) be an integer. The path of length \(m\), denoted \(P_m\), is the graph with \(m + 1\) vertices, \(V_m = \{0, \ldots, m\}\), \(m\) edges, \(E_m = \{1, \ldots, m\}\), and \(\text{ep}(i) = \{i - 1, i\}\) for \(1 \leq i \leq m\). A path of length 0 is a graph with a single vertex and no edges. Here is the path of length 4:

   ![Path of length 4]

3. **[Complete graph]** Let again \(m \geq 1\) 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 \(\text{ep}((x, y)) = \{x, y\}\). In other words, each pair of distinct vertices is joined by (exactly) one edge. Here is the complete graph \(K_5\):

   ![Complete graph K_5]

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 \(m \geq 2\); \(P_0\) is 0-regular, \(P_1\) is 1-regular (but \(P_k\) is not regular for \(k \geq 2\)); \(K_m\) is \((m - 1)\)-regular for all \(m \geq 1\).

4. **[A Cayley graph]** Our last sequence of examples is less obvious, but it illustrates the type of graphs that will become 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 [12], we fix \(n \geq 3\) and take as vertex set \(V_n\) all the possible arrangements of a deck \(D_n\) of \(n\) cards (ordered from top to bottom, 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
Thus, by definition, $G_n$ is a 3-regular graph for each $n \geq 3$, with $n!$ vertices. We have in Figure 1 an illustration of $G_3$, with the deck $D_3 = \{a, b, c\}$, and in Figure 2 one of $G_4$, with deck $D_4 = \{a, b, c, d\}$ (it is by far the most complicated graph we will draw...).

Having selected a specific, maybe artificial-looking, type of coding for graphs at least has 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 of those edges, and other sets could be equally suitable. This leads to the definition of isomorphism of graphs or, more generally, to maps of graphs:

**Definition 2.1.7 (Maps of graphs).** Let $\Gamma_1$ and $\Gamma_2$ be graphs. A map, or graph map, from $\Gamma_1$ to $\Gamma_2$ is a pair $(f, f_*)$ where

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

is a map between the vertex sets and

$$f_* : E_{\Gamma_1} \to E_{\Gamma_2}$$
is a map between the edges, such that

\[(2.1) \quad \text{ep}(f_\ast(\alpha)) = f(\text{ep}(\alpha))\]

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

We can spell out the meaning of the condition \((2.1)\): if \(x \leftrightarrow y\) (i.e., \(\text{ep}(\alpha) = \{x,y\}\); recall that \(x = y\) is permitted), then we have

\[f(x) \xleftarrow{\{\varepsilon\}} f(y).\]

Intuitively, in addition to specifying where each vertex goes, a map of graphs must say where each edge goes, as an edge, and not only specify what the endpoints of the image edge are. If the graph is simple, of course, the companion edge-map \(f_\ast\) is uniquely specified by \(f\) itself: if there is an edge between \(x\) and \(y\) (necessarily unique), there must also be an edge between \(f(x)\) and \(f(y)\).

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

**Definition 2.1.8.**

1. Let \(\Gamma\) be a graph. The identity map \(\Gamma \to \Gamma\) of \(\Gamma\) is the pair \((\text{Id}_V, \text{Id}_E)\), and is denoted \(\text{Id}_\Gamma\).
2. For any graphs \(\Gamma_1, \Gamma_2, \Gamma_3\) and maps \(\Gamma_1 \xrightarrow{(f,f_\ast)} \Gamma_2 \xrightarrow{(g,g_\ast)} \Gamma_3\), the composite map is defined by the pair \((g \circ f, g_\ast \circ f_\ast)\). 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 \(f : \Gamma_1 \to \Gamma_2\), we have

\[f \circ \text{Id}_{\Gamma_1} = f, \quad \text{Id}_{\Gamma_2} \circ f = f.\]

**Remark 2.1.9.** In the language of categories, this says that there is a category of graphs where objects are graphs, and arrows are graph maps. We will not use deeply this language, but we will use remarks (like this one) to indicate the interpretation of certain facts in these terms.

**Exercise 2.1.10.** 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. (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 2.1.11.** Let \(\Gamma = (V,E,\text{ep})\) be an arbitrary graph. We can associate to it a simple graph \(\Gamma^s = (V^s, E^s, \text{ep})\) 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

\[(2.2) \quad E^s = \{\{x,y\} \subset V \mid x \neq y\text{ and there exists }\alpha \in E\text{ with }\text{ep}(\alpha) = \{x,y\}\},\]

with \(\text{ep}(\{x,y\}) = \{x,y\}\). There is always a graph map \(\Gamma^s \to \Gamma\) which is the identity on vertices, and maps an edge in \(\Gamma^s\) to some (arbitrarily chosen) edge in \(\Gamma\) between the same extremities. This map is not uniquely defined in general (if multiple edges exist). On the other hand, if \(\Gamma\) has no loops, there is a canonical map \(\Gamma \to \Gamma^s\), which is again
the identity on the vertices, where any edge is mapped to the unique edge in \( \Gamma^* \) with the same extremities. (If \( \Gamma \) has a loop, there is no map \( \Gamma \to \Gamma^* \) at all.)

More definitions:

**Definition 2.1.12** (Isomorphism, automorphism, embedding). (1) A graph map \( f : \Gamma_1 \to \Gamma_2 \) is an isomorphism with inverse \( g \) if and only if \( f \circ g = \text{Id}_{\Gamma_2} \) and \( g \circ f = \text{Id}_{\Gamma_1} \). If \( \Gamma = \Gamma_1 = \Gamma_2 \), then \( f \) is called an automorphism of \( \Gamma \).

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

(3) An embedding \( \Gamma_1 \hookrightarrow \Gamma_2 \) is a graph map \((f, f_{\ast})\) such that \( f \) and \( f_{\ast} \) are both injective.

**Remark 2.1.13**. 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 and their edges are both in bijection, and those bijections are “compatible”. This corresponds to the intuitive idea of changing these sets to different codings while respecting the graph structure.

Similarly, to say that \((f, f_{\ast})\) is an embedding means that the vertices of \( \Gamma_1 \) can be identified with a subset of those of \( \Gamma_2 \), and that the edges between any two vertices in \( \Gamma_1 \) are then a subset of those in \( \Gamma_2 \) (where there could be more edges, of course.)

**Example 2.1.14.** (1) If \( \Gamma^* \) is the simple graph associated to a graph \( \Gamma \), as in Example 2.1.11, the maps \( \Gamma^* \to \Gamma \) described in this example are all embeddings of \( \Gamma^* \) in \( \Gamma \).

Moreover, if \( \Gamma \) is itself a simple graph, there is a canonical isomorphism \( \Gamma \sim \Gamma^* \) (although the sets of edges \( E \) and \( E^* \) defined in (2.2) might not be identical) given by the identity on vertices and

\[ f_{\ast}(\alpha) = \text{ep}(\alpha) \in E^* \]

for \( \alpha \in E \).

(2) The path \( P_k \), for \( k \geq 1 \), has a non-trivial automorphism, which is intuitively given by “reversing the path”, and can be defined formally by

\[ f(i) = m - i, \quad f_{\ast}(j) = m + 1 - j \]

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

\[ \text{ep}(f_{\ast}(j)) = \text{ep}(m + 1 - j) = \{m - j, m - j - 1\} = f(\{j, j + 1\}) = f(\text{ep}(j)), \]

and since \( f \) and \( f_{\ast} \) are both involutions, \((f, f_{\ast})\) 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) Let \( \Gamma = (V, E, \text{ep}) \) 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 \( \text{ep}(\alpha) \subset V' \) for any \( \alpha \in E' \)), the pair of inclusions \((V' \hookrightarrow V, E' \hookrightarrow E)\) is an embedding of the graph \((V', E', \text{ep})\) inside \((V, E, \text{ep})\). We then say that \((V', E', \text{ep})\) is a subgraph of \( \Gamma \).

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

\[ E' = \{ \alpha \in E \mid \text{ep}(\alpha) \subset V' \}, \]

we say that \((V', E', \text{ep})\) is a full subgraph of \( \Gamma \). Such subgraphs are therefore in one-to-one correspondence with subsets of \( V \).
Embeddings or other graph maps can frequently be used to define invariants and distinguish special families of graphs. Here is an important example:

**Definition 2.1.15 (Girth).** Let $\Gamma = (V,E,ep)$ be a graph.

1. For $m \geq 1$, a cycle of length $m$ in $\Gamma$ is an embedding $C_m \to \Gamma$.

2. The girth $girth(\Gamma)$ of $\Gamma$ is the smallest integer $m \geq 1$ such that there exists at least one cycle of length $m$ in $\Gamma$, or $+\infty$ if no cycle exists at all in $\Gamma$.

**Example 2.1.16.** The girth of $C_m$ itself is of course equal to $m$. Moreover, $\Gamma$ has girth 1 if and only if $\Gamma$ 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 $\Gamma$, i.e., three distinct vertices $x_1, x_2, x_3$ and three edges $\alpha_1, \alpha_2$ and $\alpha_3$ with $\alpha_1$ joining $x_1$ and $x_2$, $\alpha_2$ joining $x_2$ and $x_3$ and finally $\alpha_3$ 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 3 for $m \geq 3$.

Here is an example of graph with girth 5.

**Example 2.1.17 (Trees and forests).** Graphs with infinite girth have a name:

**Definition 2.1.18 (Forests (and trees)).** A graph $\Gamma$ with infinite girth (i.e., there is no embedding $C_m \to \Gamma$, for any $m \geq 1$) is called a forest. Anticipating Definition 2.2.1 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 $k \geq 1$. Here are some more interesting examples. Fix some integers $d \geq 2$ and $k \geq 1$. The finite 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 $\leq k$ in the alphabet $A = \{1, \ldots, d\}$ with no letter repeated twice in a row (including the empty word), i.e.

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

with edges between “neighbouring” 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.

We have in Figure 3 pictures of $T_{2,3}$ and $T_{4,2}$, with the vertices labelled with the corresponding words, which should clarify the matter.

One can extend this construction to infinite depth: the $d$-regular tree $T_d$, for $d \geq 2$, is the infinite graph with vertices given by all words of length $\geq 0$ in the alphabet $\{1, \ldots, d\}$, and with the edges described above.
The reader is invited to check that the number of vertices and edges of $T_{d,k}$ are given by

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

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

$$\text{ep}(\alpha) \cap V_0 \neq \emptyset, \quad \text{ep}(\alpha) \cap V_1 \neq \emptyset$$

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 2.1.20.** The complete bipartite graph $K_{m,n}$ with $m \geq 1$ inputs and $n \geq 1$ outputs is the bipartite graph defined by the vertices

$$V_0 = \mathbb{Z}/m\mathbb{Z}, \quad V_1 = \mathbb{Z}/n\mathbb{Z}, \quad V = V_0 \cup V_1$$

(a disjoint union) and edges

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

with the endpoint map $\text{ep}(\{x_0, x_1\}) = \{x_0, x_1\}$.

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 $m, n \geq 2$, while it is infinite for $m = 1$ or $n = 1$.

We now have an easy proposition:

**Proposition 2.1.21 (Bipartiteness criterion).** A graph $\Gamma = (V, E, \text{ep})$ is bipartite if and only if there exists a surjective graph map $\Gamma \rightarrow P_1$, where $P_1$ is the path of length 1:

**Proof.** We denote by $\{0, 1\}$ the two vertices of $P_1$ and by $\alpha_0$ its unique edge. If $\Gamma$ is bipartite, with a partition $V = V_0 \cup V_1$ in inputs and outputs, we can define rather
obviously a surjective 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 \( f_\ast(\alpha) = \alpha_0 \) for all \( \alpha \in E_\Gamma \). This is indeed a graph map because

\[
\text{ep}(f_\ast(\alpha)) = \{0, 1\} = f(\text{ep}(\alpha))
\]

since, by definition, \( \alpha \) has one extremity in \( V_0 \) and one in \( V_1 \).

Conversely, let \((f, f_\ast)\) be a surjective graph map from an arbitrary graph \( \Gamma \) to \( P_1 \). Defining \( V_0 = f^{-1}(0), V_1 = f^{-1}(1) \), we obtain of course a partition of \( V \), and the surjectivity implies that neither \( V_0 \) nor \( V_1 \) is empty. Then we consider an arbitrary edge \( \alpha \in E_\Gamma \). Since \( f_\ast(\alpha) \) has no choice other than to be equal to the unique edge \( \alpha_0 \), we get

\[
\{0, 1\} = \text{ep}(f_\ast(\alpha)) = f(\text{ep}(\alpha)),
\]

which is only possible if the extremities of \( \alpha \) 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 \( \Gamma \) into a bipartite graph. \( \square \)

### 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 \( \Gamma = (V, E, \text{ep}) \) be a graph.

1. A *path of length* \( k \geq 0 \) in \( \Gamma \) is a graph map \( P_k \rightarrow \Gamma \), i.e., an ordered sequence \( (x_0, \ldots, x_k) \) of vertices of \( \Gamma \), and an ordered sequence \( (\alpha_1, \ldots, \alpha_k) \) of edges of \( \Gamma \) such that

\[
\text{ep}(\alpha_i) = \{x_{i-1}, x_i\}
\]

for \( 1 \leq i \leq k \). If \( k \geq 1 \), the *extremities* of the path \( \gamma \) 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 \( \gamma \) is a path between \( x \) and \( y \), and one writes \( \ell(\gamma) = k \) for its length.

2. For any two vertices \( x, y \in V \), the *distance on* \( \Gamma \) *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.,

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

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.

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

**Proposition 2.2.2.** (1) If \( \Gamma = (V, E, \text{ep}) \) is a connected graph, the distance function \( d_\Gamma \) is a metric on \( V \), i.e., it is non-negative and satisfies

\[
\begin{align*}
d_\Gamma(x, y) &= d_\Gamma(y, x), \\
d_\Gamma(x, y) &= 0 \text{ if and only if } x = y, \\
d_\Gamma(x, y) &\leq d_\Gamma(x, z) + d_\Gamma(z, y)
\end{align*}
\]
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 \( \Gamma \) corresponding to an equivalence class \( V' \subset V \) is a connected graph such that the distance \( d_{\Gamma'} \) is the restriction of \( d_\Gamma \) 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 \( \Gamma \).

**Proof.** (1) is intuitively clear: the symmetry is because a path \( P_k \) can be taken in “reversed” order (this is the content of Example 2.1.14, (2); note in passing that this depends on the fact that the edges are unoriented), moreover a path \( \gamma \) of length 0 is \( P_0 \to \Gamma \) which has only one extremity, and finally the triangle inequality comes from the possibility of concatenating a path of length \( k_1 \) between \( x \) and \( z \) with one of length \( k_2 \) between \( z \) and \( y \) to obtain one of length \( k_1 + k_2 \) between \( x \) and \( y \).

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 extremity in \( V' \): if \( \text{ep}(\alpha) = \{x, y\} \) with \( x \in V' \), then the edge \( \alpha \) shows (by definition) that \( d_\Gamma(x, y) \leq 1 \), so that \( y \sim x \) is also in \( V' \). Thus, if \( E' \) is the set of edges with an extremity in \( V' \), \((V', E', \text{ep})\) is a full subgraph of \( \Gamma \). 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., \((V', E', \text{ep})\) is connected.

Moreover, since one can not connect elements of \( V' \) in \( \Gamma \) using edges others than those in \( E' \), we also see that the distance in \( \Gamma' \) is the restriction to \( V' \times V' \) of \( d_\Gamma \). \( \Box \)

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 2.2.3 (Diameter of a graph).** Let \( \Gamma = (V, E, \text{ep}) \) be a graph. The diameter of \( \Gamma \), denoted \( \text{diam}(\Gamma) \) is the largest distance between two vertices in \( \Gamma \), i.e., we have
\[
\text{diam}(\Gamma) = \sup_{x,y \in V} d_\Gamma(x, y) \in \{0, 1, 2, \ldots\} \cup +\infty.
\]

**Example 2.2.4.** If \( \Gamma \) is a finite, connected, graph, its diameter will be finite. One of the questions in this book is: given certain finite graphs, which are known to be connected, what can one say about their diameters? In particular, is this diameter relatively small, compared with the number of vertices?

We can here 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 \( m \geq 2 \) (\( K_1 = P_0 \) has diameter 0);
- For \( K_{m,n} \), the diameter is 2 if either \( m \) or \( n \) is \( \geq 2 \), while \( \text{diam}(K_{1,1}) = 1 \);
- The diameter of the cycle \( C_m \) is given by
\[
\text{diam}(C_m) = \begin{cases} 
\frac{m}{2} & \text{if } m \text{ is even} \\
\frac{m-1}{2} & \text{if } m \text{ is odd}.
\end{cases}
\]

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 \( \text{diam}(G_n) \ll n^2 \). Since \( |G_n| = n! \), this means that
\[
\text{diam}(G_n) \ll (\log |G_n|)^2,
\]
hence the diameter is here rather small compared with the number of vertices.

**Exercise 2.2.5.** 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 $d \geq 2$ and $k \geq 0$ is $2k$, and is achieved by the distance between any two vertices labelled with words of (maximal) length $k$.

2. Show that if $T$ is a tree, the image of all paths of length $d_T(x, y)$ between two vertices $x$ and $y$ of $T$ is the same, i.e., the set of vertices on such a path is independent of the “parametrization” of the latter.

3. If $T = T_{d,k}$ with “root” vertex $x_0 = \emptyset$ and $0 \leq j \leq k$, show that
   \[ V' = \{ x \in V_T \mid d_T(x_0, x) \leq j \} \]
   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
   \[ V'' = \{ y \in V_T \mid d_T(y, x_0) \geq d_T(y, x) \} \]
   induces a full subgraph $T''$ of $T$ which is also a tree.

5. Let $\Gamma = (V, E, ep)$ be any graph with girth $\ell \geq 1$, and let $x_0 \in V$. Show that the subgraph of $\Gamma$ induced by
   \[ V' = \{ x \in V \mid d_\Gamma(x_0, x) < \frac{\ell}{2} \} \]
   is a tree.

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

**Proposition 2.2.6.** Let $\Gamma_1$ and $\Gamma_2$ be graphs, and let $f : \Gamma_1 \to \Gamma_2$ be a graph map. Then $f$ is always distance-decreasing, i.e., we have
\begin{equation}
  d_{\Gamma_2}(f(x), f(y)) \leq d_{\Gamma_1}(x, y)
\end{equation}
for any $x, y \in \Gamma_1$. In particular, if $f$ is surjective on vertices, the diameter of $\Gamma_2$ is at most that of $\Gamma_1$, and if $f$ is an isomorphism, it is isometric.

**Proof.** This inequality follows from the observation that any path
\[ \gamma : P_k \to \Gamma_1 \]
in $\Gamma_1$, between $x, y \in \Gamma_1$, gives a corresponding one $f \circ \gamma$ in $\Gamma_2$, of the same length, between $f(x)$ and $f(y)$. The distance in $\Gamma_2$ between $f(x)$ and $f(y)$ is computed using a minimum over a set which contains these particular paths, and that implies that $d_{\Gamma_2}(f(x), f(y)) \leq d_{\Gamma_1}(x, y)$.

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
\[ d_{\Gamma_2}(z, w) = d_{\Gamma_2}(f(x), f(y)) \leq d_{\Gamma_1}(x, y) \leq \text{diam}(\Gamma_1), \]
which gives $\text{diam}(\Gamma_2) \leq \text{diam}(\Gamma_1)$.

**Exercise 2.2.7.** Here is an application of graphs and connected components to group theory, due to Bauer and Knutson (see [14, Lemma, p. 98]). Let $k \geq 2$ be an integer, $G = S_k$ 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 $H = G$. 

17
Let $\Gamma = (V, E, ep)$ 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 $\Gamma$ is a complete graph.
2. Show that it is enough to show that $\Gamma$ 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 $\Gamma$ by automorphisms. Show then that $G$ acts transitively on the set of all connected components of $\Gamma$. 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 $\Gamma$, and conclude from this.

**Exercise 2.2.8.** Let $\Gamma = (V, E, ep)$ be a graph, and let $\Gamma^s$ be the associated simple graphs. Show that $d_{\Gamma^s}(x, y) = d_{\Gamma}(x, y)$ for all $x, y \in V$.

**Exercise 2.2.9 (Uniqueness of bipartite decompositions).** (1) Let $\Gamma = (V, E, ep)$ be a connected bipartite graph with a bipartite decomposition $V = V_0 \cup V_1$. If $x_0 \in V_0$, show that
\begin{equation}
V_0 = \{ y \in V \mid \text{there is a path of even length joining } x \text{ to } y \}. \tag{2.4}
\end{equation}

(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 $\Gamma$ be an arbitrary connected graph, and let $W$ be the right-hand side in (2.4). What is $W$ when $\Gamma$ is not bipartite?

(4) Show that a forest is always bipartite.

(5) Show that if a $\Gamma$ is finite and not bipartite, its girth is finite. In fact, show that $\text{girth}(\Gamma) \leq 2 \text{diam}(\Gamma) + 1$, and that this is best possible.

### 2.3. Cayley graphs, action graphs, Schreier graphs

We can now define the most important type of graphs in the context of this text: the **Cayley graphs**, which are used to get a geometric vision of certain groups.

**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
\[ E = \{ \{g, gs\} \mid g \in G, \quad s \in S \} \subset V^{(2)} \]
and $ep$ is the identity 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 has no multiple edges, but it may have loops. In fact, this happens if and only if $1 \in S$.

**Example 2.3.2.** (1) For $m \geq 3$, the cycle $C_m$ can be seen as (i.e., it is isomorphic to) the Cayley graph $\mathcal{C}(\mathbb{Z}/m\mathbb{Z}, \{\pm 1\})$, as the reader is invited to check. For $m = 2$, this is not the case (because in that case $1 = -1$), and $\mathcal{C}(\mathbb{Z}/2\mathbb{Z}, \{1\}) \simeq P_2$. 18
Similarly, for all $m \geq 2$, the complete graph $K_m$ is also isomorphic to a Cayley graph of $\mathbb{Z}/m\mathbb{Z}$, but with respect to $S = \mathbb{Z}/m\mathbb{Z} - \{0\}$. This already shows that Cayley graphs can look quite different for the same group $G$!

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

![Cayley Graph Example](image)

Note that this graph is not connected.

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

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

\begin{equation}
S_n = \{\tau, (1 \ 2 \ \cdots \ n)^{\pm 1}\}.
\end{equation}

Indeed, if we use the deck of card $D_n = \{1, \ldots, n\}$, 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 often refer to the graphs $G_n$ simply as Cayley graphs.

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 $S_n$.

**Remark 2.3.3.** There is a slightly different definition of Cayley graphs, which is sometimes better than the one we used. However, this second definition only produces a different graph when some of the generators are non-trivial involutions, i.e., when some $s \in S$, $s \neq 1$, satisfies $s^2 = 1$. Here one takes the same vertex set $V = G$, but the edge set is

$$E = \{(g, gs) \mid g \in G \ s \in S\}/\sim$$

where the equivalence relation $\sim$ is defined by

$$(g, gs) \sim (h, ht)$$

if and only if $\{g, gs\} = \{h, ht\}$ but $s$ is not a non-trivial involution. The end points of $(g, gs)$ are $g$ and $gs$, as in the case of the previous definition. It is also clear that if $S$ contains no non-trivial involution, then the equivalence relation identifies $(g, gs)$ with $(gs, g)$ and the graph defined in this way is isomorphic to $\mathcal{C}(G, S)$.

However, if $g \in V$, and $s$ is a non-trivial involution, there are two different edges $(g, gs)$ and $(gs, g) = ((gs), (gs)s)$ with endpoints $g$ and $gs$ in this new graph. In particular, for
If \( G = \mathbb{Z}/2\mathbb{Z} \) and \( S = \{1\} \), the Cayley graph “of the second kind” is isomorphic to the cycle \( C_2 \) (compare with the first example above, where we had \( m \geq 3 \) to interpret \( C_m \) as a Cayley graph).

Similarly, for \( G = \mathfrak{S}_n \) and the generating set \( S_n \) of (4), this second definition does not give Example 2.1.6, but graphs where single edges corresponding to the generator \( \tau \) are doubled.

For the purposes of studying expander graphs, we will see that both definitions are equivalent. Since Definition 2.3.1 has some other slight advantages, we will use, and only make passing references to this second 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 \( G \), and let \( \Gamma = \mathcal{C}(G,S) \) be the corresponding Cayley graph.

1. The Cayley graph \( \Gamma \) is connected if and only if \( S \) is a generating set of \( G \).
2. Denote \( \|x\|_S = d_\Gamma(1, x) \). Then the distance \( d_\Gamma \) satisfies
   \[
   d_\Gamma(x, y) = \|x^{-1}y\|_S,
   \]
   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
   \[
   \|x\|_S = \min\{k \geq 0 \mid x = s_1 \cdots s_k \text{ for some } s_i \in S\},
   \]
   which is called the word length of \( x \) with respect to \( S \).

**Proof.** (1) Should be intuitively clear, since paths in \( \mathcal{C}(G,S) \) joint 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 \( \gamma : P_k \rightarrow \Gamma \) be a path between 1 and \( x \) (of some arbitrary length). If \( x_i \) is the element \( \gamma(i) \in G \) for \( 0 \leq i \leq k - 1 \), we have \( x_0 = 1 \) and \( x_k = x \), and by definition of the edges in \( \Gamma \), there exists \( s_i \in S \) such that
   \[
   \ep(f_*(i)) = \{x_{i-1}, x_{i-1}s_i\} = \{x_{i-1}, x_i\}
   \]
   for \( 1 \leq i \leq k \), i.e., \( x_i = x_{i-1}s_i \). By induction this gives
   \[
   x = x_k = x_{k-1}s_k = \cdots = s_1s_\cdots s_k
   \]
   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 \( k \geq 0 \) and elements \( s_i \in S \), \( 1 \leq i \leq k \), such that
   \[
   x^{-1}y = s_1 \cdots s_k,
   \]
   and then there is a path \( \gamma : P_k \rightarrow \Gamma \) defined by
   \[
   f(0) = x, \quad f(i) = xs_1 \cdots s_i, \quad 1 \leq i \leq k,
   \]
   \[
   f_*(i) = \{xs_0 \cdots s_{i-1}, xs_0 \cdots s_i\}, \quad 1 \leq i \leq k,
   \]
   which links \( f(0) = x \) to \( f(k) = y \).
The formulas (2.6) and (2.7) are implicit in what was done before: given \( x, y \in G \), there is for any \( k \geq 0 \) a bijection, which we just constructed, between paths \( \gamma : P_k \rightarrow \Gamma \) between \( x \) and \( y \), and \( k \)-tuples \((s_1, \ldots, s_k) \in S^k\) such that
\[
y = x s_1 s_2 \cdots s_k.
\]
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
\[
dr(x, y) = \|x^{-1} y\|_S,
\]
proving (2.6).

**Exercise 2.3.5.** Prove that the set \( S_n \) given by (2.5) generates \( S_n \), and hence that the graphs \( G_n \) of Example 2.3.2, (3), are all connected. In fact, show that
\[
\text{diam}(G_n) \leq C n^2
\]
for some constant \( C > 0 \) and all \( n \geq 3 \). [Hint: This is a fairly classic exercise. As described by Diaconis and Saloff-Coste [12, §3, Ex. 1], it can be convenient to think of this in turns of card shuffling.]

Cayley graphs do not only give a graphical or 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
\[
(f, f_*) : \mathcal{C}(G, S) \rightarrow \mathcal{C}(H, f(S))
\]
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.) 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
\[
\varepsilon : G \rightarrow \{\pm 1\}
\]
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 2.1.21, the proof is simple enough to spell out in full.

**Proof.** First of all, suppose \( \varepsilon \) 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 \( \varepsilon \) 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 \( G \rightarrow \{\pm 1\} \) such that \( \varepsilon(s) = -1 \) for \( s \in S \).

Certainly, \( \varepsilon \) 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 \( \varepsilon \) 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 \( \gamma \) 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 \( \{s_1, s_1s_2\} \), with one extremity in \( V_1 \), implies that \( s_1s_2 \in V_0 \), and then similarly \( s_1s_2s_3 \in V_1 \), and so on: by induction, \( s_1 \cdots s_k \) 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.  \( \square \)

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

(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 \( \mathcal{C}(G, S) \) 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 3.4, Cayley graphs are in some ways much 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 $S \subseteq G$ 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

$$f_g(x) = gx, \quad f_g, \{\{x, xs\}\} = \{gx, gxs\}$$

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 is a simple corollary, which expresses the girth of a Cayley graph in group-theoretic terms. It is one of the few places in this book where our definition of Cayley graphs imposes a restriction.

Corollary 2.3.9. Let $G$ be a group and $S \subseteq G$ a symmetric set that contains no non-trivial involution. Let $\Gamma = \mathcal{C}(G, S)$ be the corresponding Cayley graph. The girth of $\Gamma$ is then equal to the length of the shortest non-trivial relation among the elements of $S$, namely $\text{girth}(\Gamma)$ is the smallest $m \geq 1$ for which there exist

$$s_1, \ldots, s_m$$

in $S$, with $s_i s_{i+1} \neq 1$, such that

$$s_1 s_2 \cdots s_m = 1.$$

In particular, if $G$ is finite, the girth of the Cayley graph $\Gamma$ is finite.

Proof. First of all, by composing with a suitable automorphism of $\Gamma$, 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 by $k$ the girth of $\Gamma$.

(1) If $k$ is finite, and $\gamma : C_k \hookrightarrow \Gamma$ is cycle of length $k$ starting at 1, the edges $\gamma_i(\{i-1, i\})$ are of the form $\{g_i, g_is_i\}$ 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 $C_k \rightarrow \Gamma$ is an embedding, we also obtain $s_i \neq s_{i+1}$ for all $i$, hence $m \leq k$.

(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 $\mathbb{Z}/m\mathbb{Z}$ of $C_m$ with $\{0, \ldots, m-1\}$, we define

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

and

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

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

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

and hence

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

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

Remark 2.3.10. The restriction on $S$ is needed because, for instance, the Cayley graph $\mathcal{C}(\mathbb{Z}/2\mathbb{Z}, \{1\})$ 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 of the second kind, defined 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 \( X \) on which the group acts: 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 \( x, s \cdot x \) for \( s \in S \).

For reasons that we will clarify a bit later, it is important to be careful with the definition of edges in such an action graph, because of the possibility of fixed points, which may lead to multiple edges and to loops.

We begin with the following graph:

\[
\Gamma_1 = (X, S \times X, \text{ep})
\]

where

\[
\text{ep}((s, x)) = \{x, s \cdot x\}.
\]

This set of edges is too big for us, because the edges \((s, x)\) and \((s^{-1}, s \cdot x)\) (which are usually distinct) have always the same sets of endpoints \(\{x, s \cdot x\} = \{s \cdot x, s^{-1} \cdot (s \cdot x)\}\). When these are two endpoints, we wish to identify them. This leads to the definition:

**Definition 2.3.11 (Action graph).** Let \( G \) be a group, \( S \subset G \) a symmetric subset and \( X \) a set on which \( G \) acts on the left. The *action graph* of \( X \) with respect to \( S \), denoted \( A(X, S) \), is the graph

\[
A(X, S) = (X, (S \times X)/\sim, \text{ep})
\]

where the equivalence relation \( \sim \) on \( S \times X \) is defined by the only equivalences \((s, x) \sim (s, x)\) and

\[
(s, x) \sim (s^{-1}, s \cdot x)
\]

when \( x \in X \), \( s \in S \) and \( s \cdot x \neq x \), and with endpoints induced by \( \text{ep}((s, x)) = \{x, s \cdot x\} \), which is well-defined on the quotient set.

In other words, we have defined \( A(X, S) \) as obtained, from the graph \( \Gamma_1 \) defined in (2.9) by identifying only those edges of \( \Gamma_1 \) of the type previously mentioned. This identification is, in general, intermediate between \( \Gamma_1 \) and the simple graph \( \Gamma^s \): the action graph may still have multiple edges, and may have loops. We will give an example just after the next definition, which corresponds to the important special case of the action of a group on its cosets modulo a subgroup:

**Definition 2.3.12 (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 \( G/H \) with respect to \( S \) is the action graph \( \mathcal{A}(G/H, S) \), where \( G \) acts on \( G/H \) by

\[
g \cdot (xH) = (gx)H
\]

for \( g \in G \) and \( xH \in G/H \).

2. If \( H \triangleleft G \) is normal in \( G \) with quotient \( K = G/H \), we call the Schreier graph the *relative Cayley graph* of \( K \) with respect to \( S \). It is also denoted \( C(K, S) \).

**Example 2.3.13.** The following example should clarify the construction. We take \( G = S_5 \), the symmetric group on 5 letters, with symmetric generating set \( S = \{\tau, \mu, \mu^{-1}\} \) where

\[
\tau = (1 2), \quad \mu = (1 2 3 4 5), \quad \mu^{-1} = (1 5 4 3 2),
\]

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

This group acts on \( X = \{1, \ldots, 5\} \) by evaluation (i.e., \( \sigma \cdot i = \sigma(i) \)). This action is (in the natural sense) isomorphic to the multiplication action of \( G \) on right cosets of the
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 the graphical representation of the graph \( \Gamma_1 \) defined by (2.9) in this case:

![Graph Γ₁](image)

The reader should check that it is correct. For instance, there are indeed \( 15 = |S| \times |X| \) edges, and the four edges with extremities \( \{1, 2\} \) are given by

\[
(\tau, 1), \quad (\tau, 2), \quad (\mu, 1), \quad (\mu^{-1}, 2),
\]

while the loops around 3, 4 and 5 are coded by \( (\tau, i), \; 3 \leq i \leq 5 \).

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

![Graph Γ](image)

There are now only two edges between 1 and 2, since

\[
(\tau, 1) \sim (\tau, 2) \quad (\mu, 1) \sim (\mu^{-1}, 2),
\]

but the loops are preserved.

One sees also that the associated simple graph is indeed different from both \( \Gamma_1 \) and \( \mathcal{A}(X, S) \): it is simply the cycle \( C_5 \).

The reader is invited to repeat this exercise when \( S \) is replaced with \( S_1 = S \cup \{(1 2 3), (1 3 2)\} \), and check that the action graph \( \mathcal{A}(X, S_1) \) is 5-regular, with three distinct loops at the vertices 4 and 5. This shows why the equivalence relation only identifies \( (s, x) \) with \( (s^{-1}, s \cdot x) \) when \( s \cdot x \neq x \).

The following simple lemma explains why the definition of the action graph is chosen the way it is (and why it is indeed better than the simpler-looking \( \Gamma_1 \)):

**Lemma 2.3.14** (Regularity of action graphs). Let \( G \) be a group acting on a set \( X \), \( S \subset G \) a symmetric generating set of \( G \). For any \( x \in X \), there is a natural bijection between \( S \) and the set of edges \( \alpha \) in \( \mathcal{A}(G, S) \) with one extremity equal to \( x \), i.e., with \( x \in \text{ep}(\alpha) \). In particular, \( \mathcal{A}(X, S) \) is \( |S| \)-regular.

Note that in Example 2.3.13, we see that the graph \( \Gamma_1 \) is not always regular (the valency of the vertices 1 and 2 is 6, while that of 3, 4, 5 is equal to 5.) And when one
Proof. Let $E_x$ be the set of edges where $x$ is an extremity. We have a map

$$i \begin{cases} S & \mapsto E_x \\ s & \mapsto [(s, x)] \end{cases}$$

where $[(s, x)]$ denotes the equivalence class with respect to the equivalence relation $\sim$ on the edges which is used to define the action graph. We claim this is a bijection. Indeed, if $i(s) = i(t)$, we have $(s, x) \sim (t, x)$, which implies $s = t$ (the alternative is that $s \cdot x \neq x$ and $(t, x) = (s^{-1}, s \cdot x)$, which is contradictory!) Furthermore, if an edge $\alpha = [(t, y)]$ is in $E_x$, we must have $x \in \{y, t \cdot y\}$, and either $x = y$, so that $\alpha = i(t)$, or $x \neq y$ and $x = t \cdot y$, in which case $(t, y) \sim (t^{-1}, t \cdot y) = (t^{-1}, x)$, i.e., $\alpha = i(t^{-1})$.

So we have checked that $i$ is injective and surjective, concluding the proof. □

Exercise 2.3.15. (1) Continuing Exercise 2.1.5, give a definition of action graphs using the coding you defined, which corresponds to the one above.

(2) Show that for any $x, y \in X$, there is a bijection between edges in $A(X, S)$ with extremities at $x$ and $y$ and elements $s \in S$ such that $s \cdot x = y$. (In particular, this determines the adjacency matrix of $A(X, S)$.)

Naturally enough, we obtain graph maps between Cayley graphs and action graphs. However, our convention corresponds to Cayley graphs being associated to the right action on $G$ on itself by multiplication on the right, and we therefore define an action graph for a right action by an immediate adaptation of the definition: we use $(S \times X)/\sim$ as edge set with

$$(x, s) \sim (x, s), \quad (x, s) \sim (x \cdot s, s^{-1})$$

when $x \neq x \cdot s$ as equivalences, and $\text{ep}((x, s)) = \{x, x \cdot s\}$.

Proposition 2.3.16. Let $G$ be a group and $S \subseteq G$ 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

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

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

$A(X, S) \to A(Y, S)$.

Proof. (1) We need to define the map $f_s$ between edges so that $(f, f_s)$ is a graph map. Thus we must map an edge $\{g, gs\}$ between $g$ and $gs$ in the Cayley graph to an edge between $x_0 \cdot g$ and $x_0 \cdot (gs)$. The natural definition is

$$f_s(\{g, gs\}) = [(x_0 \cdot g), s],$$

but we must check that this is well-defined. Indeed, this is the case, provided $x_0 \cdot g \neq x_0 \cdot (gs)$ (since we then have $[(x_0 \cdot g), s] = [(x_0 \cdot gs, s^{-1})$, by definition of the equivalence relation.)
If \( x_0 \cdot g = x_0 \cdot (gs) \), on the other hand, we must map the edge \( \{g, gs\} \) to a loop based at \( f(g) = x_0 \cdot g = f(gs) \). We can do this, since the edges \([f(g), s]\) and \([f(gs), s^{-1}]\) are loops at \( x_0 \cdot g \) in the action graph. They are usually not the same (they are equal if and only if \( s = s^{-1} \)), so we pick one randomly to define \( f_\ast(\{g, gs\}) \).

(2) The construction is similar (and one may need to make a similar choice of loops...)

\[\square\]

**Remark 2.3.17.** With our definitions, the map \( f_\ast \) can not be specified uniquely, but this does should not create difficulties in applications.

**Exercise 2.3.18.** Let \( G \) act on the left on a set \( X \).

1. If \( S \subset G \) is a symmetric generating set, show that \( A(X, S) \) 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 \( A(G/H, S) \) has no loops if and only if \( S \cap xHx^{-1} = \emptyset \) for all \( x \in G \).

3. Let \( k \geq 2 \) 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 \( A(X, S) \) 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 in 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 Section 1.1, 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 $\Gamma_m$ given by taking the disjoint union of two copies $\Gamma$ and $\Gamma'$ of a complete graph $K_m$, for some $m \geq 2$, and adding a single edge between chosen vertices $x_1 \in \Gamma$ and $x_2 \in \Gamma'$:

We clearly have $\text{diam}(\Gamma_m) = 3$, for any $m$, which shows that $\Gamma_m$ 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.

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

(1) For any disjoint subsets of vertices $V_1, V_2 \subset V$, we denote by $E(V_1,V_2)$ or $E_\Gamma(V_1,V_2)$ the set of edges of $\Gamma$ with one extremity in $V_1$ and one extremity in $V_2$, $E(V_1,V_2) = \{ \alpha \in E \mid ep(\alpha) \cap V_1 \neq \emptyset, \ ep(\alpha) \cap V_2 \neq \emptyset\}$.

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

(2) The expansion constant $h(\Gamma)$ is defined by $h(\Gamma) = \min \left\{ \frac{|E(W)|}{|W|} \mid \emptyset \neq W \subset V, \text{ such that } |W| \leq \frac{1}{2} |\Gamma| \right\}$.

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 non-empty, but not too big, subset of vertices. 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 $\Gamma = (V,E,ep)$ be a finite graph.
(1) We have \( h(\Gamma) > 0 \) if and only if \( \Gamma \) is connected.

(2) If \( W \subset V \) is a subset of vertices with \( |W| = \delta|V|, \delta \leq \frac{1}{2} \), one must remove at least \( \delta h(\Gamma)|V| \) edges from \( \Gamma \) 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 \( \leq |\Gamma|/2 \), such that \( E(W) \) is empty. If \( x \in W \) and \( y \notin W \) are any two vertices, there can not be a path in \( \Gamma \) between \( x \) and \( y \), since it would have to cross from \( W \) to \( V - W \) at some point. Hence \( \Gamma \) is not connected.

Conversely, if \( \Gamma \) is not connected, there must be at least two connected components, and at least one of them must have size \( |W| \leq |\Gamma|/2 \). Then \( E(W) = \emptyset \), and hence \( h(\Gamma) = 0 \).

(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 \( E(W) \subset C \), i.e., all edges that go from \( W \) to “somewhere else” are contained in \( C \). Then since \( |E(W)| \geq h(\Gamma)|W| = \delta h(\Gamma)|V| \), by definition of \( h(\Gamma) \), our statement is just a reformulation...

**Example 3.1.3.** (1) Consider the complete graph \( K_m \) with \( m \geq 2 \) 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

\[
\begin{align*}
  h(K_m) &= \min_{1 \leq j \leq m/2} \frac{1}{j} |E(\{1, \ldots, j\})| = \min_{1 \leq j \leq m/2} (m - j) = m - \left\lfloor \frac{m}{2} \right\rfloor \\
\end{align*}
\]

(since there are \( j(m - j) \) edges in \( K_m \) from \( \{1, \ldots, j\} \) to its complement \( \{j + 1, \ldots, m\} \)). Note here that it is very visible that expansion “slows down” when sets of vertices larger than half of the graph are considered, so the condition \( |V| \leq |\Gamma|/2 \) is needed to obtain a good definition.

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

\[
(3.1) \quad h(C_m) = \frac{2}{\left\lfloor \frac{m}{2} \right\rfloor} \leq \frac{4}{m - 1}.
\]

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

\[
\begin{align*}
  h(C_m) &\leq \frac{|E(W)|}{|W|} \\
\end{align*}
\]

by definition for any subset \( W \).

(3) Let \( \Gamma \) be a graph like the one discussed at the beginning of this section: two copies of \( K_m \) joined by a single edge \( \alpha \). Then if we take \( W \) to be the first copy of \( K_m \), we see that \( E(W) = \{\alpha\} \), hence

\[
  h(\Gamma) \leq \frac{1}{m}.
\]

(4) Let \( T = T_{d,k} \) be a finite tree with degree \( d \geq 3 \) and depth \( k \geq 1 \). 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

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

which (see Exercise 2.2.5, (4)), can be written equivalently as

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

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

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

It is clear from a picture that $E(W)$ 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

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

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 Example (3) have quite small expansion constants.

The arguments, though they were highly elementary, also indicate that it can be 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 $\Gamma$ such that $V$ and $E$ are non-empty, we have

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

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

**Proof.** For the lower-bound, we just note that since $\Gamma$ is connected, we must have $|E(W)| \geq 1$ for any non-empty proper subset $W$ in $V$, hence

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

if $1 \leq |W| \leq |\Gamma|/2$.

On the other hand, for the upper-bound, take $W = \{x\}$ for some $x$; we have $|E(W)| = \text{val}(x)$ and hence

$$h(\Gamma) \leq \text{val}(x),$$

which gives the result. \qed

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 \( \Gamma \) be a finite non-empty connected graph. We have

\[
\text{diam}(\Gamma) \leq 2 \frac{\log |\Gamma|}{\log \left( 1 + \frac{h(\Gamma)}{v} \right)} + 3
\]

where

\[
v = \max_{x \in V} \text{val}(x).
\]

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 ratio gives a geometric 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 \( \Gamma \) be a finite non-empty connected graph and \( x \in V \). For any \( n \geq 0 \), let

\[ B_x(n) = \{ y \in V \mid d_\Gamma(x, y) \leq n \}. \]

Then, with \( v \) as above, we have

\[
|B_x(n)| \geq \min \left( \frac{|\Gamma|}{2}, \left( 1 + \frac{h(\Gamma)}{v} \right)^n \right).
\]

Proof. It is enough to show that if \( n \geq 0 \) is such that \( |B_x(n)| \leq |\Gamma|/2 \), we have

\[
|B_x(n+1)| \geq \left( 1 + \frac{h(\Gamma)}{v} \right) |B_x(n)|,
\]

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

It is possible that multiple edges \( \alpha \) starting from \( B_x(n) \) lead to the same \( y \), but since all these edges share the extremity \( y \), the maximal number of edges leading to \( y \) is \( \text{val}(y) \leq v \), so that

\[
|B_x(n+1) - B_x(n)| \geq \frac{|E(B_x(n))|}{v} \geq \frac{h(\Gamma)}{v} |B_x(n)|,
\]

by definition of \( h(\Gamma) \), using the assumption that \( |B_x(n)| \leq |\Gamma|/2 \). Then we get

\[
|B_x(n+1)| = |B_x(n)| + |B_x(n+1) - B_x(n)| \geq \left( 1 + \frac{h(\Gamma)}{v} \right) |B_x(n)|,
\]

as desired. \( \square \)

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 \( n \geq 1 \) the smallest integer such that

\[
\beta^n \geq \frac{|\Gamma|}{2}.
\]
(which is possible since $\beta > 1$, in view of the connectedness of $\Gamma$). Then by Lemma 3.1.6, applied to $x$ and $y$, we find that

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

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

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

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

we obtain the diameter bound that we stated. \qed

**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:

$$\text{diam}(K_m) \ll \log m, \quad \text{diam}(C_m) \ll m \log m$$

for $m \geq 2$. 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 $\Gamma_i = (V_i, E_i, e)$ is an expander family, or a family of expanders, if there exist constants $v \geq 1$ and $h > 0$, independent of $i$, such that:

- The number of vertices $|V_i|$ “tends to infinity”, in the sense that for any $N \geq 1$, there are only finitely many $i \in I$ such that $\Gamma_i$ has at most $N$ vertices;
- For each $i \in I$, we have

$$\max_{x \in V_i} \text{val}(x) \leq v,$$

i.e., the maximal valency of the graphs is bounded independently of $i$;
- For each $i \in I$, the expansion ratio satisfies

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

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.** Of course, 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 $\Gamma$ 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 not allowed to grow. The idea 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”, and hence repeating a single graph infinitely many times is not interesting.

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

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

The number of edges is seen here (as discussed in Section 1.1) 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 $(\Gamma_i)$ be an expander family of graphs. Then we have

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

for all $i$, where the implied constant depends only on the expansion parameters $(h,v)$ of the family. \footnote{We use $3|\Gamma_i|$ here in order to avoid problem with the exceptional $i$'s where $|\Gamma_i| = 1$, and because $\log 3 \geq 1$; this is old analytic number theory lore.}

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

**Proof.** We remove from $I$ those indices $i \in J$ such that $|\Gamma_i| \leq \frac{1}{3}e^3$, and then we just apply Proposition 3.1.5: denoting

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

and

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

we get

$$\text{diam}(\Gamma_i) \leq 2\xi \log\left(\frac{1}{2}|\Gamma_i|\right) + 3 \leq 2\xi \log\left(\frac{1}{2}|\Gamma_i|\right) + \log(3|\Gamma_i|) \leq (2\xi + 1) \log(3|\Gamma_i|),$$

for $i \notin J$, and we get an estimate valid for all $i$, e.g., by writing

$$\text{diam}(\Gamma_i) \leq C \log(3|\Gamma_i|)$$

with

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

for all $i \in I$. \hfill \Box

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 $\Gamma$ be any graph with maximal valency $\leq k$, where $k \geq 0$ is an integer. Then

$$\text{diam}(\Gamma) \geq \frac{\log(|\Gamma|)}{\log k}.$$ 

**Proof.** The idea is simply that, from distance $n$ to $n + 1$ of a fixed vertex $x$, we can at most multiply the number of vertices by the maximal valency, so the size of balls around $x$ can only grow exponentially fast. Precisely, let $x \in \Gamma$ be such a fixed vertex. We claim that

$$|B_x(n)| \leq k^n$$

for $n \geq 0$ (which is wildly pessimistic but suffices!). Indeed, from each $y \in B_x(n - 1)$, we reach at most $k$ new vertices at distance $n$, so that

$$|B_x(n)| \leq k|B_x(n - 1)|,$$

which gives the bound above by induction. Now, if $d = \text{diam}(\Gamma)$, we have $B_x(d) = \Gamma$ by definition and hence

$$|\Gamma| \leq k^d,$$

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

$$|B_x(n)| \geq \min\left(\frac{|\Gamma|}{2}, \gamma^n\right),$$

for all $x \in \Gamma_i$ and $n \geq 0$. 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 $(\Gamma_i)$ be an expander family such that all $\Gamma_i$ are $k$-regular for some fixed integer $k \geq 1$. Show that $k \geq 3$. (We will see later that $k$-regular expander families do exist for all $k \geq 3$.)

**Exercise 3.1.14 (Some Cayley graphs of $S_n$).** 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

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

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 3.4.8, another proof of this using the results of Section 3.3, 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 $\mathbb{Z}/n\mathbb{Z}$. With that interpretation, the generators $\sigma_n$ and $\sigma_n^{-1}$ act on $\mathbb{Z}/n\mathbb{Z}$ by

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

for $i \in \mathbb{Z}/n\mathbb{Z}$. 
Define then

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

(1) Show that

$$\frac{|\mathcal{E}(W_n)|}{|S_n|} \ll \frac{1}{n}.$$ 

(2) Show that

$$\frac{1}{3} \leq \liminf_{n \to +\infty} \frac{|W_n|}{|S_n|} \leq \limsup_{n \to +\infty} \frac{|W_n|}{|S_n|} \leq \frac{1}{2},$$
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

$$(3.4) \quad \tilde{h}(\Gamma) = \min_{1 \leq |W| \leq |\Gamma|/2} \frac{|\partial W|}{|W|},$$

where

$$\partial W = \{ x \in V_{\Gamma} \mid x \notin W, \ d_{\Gamma}(x,y) = 1 \text{ for some } y \in W \}$$
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 $\text{ep}(\alpha) = \{ x, y \}$, since cutting one of the 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 (3.4). This follows from an easy lemma:

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

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

In particular, a family $(\Gamma_i)$ is an expander family if and only if it has bounded valency, and there exists $\tilde{h} > 0$ such that

$$\tilde{h}(\Gamma_i) \geq \tilde{h}$$
for all $i \in I$.

**Proof.** Consider the map

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

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 $|\mathcal{E}(W)| \leq v |\partial W|$. \qed
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 \( \Gamma = (V, E) \) a finite bipartite graph with a decomposition \( V = V_0 \cup V_1 \), we let

\[
\bar{h}(\Gamma) = \min(h_0, h_1), \quad h_i = \min_{W \subseteq V_i, 1 \leq |W| \leq |V_i|/2} \frac{|\partial W|}{|W|}.
\]

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

**Lemma 3.1.16.** For \( \Gamma = (V, E) \) a finite bipartite graph with a decomposition \( V = V_0 \cup V_1 \) and with maximal valency \( v \geq 1 \). Assume that \( |V_0| = |V_1| \). We have

\[
\frac{\bar{h}(\Gamma) - 1}{2} \leq h(\Gamma) \leq v\bar{h}(\Gamma).
\]

**Proof.** The upper-bound is easy and left as an exercise. For the lower-bound, we can assume that \( \bar{h}(\Gamma) \geq 1 \), say \( \bar{h}(\Gamma) = 1 + \delta \) with \( \delta \geq 0 \); we must check that \( h(\Gamma) \geq \delta/2 \). Let then \( W \subseteq V \) be any subset with \( 1 \leq |W| \leq |V|/2 \). 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 \( \bar{h}(\Gamma) \), we can assume that \( |W_1| \leq |W_0| \), and in particular that \( |W_0| \geq |W|/2 \). We now distinguish two cases:

1. If \( |W_0| \leq |V_0|/2 \), we deduce by definition that

\[
|\partial W_0| \geq (1 + \delta)|W_0| \geq (1 + \delta)|W_1|,
\]

from which it follows that \( W_0 \) has at least \( \delta|W_1| \geq \frac{\delta}{2}|W| \) neighbors (in \( V_1 \)) which are not in \( W \). Hence

\[
\frac{|E(W)|}{|W|} \geq \frac{\delta}{2}
\]

in that case.

2. If \( |W_0| > |V_0|/2 \), we deduce

\[
|\partial W_0| \geq (1 + \delta)\frac{|V_0|}{2}
\]

(by applying the definition to a subset of \( W_0 \) of size \( \lceil |V_0|/2 \rceil \)). But \( |W_1| \leq |V_0|/2 \) since \( |W| = |W_0| + |W_1| \leq |V| = |V_0|/2 \), and therefore we again see that \( \partial W_0 \) contains at least \( \delta|V_0|/2 = \delta|V|/4 \geq \delta|W|/2 \) neighbors not in \( W \), and hence the same lower bound on \( |E(W)|/|W| \) follows. \( \square \)

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 ratios).** Let \( \Gamma_1 = (V_1, E_1, \text{ep}) \) and \( \Gamma_2 = (V_2, E_2, \text{ep}) \) 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 \( d \geq 1 \), in particular \( |V_1| = d|V_2| \);
2. There exists \( C > 0 \) for which

\[
d_2(f(x), f(y)) \leq Cd_1(x, y),
\]

for all \( x, y \in V_1 \).
Then we have
\[ h(\Gamma_2) \geq \frac{h(\Gamma_1)}{w}, \]
where \( w > 0 \) depends only on \((C, v_1, v_2)\), namely
\[ w = v_1 \sum_{j=1}^{\lfloor C \rceil} v_j^{j-1}. \]

We emphasize that \( f \) is not assumed to be a graph map, so that the condition \( d_2(f(x), f(y)) \leq d_1(x, y) \) is not automatic (as it is for graph maps, as stated in Proposition 2.2.6.)

**Proof.** Let \( W \subset V_2 \) be a non-empty set of vertices with \(|W| \leq \lfloor V_2 \rfloor/2\). Assumption (1) implies that \( W' = f^{-1}(W') \subset V_1 \) is (non-empty and) of size \( \leq |V_1|/2 \). Thus we get
\[ |\mathcal{E}(W')| \geq h(\Gamma_1)|W'| = d h(\Gamma_1)|W|. \]

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
\[ dh(\Gamma_1)|W| \leq |\mathcal{E}(W')| \leq v_1 |\partial W'| \]
(Using Lemma 3.1.15).

Using (1) once more, we have \(|\partial W'| = d|f(\partial W')|\), hence
\[ h(\Gamma_1)|W| \leq v_1 |f(\partial W')|. \]

Since any \( x \in \partial W' \) satisfies \( d_1(x, x_0) = 1 \) for some \( x_0 \in W' \), assumption (2) gives
\[ f(\partial W') \subset W'' = \{ y \in V_2 - W \mid d_2(y, W) \leq C \}. \]

By induction on \( j \geq 1 \), we have
\[ |\{ y \in V_2 \mid y \notin W, \ d_2(y, W) = j \}| \leq v_j^{j-1}|\partial W| \leq v_j^{j-1}|\mathcal{E}(W)|, \]
(Using again Lemma 3.1.15) and hence
\[ |f(\partial W')| \leq |W''| \leq \left( \sum_{j=1}^{\lfloor C \rceil} v_j^{j-1} \right) |\mathcal{E}(W)|, \]
which leads to the inequality
\[ |\mathcal{E}(W)| \geq \frac{h(\Gamma_1)}{v}|W|, \quad v = v_1 \sum_{j=1}^{\lfloor C \rceil} v_j^{j-1}, \]
and hence to the conclusion. \( \square \)

Here are applications to expanders:

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

1. If \( \Gamma'_i \) is any graph with the same vertices as \( \Gamma_i \) and with “more edges”, i.e., \( E'_i \supset E_i \), then \((\Gamma'_i)\) is also a family of expander graphs provided the maximal valency of \( \Gamma'_i \) remains bounded.

2. The family of simple graphs \((\Gamma'_s)\) is a family of expanders.

3. More generally assume that, for any \( i \in I \), we are given graphs \( \Gamma_i = (V_i, E_i, ep) \) with maximal valency bounded by \( w \), and bijections \( V_i \xrightarrow{f_i} V'_i \), such that
\[ d_{\Gamma'_i}(f_i(x), f_i(y)) \leq Cd_{\Gamma_i}(x, y) \]
for some fixed constant $C \geq 0$.

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

$$\inf_{i \in I} h(\Gamma'_i) \geq \delta \inf_{i \in I} h(\Gamma_i),$$

where

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

i.e., if $(h, v)$ are expansion parameters for $(\Gamma_i)$, then $(\delta h, w)$ are expansion parameters for $(\Gamma'_i)$.

The example to keep in mind for (3) is when $V'_i = V_i$ and $f_i$ is simply the identity. This means that $\Gamma'_i$ is a graph with the same vertices, but with edges “rearranged” in some way, and the condition (3.6) 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 $\Gamma_i$ with a graph $\Gamma'_i$, which has in each case the same set of vertices (i.e., the graphs $(\Gamma_1, \Gamma_2)$ in the Lemma are $(\Gamma_i, \Gamma'_i)$), and with $f$ being the identity (so that Condition (1) of the Lemma is automatic). In (1), because we added edges, we have

$$d_{\Gamma'_i}(x, y) \leq d_{\Gamma_i}(x, y),$$

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 2.2.8), so Condition (2) holds again with $C = 1$. Finally, in (3), Condition (2) is precisely given by (3.6).

**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 trees $T_{d,k}$ with $d \geq 3$ fixed and $k \to +\infty$), there are many subsets of vertices which do have large expansion ratio $|E(W)|/|W|$. 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 some edges to 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 Section 3.5 using probabilistic methods, as originally done by Barzdin and Kolmogorov [1], and independently by Pinsker [39]. The reader may skip to that section right now, since it is independent of what follows directly.

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, 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 \( \Gamma = (V, E, ep) \) be a graph. A random walk on \( \Gamma \) is a sequence \( (X_n)_{n \geq 0} \) of \( V \)-valued random variables, defined on a common probability space \( (\Omega, \Sigma, P) \), with joint distribution satisfying the following rule:

\[
P\{X_{n+1} = y \mid (X_n, \ldots, X_0) = (x_n, \ldots, x_0)\} = P\{X_{n+1} = y \mid X_n = x_n\}
\]

\[
= \begin{cases} 
0 & \text{if } d_\Gamma(x,y) > 1, \\
\frac{|\{\alpha \in E \mid \text{ep(\alpha)} = \{x,y\}\}|}{\text{val}(x)} & \text{if } d_\Gamma(x,y) = 0 \text{ or } 1,
\end{cases}
\]

or 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 happens 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 \( P(X_0 = x) \) 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 starts at \( x_0 \), 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 $(\Omega, \Sigma, P)$, the conditional probability of $A$ knowing $B$, denoted $P(A \mid B)$, is defined by

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

(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

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

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

$$P(x,y) = \frac{a(x,y)}{\text{val}(x)}$$

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 [30]. 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 $\mu_0$ on $V$, determined by the probabilities $\mu_0(x)$ for $x \in V$, there is a random walk on $\Gamma$ (on some probability space) for which $X_0$ has distribution $\mu_0$, i.e.

$$P(X_0 = x) = \mu_0(x).$$

This existence statement is a standard fact in probability theory, which we will not prove (except for reducing it to another standard property, in the case when $\Gamma$ is a Cayley graph, see Example 3.2.6 below). We observe, however, that this random walk is unique (given $\mu_0$) in the sense that the joint distribution of the process $(X_n)$, i.e., all values of probabilities of the type

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

for $k \geq 1$, $n_1 < n_2 < \cdots < n_k$, and $(x_1, \ldots, x_k) \in V^k$, depend only on $\mu_0$, and not on specific features of the construction. To give an example, let

$$q(w,x,y,z) = P((X_0, X_1, X_2, X_3) = (w, x, y, z))$$

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

$$q(w, x, y, z) = P(X_3 = z \mid (X_0, X_1, X_2) = (w, x, y)) \cdot P((X_0, X_1, X_2) = (w, x, y))$$

$$= P(y, z) \cdot P(X_2 = y \mid (X_0, X_1) = (w, x)) \cdot P((X_0, X_1) = (w, x))$$

$$= P(y, z) \cdot P(X_1 = x \mid X_0 = w) \cdot P(X_0 = w)$$

$$= P(y, z) \cdot P(x, y) \cdot P(w, x) \cdot \mu_0(\{w\}),$$

which is determined by $P$ and $\mu_0$, 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 $\Gamma$ also leads to a similar random walk on the graph.
Proposition 3.2.3. Let $\Gamma = (V, E, ep)$ be a graph and let $(X_n)$ be a random walk on $\Gamma$. Fix an integer $m \geq 0$, and define

$$Y_n = X_{m+n}$$

for $n \geq 0$. Then $(Y_n)$ is a random walk on $\Gamma$ with initial distribution $\mu_1$ given by the law of $X_m$, i.e., by

$$\mu_1(A) = \Pr(X_m \in A)$$

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 $\Gamma = (V, E, ep)$ be a finite graph, and $(X_n)$ a random walk on $\Gamma$.

(1) Let $\tau$ be an random variable taking non-negative integer values such that, for any $n \geq 0$, the event $\{\tau = n\}$ can be described only using $X_0, \ldots, X_n$. For $n \geq 0$, 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 $\Gamma$ with initial distribution given by the law of $X_\tau$; note that we can take $\tau$ to be constant, equal to some integer $m \geq 0$, and the result is then Proposition 3.2.3.)

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

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

has the desired property (here we allow infinity values).

Random variables $\tau$ with the property above that $\{\tau \leq n\}$ 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 [30, §6.2] for an introduction and examples.

Remark 3.2.5. 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

\begin{equation}
\Pr(X_{n+1} = y \mid X_n = x) = P(x, y)
\end{equation}

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

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

by (3.9) and then inductively

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

for $n \geq 0$, 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 (3.9) in our treatment. However, in Section 4.2, we will consider some applications of random walks where the full property (3.7) is used (see Proposition 4.2.6).
3.2.6 (Random walk on a Cayley graph). If $\Gamma = \mathcal{C}(G, S)$ is the Cayley graph of a group $G$ with respect to a symmetric set $S \subset G$, and $(X_n)$ is a random walk on $\Gamma$ starting at $x_0 = 1 \in S$, the distribution of the $n$-th step is given by

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

This formula leads to a “construction” of the random walk on $\Gamma$: let $(\xi_n)_{n \geq 1}$ be a sequence of independent $S$-valued random variables, each identically uniformly distributed, so that

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

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

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

for any $k \geq 0$, any choices of distinct indices $n_1, \ldots, n_k$ and any $(s_1, \ldots, s_k) \in S^k$. We then see that

$$X_0 = 1, \quad X_n = \xi_1 \cdots \xi_n$$

defines a sequence $(X_n)$ of $G$-valued random variables which form a random walk on $\Gamma$ with $X_0 = 1$. To obtain a random walk with another initial distribution $\mu_0$, pick any $G$-valued random variable $X_0$ with distribution given by $\mu_0$, and let

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

Exercise 3.2.7. Check that $(X_n)$ is indeed a random walk on $\mathcal{C}(G, S)$, i.e., that (3.7) holds.

We get an analogue of Proposition 3.2.3:

Proposition 3.2.8. Let $\Gamma = \mathcal{C}(G, S)$ be a Cayley graph and $(X_n)$ a random walk on $\Gamma$ starting at $X_0 = 1$ with independent increments $(\xi_n)$, as above. For any $m$ and $n \geq 0$, $Y_n = X_m^1 X_{m+n}$ 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 $\xi_1, \ldots, \xi_m$, 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 $\xi_i$ is symmetrically-distributed, so that $X_n^{-1}$ has the same distribution as $X_n$. \qed

Take for instance $G = \mathbb{Z}^r$ for some $r \geq 1$, and $S$ the vectors in the canonical basis of $G$ together with their inverses. The resulting process is called the simple random walk on $\mathbb{Z}^r$. It is historically the first example of a random walk to have been studied. Indeed, Pólya proved the first theorem in the subject: if $r = 1$ or 2, the simple random walk on $\mathbb{Z}^r$ is recurrent, in the sense that

$$P(X_n = 0 \text{ for some } n \geq 1) = 1$$
(i.e., almost surely, every random walk on $\mathbb{Z}$ or $\mathbb{Z}^2$ will return to the origin; it is then fairly simple to see that, almost surely, the walker will come back infinitely often) while it is not for $r \geq 3$ (see, e.g., [30, Ch. 21]).

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, the walker is as likely to be at any vertex at time $n$, independently of the starting point or of the steps taken to get there.

The basic philosophy of this section 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 is easy to present from scratch.

**Definition 3.2.9 (Measure and functions on a graph).** Let $\Gamma = (V, E, ep)$ be a finite graph with $E$ non-empty and with $\text{val}(x) \geq 1$ for all $x \in V$.

(1) The (normalized) graph measure on $\Gamma$ is the probability measure on $V$ defined by

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

for all $x \in V$, where

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

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

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

equipped with the Hilbert space structure induced by $\mu_\Gamma$, i.e., with inner product

$$\langle \varphi_1, \varphi_2 \rangle_\Gamma = \frac{1}{N} \sum_{x \in V} \text{val}(x) \varphi_1(x) \varphi_2(x)$$

and norm-squared

$$\|\varphi\|^2_\Gamma = \frac{1}{N} \sum_{x \in V} \text{val}(x) |\varphi(x)|^2.$$

**Exercise 3.2.10 (An alternative formula).** Let $\Gamma = (V, E, ep)$ be a finite graph, and $\varphi \in L^2(\Gamma)$ of mean zero, i.e., $\langle \varphi, 1 \rangle = 0$. Show that

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

**Remark 3.2.11.** (1) As usual, we will often drop the subscript $\Gamma$ when the context is clear.

(2) Note that if $\Gamma$ is $k$-regular for some $k \geq 1$, the measure $\mu_\Gamma$ becomes the uniform counting measure defined by

$$\mu_\Gamma(W) = \frac{|W|}{|V|}.$$
for all $W \subset V$. This will be the case for most applications so the reader may read this section first with this case in mind. In general, we have the comparison relation

\[(3.11) \quad \frac{v_- |W|}{v_+ |V|} \leq \mu_\Gamma(W) \leq \frac{v_+ |W|}{v_- |V|}\]

for all $W \subset V$, where $v_- = \min_{x \in V} \val(x)$, $v_+ = \max_{x \in V} \val(x)$.

(3) If $\Gamma$ is an infinite graph, we cannot define a probability measure in the same way (since $N$ typically diverges), but we can still speak of the measure with weight $\val(x)$, and of the associated $L^2$-space.

(4) Finally, we will also have the occasion to use the supremum norm $\| \cdot \|_\infty$ on $L^2(\Gamma)$, which is always well-defined for finite graphs:

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

Any two norms on a finite-dimensional vector space are equivalent, and in this specific case we have the inequalities

\[(3.12) \quad \| \varphi \| \leq \| \varphi \|_\infty \leq \left( \frac{N}{v_-} \right)^{1/2} \| \varphi \|,\]

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

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

(5) Here is a very technical remark: we assumed that there is no isolated vertex (i.e., $\val(x) \geq 1$) because otherwise the inner-product is not positive-definite on the space of all functions $f : V \rightarrow \mathbb{C}$. In that case, the “proper” definition of $L^2(\Gamma)$ is the quotient of the space of functions on $V$ by the subspace of functions which are zero “almost everywhere” with respect to $\mu(\Gamma)$. Such a function is one which is zero on those vertices with $\val(x) \geq 1$, and hence the isolated vertices are “invisible” from the point of view of $L^2(\Gamma)$.

**Exercise 3.2.12.** Let $\Gamma = (V, E, ep)$ be a finite graph, and let $(X_n)$ be the random walk on $\Gamma$ with initial distribution $\mu_\Gamma$. Show that, for all $n$, the random variable $X_n$ is also distributed like $\mu_\Gamma$. (Of course, $X_n$ is not the same random variable as $X_0$, it simply has the same distribution.)

The basic lemma is the following simple identity:

**Lemma 3.2.13 (Markov averaging operator).** Let $\Gamma = (V, E, ep)$ be a finite graph, and let $(X_n)$ be a random walk on $\Gamma$. For any function $\varphi \in L^2(\Gamma)$ and $n \geq 0$, we have

$$\mathbf{E}(\varphi(X_{n+1})) = \mathbf{E}((M\varphi)(X_n))$$

where

$$M : \begin{cases} L^2(\Gamma) & \rightarrow & L^2(\Gamma) \\ \varphi & \mapsto & M\varphi \end{cases}$$

is the Markov averaging operator given by

$$(M\varphi)(x) = \frac{1}{\val(x)} \sum_{\alpha \in E, ep(\alpha) = \{x, y\}} \varphi(y) = \frac{1}{\val(x)} \sum_{y \in V, d_\Gamma(x, y) \leq 1} a(x, y) \varphi(y)$$
where
\[ a(x,y) = \left| \{ \alpha \in E \mid \text{ep}(\alpha) = \{x,y\} \} \right|. \]

The notation in the first formula for \( M \phi \) should be clear: the value at \( x \) of the function \( M \phi \) is the average, over all edges of which \( x \) is an extremity, of the values of \( \phi \) at the “other” extremity, \emph{including} the values \( \phi(x) \) corresponding to the possible loops at \( x \).

For instance, if \( \Gamma \) is the action graph of Example 2.3.13, we have
\[
(M \phi)(3) = \frac{1}{3}(\phi(3) + \phi(2) + \phi(4)),
\]
and
\[
(M \phi)(1) = \frac{1}{3}(2\phi(2) + \phi(5)).
\]

Note that the condition \( d_\Gamma(x,y) \leq 1 \) can be omitted in the second expression for \( M \phi \), since the quantity
\[ a(x,y) = \left| \{ \alpha \in E \mid \text{ep}(\alpha) = \{x,y\} \} \right| \]
is zero unless \( d_\Gamma(x,y) \leq 1 \). In terms of transition probabilities (3.8), we can write
\[
(M \phi)(x) = \sum_{y \in V} P(x,y) \phi(y).
\]

**Proof.** One proof is to say that this statement is linear in terms of \( \phi \), and that the defining condition (3.7) of a random walk implies that the identity holds for \( \phi \) the characteristic function of a singleton \( y \). Since these functions form a basis of \( L^2(\Gamma) \), the result follows.

However, it is worth spelling this out as an exercise in handling this type of averages: we find
\[
E(\phi(X_{n+1})) = \sum_{y \in V} \phi(y) P(X_{n+1} = y)
= \sum_{y \in V} \phi(y) \sum_{x \in V} P(X_n = x) P(X_{n+1} = y \mid X_n = x)
= \sum_{y \in V} \phi(y) \frac{1}{\text{val}(x)} \sum_{x \in V} P(X_n = x) \left| \{ \alpha \in E \mid \text{ep}(\alpha) = \{x,y\} \} \right|
= \sum_{x \in V} \psi(x) P(X_n = x) = E(\psi(X_n))
\]
where
\[
\psi(x) = \frac{1}{\text{val}(x)} \sum_{y \in V} \phi(y) \left| \{ \alpha \in E \mid \text{ep}(\alpha) = \{x,y\} \} \right| = (M \phi)(x).
\]

This is the second formula for \( M \phi \). If we sum over edges instead of vertices, and denote temporarily
\[
e_x(\alpha) = \begin{cases} y & \text{if } \text{ep}(\alpha) = \{x,y\}, \quad y \neq x \\ x & \text{if } \text{ep}(\alpha) = \{x\}, \end{cases}
\]
for edges of which \( x \) is an extremity, we get
\[
\sum_{y \in V} \phi(y) \left| \{ \alpha \in E \mid \text{ep}(\alpha) = \{x,y\} \} \right| = \sum_{\alpha \in E \atop x \in \text{ep}(\alpha)} \phi(e_x(\alpha)),
\]
which is the first formula. \( \square \)
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 3.2.14.** Let \( \Gamma = (V, E, \text{ep}) \) be a finite graph, and let \( (X_n) \) be a random walk on \( \Gamma \). For any function \( \varphi \in L^2(\Gamma) \) and \( n \geq 1 \), we have

\[
E(\varphi(X_n)) = E((M^n\varphi)(X_0)),
\]

In particular, if the random walk starts at \( x_0 \), we have

\[
E(\varphi(X_n)) = (M^n\varphi)(x_0).
\]

We are now led to the investigation of the averaging operator \( M \). This will be done by diagonalizing it, and this is where the specific shape of the inner product on \( L^2(\Gamma) \) becomes important (the self-adjoint property depends on this choice!)

**Proposition 3.2.15 (Spectral properties of the Markov operator).** Let \( \Gamma = (V, E, \text{ep}) \) be a finite graph and let \( M \) be the Markov averaging operator for \( \Gamma \).

1. For any function \( \varphi \in L^2(\Gamma) \), we have

\[
\langle (\text{Id} - M)\varphi, \varphi \rangle = \frac{1}{2N} \sum_{x,y \in V} a(x,y)|\varphi(x) - \varphi(y)|^2,
\]

2. The operator \( M \) is self-adjoint, bounded from above by the identity and below by minus the identity, and has norm \( 1 \), i.e.,

\[
\langle M\varphi_1, \varphi_2 \rangle = \langle \varphi_1, M\varphi_2 \rangle
\]

for any functions \( \varphi_1, \varphi_2 \in L^2(\Gamma) \), moreover

\[
-\langle \varphi, \varphi \rangle \leq \langle M\varphi, \varphi \rangle \leq \langle \varphi, \varphi \rangle
\]

for all \( \varphi \in L^2(\Gamma) \), and finally

\[
\sup_{\varphi \neq 0} \frac{\|M\varphi\|_{\Gamma}}{\|\varphi\|_{\Gamma}} = 1.
\]

In particular, \( M \) is diagonalizable in an orthonormal basis of \( L^2(\Gamma) \), its eigenvalues are real numbers, and all eigenvalues have absolute value at most 1.

3. The \( 1 \)-eigenspace \( \ker(M - 1) \) of \( M \) has dimension equal to the number of connected components of \( \Gamma \), and is spanned by the characteristic functions of these connected components. In particular, \( \ker(M - 1) = C \), spanned by constant functions, if \( \Gamma \) is connected.

4. If \( \Gamma \) is connected, the \( (-1) \)-eigenspace \( \ker(M + 1) \) is zero unless \( \Gamma \) 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 \( V \).
Proof. We start by proving the self-adjointness, which is a key property. In fact, it is quite easy: we have by definition
\[
\langle M\varphi_1, \varphi_2 \rangle = \frac{1}{N} \sum_{x \in V} \text{val}(x) (M\varphi_1)(x)\overline{(M\varphi_2)(x)} \\
= \frac{1}{N} \sum_{x \in V} \varphi_2(x) \sum_{y \in V} \varphi_1(y)a(x, y) \\
= \frac{1}{N} \sum_{x, y \in V} a(x, y)\varphi_1(y)\overline{\varphi_2(x)}
\]
and since \(a(x, y) = a(y, x)\), this is also \(\langle \varphi_1, M\varphi_2 \rangle\).

From the self-adjointness, it already follows that \(M\) is diagonalizable in an orthonormal basis of \(L^2(\Gamma)\), and that the eigenvalues of \(M\) are real.

We will now prove the formulas (3.14) and (3.15), since they are very useful for later applications, and lead immediately to (3.16). Both are very similar and we deal only with the first one. It is easiest to start from the right-hand side: by expanding the square, we get
\[
\sum_{x, y \in V} a(x, y)|\varphi(x) - \varphi(y)|^2 = \sum_{x, y \in V} a(x, y)|\varphi(x)|^2 + \sum_{x, y \in V} a(x, y)|\varphi(y)|^2 \\
- 2 \sum_{x, y \in V} a(x, y)\varphi(x)\overline{\varphi(y)} \\
= 2 \sum_{x, y \in V} a(x, y)|\varphi(x)|^2 - 2 \sum_{x, y \in V} a(x, y)\varphi(x)\overline{\varphi(y)} \\
= 2N(\langle \varphi, \varphi \rangle - \langle M\varphi, \varphi \rangle)
\]
where we used the symmetry of the adjacency matrix as well as (3.17) and the formula
\[
\|\varphi\|^2 = \frac{1}{N} \sum_{x \in V} \text{val}(x)|\varphi(x)|^2 = \frac{1}{N} \sum_{x \in V} |\varphi(x)|^2 \sum_{y \in V} a(x, y)
\]
(3.18)
\[
= \frac{1}{N} \sum_{x, y \in V} a(x, y)|\varphi(x)|^2.
\]

These formulas (3.14) and (3.15) give (3.16), as already observed, using the positivity of squares. From this, we get
\[
|\langle M\varphi, \varphi \rangle| \leq \|\varphi\|^2
\]
for all \(\varphi \in L^2(\Gamma)\), and since it is standard that
\[
\|M\| = \sup_{\varphi \neq 0} \frac{\|\langle M\varphi, \varphi \rangle\|}{\|\varphi\|^2}
\]
for a self-adjoint operator, this gives \(\|M\| \leq 1\).

Our next step is to investigate \(\ker(M - 1)\). We do this using (3.14), though there is a nice “geometric” proof, which is explained in 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 (3.14). By positivity, this is equivalent with
\[
\varphi(x) = \varphi(y)
\]
47
whenever $a(x,y) \neq 0$, i.e., $\varphi$ 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 $\varphi$ is constant on each connected component of $\Gamma$. The converse is easy: if $\varphi$ is of this form, it does satisfy $M\varphi = \varphi$, and hence ker($M - 1$) is the space spanned by characteristic functions of connected components in the graph.

We now 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 $\Gamma$ is bipartite, the function $\varepsilon_{\pm}$ defined in the statement of the theorem is indeed in ker($M+1$).

We now proceed to show that this is the only case that arises.

If $M\varphi = -\varphi$, we get from (3.15) this time that

$$\varphi(x) = -\varphi(y)$$

for all $x$ and $y$ connected by an edge. If $\gamma : P_2 \rightarrow \Gamma$ is any path of length 2 with $\gamma(0) = x$, $\gamma(2) = y$, it follows that

$$\varphi(x) = -\varphi(\gamma(1)) = \varphi(y).$$

Similarly, $y = \gamma(2k)$ for a path of even length $2k$. Now we fix some $x_0 \in V$, and let $W$ be the set of vertices in $\Gamma$ which are the other extremity of a path $\gamma : P_{2k} \rightarrow \Gamma$ of even length with $\gamma(0) = x_0$ (in particular, $x_0 \in W$ using a path of length 0). We see that $\varphi$ is constant, equal to $\varphi(x_0)$, on all of $W$. If $W = V$, it follows that $\varphi$ 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 $\alpha$; (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 $\Gamma$ 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 below 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 finishes the proof of bipartiteness of $\Gamma$. It is now easy to finish determining $\varphi$: 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 $\varphi(x_0)\varepsilon_{\pm}$.

□

**Exercise 3.2.16.** The following proof of $\|M\| \leq 1$, which does not use (3.16) is also useful to keep in mind. (It is easier to remember, but one should remember that the comparison bounds (3.16) are extremely useful, and are stronger than the assertion $\|M\| \leq 1$).

(1) Explain why the norm of $M$ is the maximum of the absolute values of its eigenvalues.

(2) If $\lambda$ is an eigenvalue, show directly that $|\lambda| \leq 1$. [Hint: Use the maximum norm instead of the $L^2$-norm.]

**Exercise 3.2.17** (Maximum modulus principle). This exercise discusses the “geometric” computation of ker($M - 1$).
(1) Show that if $\varphi$ is the characteristic function of a connected component of $\Gamma$, 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 $\Gamma$.

(3) Let $W \subset V$ be a connected component. Let $\varphi$ 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 $\varphi$ is equal to $m$ on all of $W$ and conclude.

(5) Using similar methods, determine $\ker(M + 1)$.

**Exercise 3.2.18 (Both sides have equal weight).** Let $\Gamma$ be a connected non-empty finite bipartite graph, partitioned as $V = V_0 \cup V_1$. 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 connected graph $\Gamma$. We define first the relevant invariant:

**Definition 3.2.19 (Equidistribution radius).** Let $\Gamma = (V, E, ep)$ be a connected, non-empty, finite graph. The equidistribution radius of $\Gamma$, denoted $\varrho_\Gamma$, is the maximum of the absolute values $|\lambda|$ for $\lambda$ an eigenvalue of $M_\Gamma$ which is different from $\pm 1$. Equivalently, $\varrho_\Gamma$ is the spectral radius of the restriction of $M$ to the subspace

$$L^2_0(\Gamma) = (\ker(M - 1) \oplus \ker(M + 1))^\perp,$$

i.e., (1) if $\Gamma$ is not bipartite, to the space of $\varphi \in L^2(\Gamma)$ such that

$$\langle \varphi, 1 \rangle = \frac{1}{N} \sum_{x \in V} \text{val}(x) \varphi(x) = 0,$$

and (2) if $\Gamma$ is bipartite with bipartite partition $V_0 \cup V_1 = V$, to the space of $\varphi \in L^2(\Gamma)$ such that

$$\frac{1}{N} \sum_{x \in V} \text{val}(x) \varphi(x) = 0, \quad \frac{1}{N} \sum_{x \in V_0} \text{val}(x) \varphi(x) = \frac{1}{N} \sum_{x \in V_1} \text{val}(x) \varphi(x).$$
The equivalence of the stated definitions of $\varrho_\Gamma$, and of the subspace $L^2_0(\Gamma)$, are direct consequences of Proposition 3.2.15 (taking into account the assumption that $\Gamma$ is connected). The following are also almost part of the definition:

**Lemma 3.2.20.** Let $\Gamma = (V, E, \text{ep})$ be a connected, non-empty, finite graph. We have $0 \leq \varrho_\Gamma < 1$ and $\varrho_\Gamma$ is given by

$$\varrho_\Gamma = \max_{0 \neq \varphi \in L^2_0(\Gamma)} \frac{|\langle M \varphi, \varphi \rangle|}{\|\varphi\|^2}. \quad (3.19)$$

**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 $L^2_0(\Gamma)$ of the eigenvalues $\pm 1$, all eigenvalues (the largest of which is $\varrho_\Gamma$) are $< 1$.

Similarly, the self-adjoint operator $M$ stabilizes the space $\ker(M - 1) \oplus \ker(M + 1)$, and hence acts on $L^2_0(\Gamma)$. On this space, it remains self-adjoint, and its norm is $\varrho_\Gamma < 1$, so that the formula (3.19) follows by looking at an eigenvalue decomposition of a function $\varphi \in L^2_0(\Gamma)$. □

**Example 3.2.21.** The simplest example of computation of $\varrho_\Gamma$ is the case of the $m$-cycle $C_m$, $m \geq 2$. We will compute all eigenvalues of $M$ and describe the eigenfunctions in that case. For a function $\varphi : \mathbb{Z}/m\mathbb{Z} \rightarrow \mathbb{C}$ (recall that we use $\mathbb{Z}/m\mathbb{Z}$ as vertex set for the cycle) and a vertex $x \in \mathbb{Z}/m\mathbb{Z}$, 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 $\mathbb{Z}/m\mathbb{Z}$, which is the linear map

$$L^2(C_m) \rightarrow L^2(C_m) \quad \varphi \mapsto \hat{\varphi}$$

defined by

$$\hat{\varphi}(a) = \frac{1}{m} \sum_{x \in \mathbb{Z}/m\mathbb{Z}} \varphi(x) e\left(-\frac{ax}{m}\right)$$

for $a \in \mathbb{Z}/m\mathbb{Z}$, where $e(z) = e^{2\pi i z}$ for $z \in \mathbb{C}$. We can also write

$$\hat{\varphi}(a) = \langle \varphi, \chi_a \rangle$$

where $\chi_a(x)$ is the function $\chi_a(x) = e(ax/m)$, which is distinguished by being a character of $\mathbb{Z}/m\mathbb{Z}$, i.e., by $\chi_a(x + y) = \chi_a(x)\chi_a(y)$.

Because of this, we see by a change of variable that if $\psi(x) = \varphi(x + b)$ for some fixed $b \in \mathbb{Z}/m\mathbb{Z}$, we have

$$\hat{\psi}(a) = \frac{1}{m} \sum_{x \in \mathbb{Z}/m\mathbb{Z}} \varphi(x + b) e\left(-\frac{ax}{m}\right) = \chi_a(b) \frac{1}{m} \sum_{y \in \mathbb{Z}/m\mathbb{Z}} \varphi(y) e\left(-\frac{ay}{m}\right),$$

i.e., $\hat{\psi} = \chi_a(b) \hat{\varphi}$. In particular, it follows that

$$\overline{M \varphi}(a) = \frac{\chi_a(1) + \chi_a(-1)}{2} \hat{\varphi}(a) = \cos\left(\frac{2\pi a}{m}\right) \hat{\varphi}(a).$$

In other words, $M$ acts diagonally on Fourier transforms. But the Fourier transform is an isomorphism – indeed, we have the formula

$$\varphi(x) = \sum_{a \in \mathbb{Z}/m\mathbb{Z}} \hat{\varphi}(a) \chi_a(x).$$
which means that $\varphi$ is the Fourier transform of $a \mapsto m\hat{\varphi}(-a)$, and so we have found an explicit diagonalization of $M$ for the cycle graphs. In fact, we find also directly that $M\chi_b = \cos\left(\frac{2\pi b}{m}\right)\chi_b$.

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 $\chi_b$ and $\chi_{-b}$), while if $m$ is even, all eigenvalues except for 1 and $-1$ (which have 1-dimensional eigenspaces, in the second case because $\mathbb{Z}/m\mathbb{Z}$ is then bipartite) have a 2-dimensional eigenspace. In any case, we get $\varrho_{C_m} = \cos\left(\frac{2\pi}{m}\right) = 1 - \frac{2\pi^2}{m^2} + O(m^{-4})$ for $m \geq 2$.

**Exercise 3.2.22.** Let $G_3$ be the Cayley graph $\mathcal{G}(S_3, S_3)$ 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 3.2.23 (Convergence to equilibrium in random walks).** Let $\Gamma = (V, E, \epsilon p)$ be a connected, non-empty, finite graph, and let $(X_n)$ be a random walk on $\Gamma$.

(1) If $\Gamma$ is not bipartite, then for any function $\varphi \in L^2(\Gamma)$, we have

$$\left|\mathbb{E}(\varphi(X_n)) - \langle \varphi, 1 \rangle\right| \leq g_n^0\left(\frac{N}{v_-}\right)^{1/2}\|\varphi\|,$$

where $v_- = \min \text{val}(x)$, and in particular

$$\lim_{n \rightarrow +\infty} \mathbb{P}(X_n = x) = \mu_\Gamma(x) = \frac{\text{val}(x)}{N}$$

for all $x \in V$.

(2) If $\Gamma$ is bipartite with partition $V = V_0 \cup V_1$, then for any function $\varphi \in L^2(\Gamma)$, we have

$$\left|\mathbb{E}(\varphi(X_n)) - \left\{m_0 + m_1 + (-1)^n(m_0 - m_1)(p_0 - p_1)\right\}\right| \leq g_n^0\left(\frac{N}{v_-}\right)^{1/2}\|\varphi\|,$$

where

$$p_0 = \mathbb{P}(X_0 \in V_0), \quad p_1 = \mathbb{P}(X_0 \in V_1),$$

$$m_0 = \frac{1}{N} \sum_{x \in V_0} \text{val}(x)\varphi(x), \quad m_1 = \frac{1}{N} \sum_{x \in V_1} \text{val}(x)\varphi(x).$$

**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 $\varphi$ and, by definition, we have $\varphi_0 \in L^2(\Gamma)$. Applying Corollary 3.2.14 and the fact that $\alpha$ is an eigenfunction of $M$ with eigenvalue 1, we get

$$\mathbb{E}(\varphi(X_n)) = \mathbb{E}(\langle M^n\varphi \rangle(X_0)) = \alpha + \mathbb{E}(\langle M^n\varphi_0 \rangle(X_0)).$$
By writing
\[ E((M^n \varphi_0)(X_0)) = \sum_{x \in V} P(X_0 = x)(M^n \varphi_0)(x), \]
we get
\[ |E((M^n \varphi_0)(X_0))| \leq \|(M^n \varphi_0)\|_\infty, \]
and we are almost done. The last step is to compare this maximum norm with the \( L^2 \) norm, which we do with (3.12), from which we get
\[ |E((M^n \varphi_0)(X_0))| \leq \left( \frac{N}{v} \right)^{1/2} \|(M^n \varphi_0)\|. \]

Now the definition of \( \varrho_1 \) (and the fact that \( M \) sends \( L^2_0(\Gamma) \) to itself) leads immediately, by induction, to
\[ \|(M^n \varphi_0)\| \leq \varrho^n \|\varphi_0\|, \]
from which the inequality (3.20) follows. The limit (3.21) is simply the special case when \( \varphi \) is the characteristic function of \( x \in V \), in which case \( \langle \varphi, 1 \rangle = \text{val}(x)/N \).

(2) The bipartite case is essentially 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 \( \varphi_0 \in L^2_0(\Gamma) \), and obtain
\[ E(\varphi(X_n)) = E((M^n \varphi)X_0) = \langle \varphi, 1 \rangle + (-1)^n \langle \varphi, \varepsilon_\pm \rangle E(\varepsilon_\pm(X_0)) + E((M^n \varphi_0)X_0). \]

The last term is estimated exactly as before:
\[ |E((M^n \varphi_0)X_0)| \leq \left( \frac{N}{v} \right)^{1/2} \varrho^n \|\varphi\|, \]
using \( \varphi_0 \in L^2_0(\Gamma) \). We now note that
\[ \langle \varphi, 1 \rangle = m_0 + m_1, \quad \langle \varphi, \varepsilon_\pm \rangle = m_0 - m_1 \]
and that
\[ E(\varepsilon_\pm(X_0)) = p_0 - p_1, \]
to deduce that
\[ \langle \varphi, 1 \rangle + (-1)^n \langle \varphi, \varepsilon_\pm \rangle E(\varepsilon_\pm(X_0)) = m_0 + m_1 + (-1)^n(m_0 - m_1)(p_0 - p_1). \]

\[ \square \]

**Remark 3.2.24.** (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 \( E(\varphi(X_n)) \) by expanding \( \varphi \) 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 \( E(\varphi(X_n)) \) becomes
\[ m_0 + m_1 + (-1)^n(m_0 - m_1) = (1 + (-1)^n)m_0 + (1 - (-1)^n)m_1 \]
\[ = \begin{cases} 2m_0 & \text{if } n \text{ is even}, \\ 2m_1 & \text{if } n \text{ is odd}. \end{cases} \]

In this case, the sequence \( E(\varphi(X_n)) \) does not converge in general (unless \( m_0 = m_1 \)): it oscillates between the even terms which converge to twice the average of \( \varphi \) on \( V_0 \), and the odd ones which converge to twice the average of \( \varphi \) on \( V_1 \).
In particular, if \( \varphi \) 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 \( 2\mu (x_1) \). The factor 2 can be interpreted as follows: since we know \textit{a priori} that \( X_n \) is in \( V_1 \) for \( n \) odd, and \( \mu (V_1) = \frac{1}{2} \) (Exercise 3.2.18), 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 3.2.25 (Time to reach equilibrium).** For the characteristic function \( \delta_x \) of a fixed vertex, we have

\[
\| \delta_x \|^2 = \mu (x) = \frac{\text{val}(x)}{N},
\]

so that, if \( \Gamma \) is not bipartite, the precise statement is

\[
\left| P(X_n = x) - \frac{\text{val}(x)}{N} \right| \leq \left( \frac{\text{val}(x)}{v_-} \right)^{1/2} \varrho_1^n
\]

for \( n \geq 1 \). 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 \( \frac{1}{2} \mu (x) \): for such \( n \), the probability \( P(X = n = x) \) 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

\[
\varrho_1^n \leq \frac{(\text{val}(x)v_-)^{1/2}}{2N},
\]

and in the case of a \( k \)-regular graph, this means

\[
n \geq \frac{\log(2|V|)}{\log(\varrho_1^{-1})}.
\]

Hence, for \( n \) larger than some (possibly large, depending on how close \( \varrho_1 \) is to 1) multiple of \( \log |V| \), a random walk \( (X_n) \) becomes significantly well-distributed. This type of considerations turns out to play an important role in understanding the arguments of the next chapter.

We see from this corollary that \( \varrho_1 \) controls the rate of convergence of a random walk on \( \Gamma \) to the normalized graph measure \( \mu_\Gamma \). The intuition that the speed of convergence should be greater when the expansion constant is also large leads to the suspicion that bounding \( \varrho_1 \) away from 1 should be related to bounding \( h(\Gamma) \) away from 0. This is indeed the case up to a minor technical point: being or not bipartite (or “very close”, in the sense that there is an eigenvalue of \( M \) 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 whimsical definition:

**Definition 3.2.26 ("Absolute Expanders").** Let \( (\Gamma_i) \) be a family of finite, non-empty, connected graphs \( \Gamma_i = (V_i, E_i, \epsilon_p) \) with bounded maximal valency \( \leq v \), such that the number of vertices tends to infinity; in the same sense as in Definition 3.1.8. We say that \( (\Gamma_i) \) is a family of \textit{absolute expanders} if and only if there exists \( \varrho < 1 \) such that

\[
\varrho_1 \leq \varrho < 1
\]

for all \( i \in I \). When this is true, we say that \( (\varrho, v) \) are \textit{equidistribution parameters} for the absolute expander family.

The precise link between expanders and absolute expanders is the following result:
Theorem 3.2.27 (Random walk definition of expanders). (1) A family of absolute expanders is an expander family.

(2) Conversely, let \((\Gamma_i)\) be an expander family with \(\Gamma_i = (V_i, E_i, \varphi)\). Let \(\tilde{\Gamma}_i\) be the “relaxed” graphs obtained from \(\Gamma_i\) by adding a loop at each vertex, i.e.,

\[
\tilde{\Gamma}_i = (V_i, E_i \cup V_i, \varphi')
\]

with \(\varphi'(\alpha) = \varphi(\alpha)\) for \(\alpha \in E_i\) and \(\varphi'(x) = \{x\}\) for \(x \in V_i\). Then \((\tilde{\Gamma}_i)\) is a family of absolute expanders.

Remark 3.2.28. 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 \(E(W_1, W_2)\) for any subsets \(W_1, W_2 \subset V\), we have \(h(\Gamma_i) = h(\tilde{\Gamma}_i)\).

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 \((\Gamma_i)\) is an expander family if and only if \((\tilde{\Gamma}_i)\) is an expander family. On the other hand, because we added loops, \(\tilde{\Gamma}_i\) 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 \(\varrho\Gamma\), 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, \(\varrho\Gamma\) is either the largest eigenvalue \(<1\) of \(M\), or the negative of the smallest eigenvalue which is \(> -1\). A very convenient way to express this is to basically give names to these two quantities, or rather to the distance to 1 or \(-1\):

Definition 3.2.29 (Normalized spectral gaps). Let \(\Gamma\) be a finite non-empty connected graph. The normalized spectral gap \(\lambda_1(\Gamma)\) is the smallest non-zero eigenvalue of \(\text{Id} - M\). The complementary normalized spectral gap \(\mu_1(\Gamma)\) is the smallest non-zero eigenvalue of \(\text{Id} + M\).

The largest eigenvalue \(<1\) of \(M\) is therefore \(1 - \lambda_1\), and the smallest \(> -1\) is \(-1 + \mu_1\). Thus we have

\[
\varrho_{\Gamma} = \max(1 - \lambda_1(\Gamma), \mu_1(\Gamma) - 1).
\]

Moreover we have

\[
\lambda_1(\Gamma) = \min_{0 \neq \varphi \perp 1} \frac{\langle (\text{Id} - M)\varphi, \varphi \rangle}{\langle \varphi, \varphi \rangle}
= \min_{\varphi \text{ not constant}} \frac{\langle (\text{Id} - M)\varphi, \varphi \rangle}{\|\varphi - \langle \varphi, 1 \rangle 1\|_2^2},
\]

where the equality between these two characterizations follows from the fact that

\[
\langle (\text{Id} - M)\varphi, \varphi \rangle = \langle (\text{Id} - M)\varphi_0, \varphi_0 \rangle
\]

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 (3.14) which shows that the numerator is determined by the difference in values of \(\varphi\) on adjacent
vertices, so that suitable choices of $\varphi$ lead to the quantity $\mathcal{E}(W)$, as the following lemma shows:

**Lemma 3.2.30.** Let $\Gamma$ be a finite non-empty connected graph. Let $W \subset V$ be a subset of vertices, $W' = V - W$, and let

$$\varphi = 1_W - \mu_{\Gamma}(W),$$

the “centered” characteristic function of $W$. Then

$$\langle (I - M) \varphi, \varphi \rangle = \langle (I - M) 1_W, 1_W \rangle = \frac{|\mathcal{E}(W)|}{N}$$

and

$$\|\varphi\|^2 = \mu_{\Gamma}(W) \mu_{\Gamma}(W').$$

**Proof.** The formula (3.14) gives

$$\langle (I - M) \varphi, \varphi \rangle = \frac{1}{2N} \sum_{x,y \in V} a(x, y)(\varphi(x) - \varphi(y))^2$$

which makes it clear that the constant term $\mu_{\Gamma}(W')$ can be removed, so that

$$\langle (I - M) \varphi, \varphi \rangle = \frac{1}{2N} \sum_{x,y \in V} a(x, y)(1_W(x) - 1_W(y))^2.$$  

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

$$\langle (I - M) \varphi, \varphi \rangle = \frac{1}{N} \sum_{x \in W, y \notin W} a(x, y) = \frac{|\mathcal{E}(W)|}{N}.$$  

The formula for $\|\varphi\|^2$ is a simple computation: since $\varphi$ is orthogonal to constants, we have

$$\|\varphi\|^2 = \|1_W\|^2 - \mu_{\Gamma}(W)^2 = \mu_{\Gamma}(W) - \mu_{\Gamma}(W')^2 = \mu_{\Gamma}(W) \mu_{\Gamma}(W').$$

We can now immediately prove (1) in Theorem 3.2.27, since it follows from the next proposition:

**Proposition 3.2.31 (Expansion and equidistribution).** Let $\Gamma = (V, E, \xi_{\Gamma})$ be a connected, non-empty, finite graph. We have

$$1 - \varrho_{\Gamma} \leq \lambda_1(\Gamma) \leq \left(\frac{2v_+}{v_-^2}\right) h(\Gamma)$$

where, as before, we denote

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

In particular, if $\Gamma$ is $k$-regular, we have

$$1 - \varrho_{\Gamma} \leq \lambda_1(\Gamma) \leq \frac{2}{k} h(\Gamma).$$

55
Proof. Because of (3.24), we can estimate \( \lambda_1(\Gamma) \) from above by
\[
\lambda_1(\Gamma) \leq \frac{\langle (\text{Id} - M)\varphi, \varphi \rangle}{\langle \varphi, \varphi \rangle}
\]
for any suitable function with mean zero \( \varphi \). Applying the lemma with \( W \) a non-empty subset \( W \subset V \) with \( |W| \leq |\Gamma|/2 \) such that
\[
h(\Gamma) = \frac{|E(W)|}{|W|},
\]
we get
\[
\lambda_1(\Gamma) \leq \frac{|E(W)|}{N} \frac{1}{\|\varphi\|^2} = \frac{1}{N} \mu_\Gamma(W)\mu_\Gamma(W').
\]
We now use (3.11) in order to make the exact ratio \( |E(W)|/|W| \) appear, obtaining
\[
N\mu_\Gamma(W)\mu_\Gamma(W') \geq v_-|W| \times \frac{v_-|W'|}{v_+|V|} \geq \frac{v^2}{2v_+}|W|,
\]
and the inequality (3.26) follows.

Before going to the other direction, it is worth remarking that this is the right result to obtain lower bounds for the expansion constant of a graph. In Chapter 5, we will construct “deterministic” families of expander graphs using exactly this inequality.

Remark 3.2.32. This result is also useful numerically: since \( \lambda_1(\Gamma) \) is an eigenvalue of the linear operator \( \text{Id} - M \) acting on \( L^2(\Gamma) \), which is a finite-dimensional vector space, of dimension \( |V| \), the problem of determining \( \lambda_1(\Gamma) \) (or indeed \( \varrho_\Gamma \) 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 \( \varrho_\Gamma \) is the largest absolute value of any eigenvalue on \( L^2(\Gamma) \) also leads to the possibility of applying various approximation algorithms for this specific problem.

We can now start investigating the converse of (3.26), we may note already that it can not be a simple relation stating that \( \lambda_1 \) (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 (3.1) that \( h(C_m) \propto 1/m \) for \( m \) large, while \( 1 - \varrho_{C_m} \propto 1/m^2 \) by Example 3.2.21, which is much smaller. However, this is as bad as it can get:

Proposition 3.2.33 (Discrete Cheeger inequality). Let \( \Gamma = (V, E, ep) \) be a connected, non-empty, finite graph. We have
\[
h(\Gamma) \leq v_+\sqrt{2\lambda_1(\Gamma)}.
\]
We will basically follow an argument of L. Trevisan [48, Handout 4], which highlights a practical algorithmic interpretation of this inequality. The idea is to study the expansion of sets of the type
\[W_{\varphi, t} = \varphi^{-1}(-\infty, t], \quad \{ x \in V \mid \varphi(x) \leq t \}\]
for a real-valued function \( \varphi : V \to \mathbb{R} \) and a real number \( t \), and show that some of them satisfy
\[
\frac{|E(W_{\varphi, t})|}{|W_{\varphi, t}|} \leq v_+\sqrt{2\lambda_1(\Gamma)},
\]
(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 \( E(W_{\varphi, t}) \) for a given function, and deduce the existence of sets with certain expansion ratios. The following lemma performs this computation:

56
Lemma 3.2.34 (Expansion of sublevel sets). Let $\Gamma = (V, E, ep)$ be a finite non-empty connected graph and let $\varphi : V \rightarrow \mathbb{R}$ be a real-valued, non-constant, function on $V$. Let 

$$a = \min_{x \in V} \varphi(x), \quad b = \max_{x \in V} \varphi(x),$$

and let $t_0 \in \mathbb{R}$ be such that

$$|W_{\varphi,t}| \leq |V|/2$$

if and only if $t < t_0$.

Then for any choice of a probability measure $\nu$ with support equal to $[a, b]$ and without atoms, we can find $t \in \mathbb{R}$ such that either $W = W_{\varphi,t}$ or $W = V - W_{\varphi,t}$ satisfies $|W| \leq |V|/2$ and

$$\frac{|\mathcal{E}(W)|}{|W|} \leq \frac{A}{B}$$

where

$$A = \frac{1}{2} \sum_{x,y \in V} a(x,y)\nu([\varphi(x), \varphi(y)]),$$

$$B = \sum_{x \in V} \nu([t_0, \varphi(x)]).$$

Proof. We denote $W_t = W_{\varphi,t}$ for simplicity. Noting that an edge $\alpha$ with $ep(\alpha) = \{x, y\}$ is in $\mathcal{E}(W_t)$ if and only if $t$ lies in the interval $I_\alpha$ between $\varphi(x)$ and $\varphi(y)$ where the largest is excluded, i.e.,

$$I_\alpha = [\min(\varphi(x), \varphi(y)), \max(\varphi(x), \varphi(y))],$$

we compute the average of $|\mathcal{E}(W_t)|$ as

$$\int_{\mathbb{R}} |\mathcal{E}(W_t)| d\nu(t) = \sum_{\alpha \in E} \nu(t \mid t \text{ is in the interval } I_\alpha)$$

$$= \sum_{\alpha \in E} \nu(I_\alpha) = \frac{1}{2} \sum_{x,y \in V} a(x,y)\nu([\varphi(x), \varphi(y)]) = A$$

since $\nu$ has no atom.

We want to compare this with the number of elements of $W_t$, or rather with the minimum $\min(|W_t|, |V - W_t|) \leq |V|/2$ (with the idea of using either $W_t$ or $V - W_t$ to test the expansion constant).

Since the size of $W_t$ is non-decreasing as a function of $t$, the real number $t_0$ such that $|W_t| \leq |V|/2$ if and only if $t < t_0$ exists. Then (again using the fact that $\nu$ has no atoms) we have

$$\int_{\mathbb{R}} \min(|W_t|, |V - W_t|) d\nu(t) = \int_{t < t_0} |W_t| d\nu(t) + \int_{t \geq t_0} |V - W_t| d\nu(t)$$

$$= \sum_{x \in V} \nu(t \mid \varphi(x) \leq t < t_0) + \sum_{x \in V} \nu(t \mid t_0 \leq t \leq \varphi(x))$$

$$= \sum_{x \in V} \nu([t_0, \varphi(x)]) = B.$$

---

2 This means that $t_0$ is the “median” of the values of $\varphi$.

3 We use the convention that $\nu([a,b]) = \nu([b,a])$ if $a > b$.  

57
We now argue simply that since
\[ \int_{\mathbb{R}} \left( B |\mathcal{E}(W_t)| - A \min(|W_t|, |V - W_t|) \right) d\nu(t) = 0, \]
there must exist some \( t \in [a, b] \) for which
\[ B |\mathcal{E}(W_t)| - A \min(|W_t|, |V - W_t|) \leq 0, \]
which is the desired conclusion!

We are now led to an attempt to select a measure \( \nu \) and then find a function \( \varphi \) 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 which case
\begin{equation}
A = \frac{1}{2} \sum_{x,y \in V} a(x, y) \left| \varphi(x) - \varphi(y) \right| \frac{1}{b - a}, \quad B = \sum_{x \in V} \left| \varphi(x) - t_0 \right| \frac{1}{b - a},
\end{equation}
and the problem looks similar, in a rather more \( L^1 \)-ish sense, to the computation of \( \lambda_1 \) using the minimization (3.24). 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 also mentioned by Trevisan [49]; but see Example 3.2.35 below for an example, and Proposition 3.4.8 for a case where this is sharper than Cheeger’s inequality.) So we use instead, as in [48], the measure \( \nu \) 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
\[ \nu([t_0, \varphi(x)]) = \frac{1}{2S} \left| \varphi(x) - t_0 \right|^2 \]
for all \( x \) and a second’s thought shows that
\[ \nu(|\varphi(x), \varphi(y)|) \leq \frac{1}{2S} \left| \varphi(x) - \varphi(y) \right| \times (|\varphi(x) - t_0| + |\varphi(y) - t_0|). \]

Hence we find in this way a set \( W \) for which
\[ h(\Gamma) \leq \frac{|\mathcal{E}(W)|}{|W|} \leq \frac{A'}{B'} \]
where
\[ A' = \frac{1}{2} \sum_{x,y \in V} a(x, y) \left\{ |\varphi(x) - t_0| + |\varphi(y) - t_0| \right\} |\varphi(x) - \varphi(y)|, \]
\[ B' = \sum_{x \in V} \left| \varphi(x) - t_0 \right|^2. \]

We can now estimate further in terms of quantities related to \( M \). First, we write
\[ B' = \sum_{x \in V} \left| \varphi(x) - t_0 \right|^2 \geq \frac{1}{v_+} \sum_{x \in V} \text{val}(x) |\varphi(x) - t_0|^2 = \frac{N}{v_+} \left\| \varphi - t_0 \right\|^2 \]
while, by the Cauchy-Schwarz inequality and the formulas (3.14) and (3.15), we have
\[ (A')^2 \leq \frac{1}{2} \sum_{x,y} a(x, y) |\varphi(x) - \varphi(y)|^2 \left( \frac{1}{2} \sum_{x,y} a(x, y) \left\{ |\varphi(x) - t_0| + |\varphi(y) - t_0| \right\}^2 \right) = N \langle (\text{Id} - M)\varphi, \varphi \rangle \times N \langle (\text{Id} + M)\varphi - t_0, |\varphi - t_0| \rangle. \]
Since \( \| \operatorname{Id} + M \| \leq 2 \), we obtain
\[
\frac{A'}{B'} \leq v_+ \left( \frac{2 \langle (\operatorname{Id} - M) \varphi, \varphi \rangle}{\| \varphi - t_0 \|^2} \right)^{1/2} \leq v_+ \left( \frac{2 \langle (\operatorname{Id} - M) \varphi, \varphi \rangle}{\| \varphi \|^2} \right)^{1/2}
\]
(because \( \varphi \) is the orthogonal projection of \( \varphi - t_0 \) to the orthogonal complement of the constants, we have \( \| \varphi - t_0 \| \geq \| \varphi \| \)).

We finally select \( \varphi \) to be an eigenfunction of \( \operatorname{Id} - M \) with eigenvalue \( \lambda_1 \), and we obtain the inequality
\[
h(\Gamma) \leq v_+ \sqrt{2 \lambda_1(\Gamma)}
\]
(note that there always exists a real-valued eigenfunction of \( \operatorname{Id} - M \), since the real and imaginary parts of an eigenfunction \( \varphi \) are still eigenfunctions with the same eigenvalue, and one at least must be non-zero if \( \varphi \neq 0 \ldots \)) This finishes the proof of the Cheeger inequality!

**Example 3.2.35** (The cycles again). Let \( \Gamma = C_m \) with \( m \geq 2 \). In Example 3.2.21, we have shown that \( \lambda_1(C_m) = 1 - \cos(2\pi/m) \sim (2\pi^2)/m^2 \) as \( m \to +\infty \). A real-valued \( \lambda_1 \)-eigenfunction is given by
\[
\varphi(x) = \Re \left( e \left( \frac{x}{m} \right) \right) = \cos \left( \frac{2\pi x}{m} \right)
\]
for \( x \in \mathbb{Z}/m\mathbb{Z} \). It follows that \( W_0 = \{ x \mid \varphi(x) \leq 0 \} \) is roughly the image modulo \( m \) of the integers between \( m/4 \) and \( 3m/4 \), which we saw 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 applications of Lemma 3.2.34 for this function, with the uniform probability measure, shows that the existence of some set \( W \) with
\[
h(C_m) \leq \frac{|E(W)|}{|W|} \leq \frac{A}{B}
\]
where, spelling out (3.28), we have
\[
A = \sum_{0 \leq x \leq m - 1} \left| \cos \left( \frac{2\pi x}{m} \right) - \cos \left( \frac{2\pi(x + 1)}{m} \right) \right|,
\]
\[
B = \sum_{0 \leq x \leq m - 1} \left| \cos \left( \frac{2\pi x}{m} \right) \right|.
\]

It is elementary (looking at the graph of the cosine) that \( A \) converges to 4 as \( m \) tends to infinity, while \( B \sim 2\pi m \). Thus the bound \( h(C_m) \leq A/B \sim 2\pi/m \) is of the right order of magnitude in that case.

We can now also conclude the proof of part (2) in Theorem 3.2.27. Given a family \( (\Gamma_i) \) of expanders, we see from Cheeger’s inequality that the relaxed graph satisfy
\[
v \sqrt{2 \lambda_1(\tilde{\Gamma}_i)} \geq h(\tilde{\Gamma}_i) = h(\Gamma_i).
\]

This shows that the normalized spectral gap is bounded away from zero. Hence it is now enough to prove that \( \tilde{\Gamma}_i \) can not have an eigenvalue too close to \(-1 \). But the
definition of $\tilde{\Gamma}_i$ with its added loops leads to the formula

$$\langle (Id + M_i)\varphi, \varphi \rangle = \frac{1}{2N_i} \sum_{x,y \in V_i} \tilde{a}(x,y)|\varphi(x) + \varphi(y)|^2$$

$$= \frac{1}{2N_i} \left( \sum_{x,y \in V_i} a(x,y)|\varphi(x) + \varphi(y)|^2 + 4 \sum_{x \in V_i} |\varphi(x)|^2 \right)$$

and since $\tilde{N}_i = N_i + |V_i| \leq 2N_i$ and $\text{val}(x) \leq v_+ \leq v$, we get by positivity

$$\langle (Id + M_i)\varphi, \varphi \rangle \geq \frac{1}{N} \sum_{x \in V_i} |\varphi(x)|^2 \geq \frac{1}{v} \|\varphi\|^2,$$

which implies that $\tilde{M}_i$ has no eigenvalue $<-1 + v^{-1}$. Hence we derive

$$\rho_{\tilde{\Gamma}_i} \leq 1 - \min\left(\frac{h^2}{2v}, \frac{1}{v}\right) < 1$$

for all $i$, giving equidistribution parameters of the relaxed graphs in terms of the expansion parameters $(h, v)$ of $(\Gamma_i)$. (Typically, $h^2/2$ is less than 1, of course, so we can replace this expression by $1 - h^2/(2v)$.)

**Example 3.2.36 (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 $(\Gamma_n)$ of bipartite expanders – their existence will be proved in Section 3.5. Then, for each $n$, we attach a single loop at some (arbitrarily chosen) vertex $x_n$ of $\Gamma_n$, obtaining a new sequence $(\Gamma'_n)$ 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 $\mu_1(\Gamma_n)$ and its characterization

$$\mu_1(\Gamma) = \min_{0 \neq \varphi \in L^2(\Gamma)} \frac{\langle (Id + M)\varphi, \varphi \rangle}{\langle \varphi, \varphi \rangle}$$

for a non-bipartite graph, so there is no condition required about $\varphi$, in the absence of an eigenfunction of eigenvalue $-1$ of $M$.

For $\Gamma_n$, the function $\varepsilon_\pm$ defined in Proposition 3.2.15 minimizes this expression. For $\Gamma'_n$, it is natural enough to expect that it should be also close (if not equal!) to the minimum. We have, with obvious notation, $N'_n = N_n + 1$ and

$$\langle (Id + M'_n)\varphi, \varphi \rangle = \frac{1}{2N'_n} \sum_{x,y \in V_n} a(x,y)(\varepsilon_+(x) + \varepsilon_-(y))^2,$$

which only differs from the corresponding quantity for $\Gamma_n$ 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 $Id + M_n$, this gives

$$\langle (Id + M'_n)\varphi, \varphi \rangle = \frac{2}{N'_n} = \frac{2}{N_n + 1}.$$
On the other hand, since $|\epsilon_\pm(x)| = 1$, the norm squared of $\epsilon_\pm$ is still one, and we get
\[
\lim_{n \to +\infty} \mu_1(\Gamma_n) = 0
\]
since $N_n \geq |V_n| \to +\infty$. Hence the graphs $(\Gamma'_n)$ are not absolute expanders.

**Remark 3.2.37 (Trivial lower bound).** From Lemma 3.1.4 for $h(\Gamma)$ and Proposition 3.2.33, we see that there is a universal “trivial” lower bound
\[
(3.29) \quad \lambda_1(\Gamma) \geq \frac{1}{2v_+^2 |\Gamma|^2}
\]
for a finite connected graph $\Gamma$. The example of the cycles $C_m$ with $v_+ = 2$ and $\lambda_1(C_m) \asymp m^{-2} = |C_m|^{-2}$ shows that the order of magnitude can not be improved.

### 3.3. 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 $\operatorname{Id} - M$. 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 3.3.1.** Let $\Gamma = (V,E,ep)$ be a finite graph. The normalized Laplace operator of $\Gamma$, denoted $\Delta_\Gamma$, is the linear operator
\[
\Delta_\Gamma \left\{ \begin{array}{c}
L^2(\Gamma) \longrightarrow L^2(\Gamma) \\
\varphi \mapsto (\operatorname{Id} - M)\varphi
\end{array} \right.
\]
where $M$ is the Markov operator defined in Lemma 3.2.13, i.e.
\[
(3.30) \quad (M\varphi)(x) = \frac{1}{\operatorname{val}(x)} \sum_{y \in V} a(x,y)\varphi(y).
\]

If $\Gamma$ is $k$-regular, the Laplace operator of $\Gamma$ is defined by
\[
\Delta_\Gamma = k \Delta_\Gamma,
\]
and its spectral gap $\lambda_1(\Gamma)$ is its smallest non-zero eigenvalue, which is equal to $k \lambda_1(\Gamma_1)$.

**Remark 3.3.2.** In many sources (e.g., [31, §4.2]), a Laplace operator is defined for an arbitrary finite graph by
\[
(3.31) \quad \Delta_\Gamma \varphi(x) = \operatorname{val}(\cdot) \circ (\operatorname{Id} - M) \varphi(x) - \sum_{y \in V} a(x,y)\varphi(y),
\]
i.e., by $\Delta_\Gamma = \operatorname{val}(\cdot) \circ (\operatorname{Id} - M)$ where $\operatorname{val}(\cdot)$ represents the operator of multiplication by the valency function.

However, when the valency is not constant, passing from the spectrum of $\operatorname{Id} - M$ to that of $\Delta_\Gamma$ is not automatic. For instance, it is not easy to translate the formula (3.14) for the combinatorial Laplace operator (see Exercise 3.3.5 for a similar formula). But one can still prove that the smallest non-zero eigenvalue of $\Delta$ satisfies inequalities similar to that of $\operatorname{Id} - M$, taking the renormalization into account (see [31, Prop. 4.2.4, 4.2.5]). For general graphs, we will use the normalized Laplace operator in this book.

Here is a simple random example. For the graph in the figure below the matrices...
representing $\Delta_\Gamma$ and $\Delta_\Gamma$ in the basis of characteristic functions of single points are given, respectively, by

$$
\begin{pmatrix}
4 & -2 & 0 & -1 & -1 \\
-2 & 3 & -1 & 0 & 0 \\
0 & -1 & -2 & -1 & 0 \\
-1 & 0 & -1 & 3 & -1 \\
-1 & 0 & 0 & -1 & 2
\end{pmatrix},
$$

$$
\begin{pmatrix}
1 & -1/2 & 0 & -1/4 & -1/4 \\
-2/3 & 1 & -1/3 & 0 & 0 \\
0 & -1/2 & 1 & -1/2 & 0 \\
-1/4 & 0 & -1/4 & 3/4 & -1/4 \\
-1/3 & 0 & 0 & -1/3 & 2/3
\end{pmatrix}
$$

Their approximate eigenvalues, in increasing order, are

$$
0, 1.68373, 2.42058, 3.86537, 6.03032
$$

$$
0, 0.608079, 0.793860, 1.21495, 1.79978
$$

(respectively), and the ratios of the (non-zero) eigenvalues of $\Delta_\Gamma$ to the corresponding one of $\Delta_\Gamma$ are

$$
2.76894, 3.04912, 3.18150, 3.35059,
$$

approximately. Incidentally, one can see also that

$$
h(\Gamma) = 1
$$

(the optimal set is $W = \{1, 4, 5\}$ with $|E(W)| = 3$), and that the encaement provided by the normalized Laplace operator is

$$
0.304040 = \frac{2^2}{2 \cdot 4} \lambda_1(\Gamma) \leq 1 = h(\Gamma) \leq 4\sqrt{2} \lambda_1(\Gamma) = 4.41118.
$$

Thus, in that case, the lower bound is closer to the truth.

Here is a summary of the results of the previous section in terms of the combinatorial Laplace operator, for regular graphs.

**Proposition 3.3.3 (Properties of $\Delta_\Gamma$).** Let $\Gamma = (V, E, \text{ep})$ be a finite connected $k$-regular graph.

(1) The Laplace operator is self-adjoint and non-negative; its kernel is one-dimensional and spanned by the constant functions. Moreover

$$
\langle \Delta_\Gamma \varphi, \varphi \rangle = \frac{1}{2|V|} \sum_{x,y \in V} a(x, y)|\varphi(x) - \varphi(y)|^2
$$

for all $\varphi \in L^2(\Gamma)$.
(2) We have

\[ \lambda_1(\Gamma) = \min_{\varphi \in L^2(\Gamma)} \frac{\langle \Delta_\Gamma \varphi, \varphi \rangle}{\langle \varphi, \varphi \rangle} \]

and

\[ \frac{\lambda_1(\Gamma)}{2} \leq h(\Gamma) \leq \sqrt{2k\lambda_1(\Gamma)}. \]

These are immediate consequences of the previous discussion. Similarly, we state for completeness the characterization of expander graphs in terms of \( \lambda_1(\Gamma) \) and \( \lambda_1(\Gamma) \).

Theorem 3.3.4 (Spectral definition of expanders). Let \((\Gamma_i)_{i \in I}\) be a family of finite graphs with \(|\Gamma_i| \to +\infty\) and bounded valency \(\max_i \max_x \text{val}(x) \leq v\). Then \((\Gamma_i)\) is an expander family if and only if there exists \(\lambda > 0\) such that

\[ \lambda_1(\Gamma_i) \geq \lambda > 0 \]

for all \(i \in I\).

If each \(\Gamma_i\) is \(k\)-regular for a fixed \(k \geq 3\), then \((\Gamma_i)\) is an expander if and only if there exists \(\lambda' > 0\) such that

\[ \lambda_1(\Gamma_i) \geq \lambda' > 0 \]

for all \(i \in I\).

We call \((\lambda, v)\), or \((\lambda', k)\), the spectral expansion parameters of the expanders. For \(k\)-regular graphs, one can of course take \(\lambda' = k\lambda\).

Exercise 3.3.5 (The general Laplace operator). Let \(\Gamma = (V,E,\text{ep})\) be a finite connected graph. 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

\[ \langle f_1, f_2 \rangle_E = \frac{1}{2N} \sum_{\alpha \in E} |\text{ep}(\alpha)| f_1(\alpha) \overline{f_2(\alpha)}. \]

An orientation of \(\Gamma\) is the data of two maps

\[ b, e : E \to V \]

such that \(\text{ep}(\alpha) = \{b(\alpha), e(\alpha)\}\) for all \(\alpha \in E\) (in other words, if \(\alpha\) has two extremities, a choice of a “beginning” \(b(\alpha)\) and an “end” \(e(\alpha)\)). Given such an orientation, one can define a linear map

\[ d : \left\{ \begin{array}{l} L^2(\Gamma) \\ \varphi \end{array} \right\} \mapsto L^2(E) \]

where

\[ d\varphi(\alpha) = \varphi(b(\alpha)) - \varphi(e(\alpha)). \]

(1) Show that for any \(\varphi_1, \varphi_2 \in L^2(\Gamma)\), we have

\[ \langle \Delta_\Gamma \varphi_1, \varphi_2 \rangle = \langle d\varphi_1, d\varphi_2 \rangle_E, \]

where \(\Delta_\Gamma\) is given by (3.31).

(2) Deduce that the smallest eigenvalue \(\lambda_1(\Gamma)\) of the Laplace operator satisfies

\[ \lambda_1(\Gamma) \leq 2h(\Gamma). \]
The definition of expanders using the Laplace operator is qualitatively equivalent to that based on the expansion ratio, 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 \((\lambda, v)\) 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 4.3. 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 \(\lambda_1(\Gamma)\) might differ quite significantly.

To give an example: suppose one requires the spectral gap for a sequence of \(k'\)-regular graphs \((\Gamma'_n)\), which is obtained by “perturbation”, as in Corollary 3.1.18, of a family of \(k\)-regular graphs \((\Gamma_n)\). If one knows a lower-bound for the spectral gap of \((\Gamma_n)\), we already know how to deduce one for \(\Gamma'_n\), namely
\[
\lambda_1(\Gamma'_n) \geq \frac{h(\Gamma'_n)^2}{2k' \geq c \frac{h(\Gamma_n)^2}{2k'} \geq \lambda_1(\Gamma_n)^2,
\]
where \(c > 0\) is given by Corollary 3.1.18. If the spectral gap \(\lambda_1(\Gamma_n)\) 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 ratio.

For the spectral gap of the normalized Laplace operator, there are a number of results and techniques towards this goal, as explained in \[30, Ch. 13\]. We will apply the following bound (which corresponds to \[30, Th. 13.23\] and is due to Diaconis and Saloff-Coste) in the special case of Cayley graphs, where it becomes a bit simpler, but the general case is also instructive.

**Proposition 3.3.6 (Perturbing the Laplace operator).** Let \(\Gamma = (V, E, \text{ep})\) be a finite non-empty connected graph, and let \(\Gamma' = (V, E', \text{ep}')\) 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 \(\Gamma'\), of length \(\ell(x, y) \geq 1\), from \(x\) to \(y\). For each pair of distinct vertices \((s, t)\) \(\in V \times V\), let then \(A_{s,t}\) 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 \(i, 0 \leq i < \ell(x, y)\), with
\[
\varphi(\gamma_{x,y}(i)) = s, \quad \varphi(\gamma_{x,y}(i + 1)) = t.
\]

We then have
\[
\lambda_1(\Gamma') \geq \frac{1}{c_1 c_2} \lambda_1(\Gamma)
\]
where
\[
c_1 = \max_{x \in V} \frac{\mu_{\Gamma'}(x)}{\mu_{\Gamma}(x)} = \max_{x \in V} \frac{N \text{val}_{\Gamma'}(x)}{N' \text{val}_\Gamma(x)},
\]
\[
c_2 = \frac{N'}{N} \max_{(s, t) \in V \times V} \frac{1}{a'(s, t)} \sum_{(x, y) \in A_{s,t}} \ell(x, y) a(x, y).
\]
Example 3.3.7. 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

\[(3.34) \quad \langle (\text{Id} - M) \varphi, \varphi \rangle_{\Gamma} \leq c_2 \langle (\text{Id} - M') \varphi, \varphi \rangle_{\Gamma'},\]

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, on the other hand, we have

\[\| \varphi \|_{\Gamma}^2 \geq c_1^{-1} \| \varphi \|_{\Gamma'}^2,\]

we see using (3.25) that such a bound immediately gives the stated inequality.

To prove (3.34), we begin again with (3.14):

\[\langle (\text{Id} - M) \varphi, \varphi \rangle_{\Gamma} = \frac{1}{2N} \sum_{x,y \in V} a(x,y) |\varphi(x) - \varphi(y)|^2.\]

The non-zero terms are those corresponding to adjacent vertices in $\Gamma$. To introduce the edges of $\Gamma'$ 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 Cauchy-Schwarz, we get

\[|\varphi(x) - \varphi(y)|^2 \leq \ell(x,y) \sum_{i=0}^{\ell(x,y)-1} |\varphi(\gamma_{x,y}(i+1)) - \varphi(\gamma_{x,y}(i))|^2,\]

where the successive differences are between points which are, by definition, adjacent in $\Gamma'$. And from then on, we basically just gather things up as they flow: we write

\[\sum_{x,y \in V} a(x,y) |\varphi(x) - \varphi(y)|^2 \leq \sum_{x,y \in V} a(x,y) \ell(x,y) \sum_{i=0}^{\ell(x,y)-1} |\varphi(\gamma_{x,y}(i+1)) - \varphi(\gamma_{x,y}(i))|^2 = \sum_{s,t \in V} a'(s,t) \beta(s,t) |\varphi(t) - \varphi(s)|^2\]

with $\beta(s,t) = 0$ unless $a'(s,t) = 0$ and otherwise, by definition, we have

\[\beta(s,t) = \frac{1}{a'(s,t)} \sum_{(x,y) \in A_{s,t}} a(x,y) \ell(x,y).\]

This leads to

\[\langle (\text{Id} - M) \varphi, \varphi \rangle_{\Gamma} \leq \frac{N'}{N} (\max_{s,t} \beta(s,t)) \langle (\text{Id} - M') \varphi, \varphi \rangle_{\Gamma'},\]

which is the same as (3.34). \qed
3.4. 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:

- If $G$ is a finite group, $S \subset G$ a symmetric subset of generators of $G$, and $\Gamma = \mathcal{C}(G, S)$ is the corresponding Cayley graph, we have

$$h(\Gamma) = \min_{\emptyset \neq W \subset G, |W| \leq |G|/2} \frac{|E(W)|}{|W|}$$

with

$$|E(W)| = \left| \{(g, s) \in W \times S \mid gs \notin W\} \right|$$

(the two sets are in bijection, from right to left, by $(g, s) \mapsto \{g, gs\} \in E_{\Gamma}$).

- The space $L^2(\Gamma)$ is identical with the space $L^2(G)$ of complex-valued functions on $G$, with the uniform inner-product

$$\langle \varphi_1, \varphi_2 \rangle = \frac{1}{|G|} \sum_{g \in G} \varphi_1(g) \overline{\varphi_2(g)}$$

and 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

$$\Delta \varphi(g) = |S| \varphi(g) - \sum_{s \in S} \varphi(gs),$$

$$\langle \Delta \varphi, \varphi \rangle = \frac{1}{2|G|} \sum_{g \in G, s \in S} |\varphi(gs) - \varphi(g)|^2$$

for all $\varphi \in L^2(G)$ and, as usual, the minimization formula

$$\lambda_1(\Gamma) = |S| \lambda_1(\Gamma) = \min_{\varphi \perp 1} \langle \Delta \varphi, \varphi \rangle \|\varphi\|^2.$$

The most important feature distinguishing Cayley graphs from “general” regular graphs, is their symmetry (recall that $G$ acts by graph automorphisms on $\Gamma$, see Proposition 2.3.8). In particular, “every vertex looks the same”. This has important consequences when applying Proposition 3.3.6 for Cayley graphs. Indeed, we obtain the following version of this result:

**Proposition 3.4.1 (Perturbing Cayley graphs).** Let $G$ be a finite group, $S, S' \subset G$ two 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

$$\lambda_1(\Gamma') \geq c^{-1} \lambda_1(\Gamma)$$

where

$$c = |S'| \max_{s \in S} \ell_{S'}(s)^2.$$
Proof. We apply Proposition 3.3.6 to the two Cayley graphs. The quantity $c_1$ is then equal to 1, and for any $x, y \in V$, we take a path $\gamma_{x,y}$ in $\Gamma'$ obtained by concatenating to $x$ a path representing an expression

$$x^{-1}y = w_1 \cdots w_m, \quad w_i \in S',$$

as a word in the generators from $S'$ (this exploits the homogeneity). 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 $\Gamma'$ if $h = gs'$ for some $s' \in S'$. Similarly, $x, y$ are joined by an edge of $\Gamma$ (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 $A_{g,h}$ if the edge joining $g$ to $h$ appears in the path $\gamma_{x,y}$. In terms of the decomposition (3.35) of $x^{-1}y = s$, this happens exactly when $w_i = s'$ for some $i$, with

$$x = g(w_1 \cdots w_{i-1})^{-1}, \quad y = h(w_{i+1} \cdots w_m).$$

For each $s \in S$ we get therefore as many elements in $A_{g,h}$ as there are occurrences of $s' = g^{-1}h$ in the $S'$-decomposition of $s$, say $\ell_{g^{-1}h}(s) \geq 0$ times, and we obtain

$$\sum_{(x,y) \in A_{g,h}} \ell(x,y)a(x,y) \leq \sum_{s \in S} \ell_{S'}(s)\ell_{g^{-1}h}(s) \leq |S|\ell_{S'}(s)^2,$$

for all $g$ and $h$. Referring to (3.33), this is precisely the desired formula since

$$\frac{N'}{N} = \frac{|S'|}{|S|}.$$

□

Remark 3.4.2. 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 enough gives a fairly good control of the spectral gap, and hence of the expansion constant!

Corollary 3.4.3 (Bounding the spectral gap from the diameter). Let $G$ be a finite group, $S \subset G$ a finite symmetric generating set of $G$, and denote $\Gamma = \mathcal{C}(G,S)$. We have then

$$\lambda_1(\Gamma) \geq \frac{1}{\text{diam}(\Gamma)^2},$$

and hence

$$h(\Gamma) \geq \frac{1}{2\text{diam}(\Gamma)^2}.$$
Proof. It is equivalent to show that

\[ \lambda_1(\Gamma) \geq \frac{1}{|S| \text{diam}(\Gamma)^2}, \]

and we do it by comparing \( \Gamma \) to the “perturbed” Laplace operator of \( \tilde{\Gamma} = \mathcal{C}(G, T) \) with \( T = G \). More precisely we take \( (\tilde{\Gamma}, \Gamma) \) for \( (\Gamma, \Gamma') \) in Proposition 3.4.1 (with apologies for the possible confusion); we have \( \ell_s(g) \leq \text{diam}(\Gamma) \) for all generators \( g \in T \), and therefore we get

\[ \lambda_1(\Gamma) \geq c^{-1} \lambda_1(\tilde{\Gamma}) \]

with \( c = |S|(\text{diam}(\Gamma))^{\frac{3}{2}} \). But \( \lambda_1(\tilde{\Gamma}) = 1 \), which finishes the proof! Indeed, \( \tilde{\Gamma} \) is a complete graph with vertex set \( G \), with an added loop at each vertex. Therefore, if we start from \( X_0 = 1 \), a random walk on \( \tilde{\Gamma} \) is exactly uniformly distributed at each step \( X_k, k \geq 1 \), and this means that \( M \) has no non-zero eigenvalue different from 1.

Analytically, we see this by noting that we have

\[ \tilde{M}\varphi(x) = \frac{1}{|G|} \sum_{g \in G} \varphi(xg) = \frac{1}{|G|} \sum_{g \in G} \varphi(g), \]

i.e., \( \tilde{M} \) is the orthogonal projection on the constants, and therefore its only eigenvalues are 0 and 1. □

Example 3.4.4. (1) The example of a finite \( d \)-regular tree \( T_{d,k} \), with \( d \geq 3 \), shows how far from being general this fact is: in that case, for fixed \( d \), we have

\[ h(T_{d,k}) \leq \frac{d}{|T_{d,k}| - 1} \ll \frac{1}{|T_{d,k}|}, \]

for \( k \geq 1 \) (by Example 3.1.3, (4); the implied constant depends on \( d \)), hence also

\[ \lambda_1(T_{d,k}) \ll |T_{d,k}|^{-1}, \]

while the diameter only grows logarithmically as a function of \( |T_{d,k}| \). The bounds of the corollary therefore fail dramatically in this (non-homogeneous) case. The reader is invited to analyze what happens when one tries to imitate for \( T_{d,k} \) the comparison in the proof of Corollary 3.4.3 using Proposition 3.3.6.

(2) On the other hand, for Cayley graphs, the result is essentially sharp. Consider indeed the cycles \( C_m = \mathcal{C}(\mathbb{Z}/m\mathbb{Z}, \{\pm 1\}) \), for which we have \( \text{diam}(C_m) \sim m/2 \) and

\[ \lambda_1(\Gamma_m) \sim \frac{4\pi^2}{m^2} \sim \frac{\pi^2}{(\text{diam } C_m)^2} \]

as \( m \to +\infty \) by Example 3.2.21 (taking into account that \( C_m \) is 2-regular.)

(3) An immediate consequence of Corollary 3.4.3, 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

\[ \lambda_1(\mathcal{C}(G, S)) \geq \frac{1}{|S||G|^2} \geq \frac{1}{|G|^3}. \]

This is slightly better than the general lower bound (3.29). This remark will be useful in the next chapter, 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.

68
Using Corollary 3.4.3, 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 3.4.5 (Esperantist graphs).** A family \((\Gamma_i)_{i \in I}\) of finite non-empty connected graphs \(\Gamma_i = (V_i, E_i, \epsilon_i)\) is an esperantist family if there exist constants \(v \geq 1, c > 0, A \geq 0\), independent of \(i\), such that the number of vertices \(|V_i|\) tends to infinity, the maximal valency of \(\Gamma_i\) is at most \(v\) for all \(i\), and the expansion constant satisfies
\[
h(\Gamma_i) \geq \frac{c}{(\log 2|\Gamma_i|)^A}.
\]

We will say that a triple \((c, A, v)\) for which the two properties above hold are esperanto parameters of the family.

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. We will discuss this briefly in Section 4.4, following [13]. As we will already be able to show in Example 3.4.7, there are quite simple examples of interesting sequences of Cayley graphs which are esperantist but not expanders.

From Corollary 3.4.3, 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 3.4.6 (Forms of esperantism).** A family \((\Gamma_i)_{i \in I}\) of finite non-empty Cayley graphs \(\Gamma_i = C(G_i, S_i)\), with \(|G_i|\) tending to infinity and \(|S_i| \leq v\) for all \(i\), is an esperantist family if and only if one of the two properties below hold:

1. For some \(c > 0\) and \(A \geq 1\), we have
   \[
diam(\Gamma_i) \leq C(\log 2|\Gamma_i|)^A,
   \]
   and one may take \((\frac{1}{2C^2}, 2A, v)\) as esperanto parameters.

2. For some \(c' > 0\) and \(A' \geq 0\), we have
   \[
   \lambda_1(\Gamma_i) \geq c'(\log 2|\Gamma_i|)^{-A'},
   \]
   and one may take \((\frac{c'}{2}, A, v)\) as esperanto parameters.

**Proof.** One direction of the equivalence of esperantism with (1) follows from Corollary 3.4.3, namely
\[
h(\Gamma_i) \geq \frac{1}{2 \diam(\Gamma_i)^2} \geq \frac{1}{2C^2 (\log 2|\Gamma_i|)^2A}.
\]
Conversely, by Proposition 3.1.5, we have
\[
diam(\Gamma_i) \leq 2 \frac{\log |\Gamma_i|}{\log \left(1 + \frac{h(\Gamma_i)}{v}\right)} + 3,
\]
and for an esperantist family we can apply
\[
\log(1 + x) \geq \min\left(\frac{x}{2}, \log(2)\right)
\]
for \(x \geq 0\) to obtain
\[
\log \left(1 + \frac{h(\Gamma_i)}{v}\right) \geq \min\left(\log 2, \frac{1}{2v} h(\Gamma_i)\right) \geq \min\left(\log 2, \frac{1}{2v} \left(\frac{c}{(\log 2|\Gamma_i|)^A}\right)\right),
\]
and hence
\[
diam(\Gamma_i) \ll (|\Gamma_i|)^{A+1},
\]
69
which gives the “polylogarithmic growth” of the diameter.

As for (2), the equivalence follows immediately from (3.32).

Example 3.4.7 (Symmetric groups as an esperantist family). Using these results, we can already exhibit an explicit esperantist family. Let $G_n = C(S_n, S_n)$ for $n \geq 3$ as in Example 2.3.2, where we see again $\mathfrak{S}_n$ as acting on $\mathbb{Z}/n\mathbb{Z}$. We already know that $\text{diam}(G_n) \ll n^2$ (by Exercise 2.3.5), and therefore we derive

$$
\lambda_1(G_n) \gg \frac{1}{n^4} \gg \frac{1}{(\log |G_n|)^4}
$$

by Corollary 3.4.3, 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 3.4.8.** For the graphs $(G_n)$ above, we have

$$
h(C(S_n, S_n)) \ll \frac{1}{n^2}, \quad \lambda_1(C(S_n, S_n)) \asymp \frac{1}{n^3},
$$

for $n \geq 3$.

**Proof.** The random walk on these Cayley graphs is analyzed by Diaconis and Saloff-Coste in [12, §3, Ex. 1]. We will first show that

$$
\lambda_1(G_n) \ll n^{-3},
$$

while the corresponding lower-bound $\lambda_1(G_n) \gg n^{-3}$ is proved in [12], and we defer it to an exercise below. We then use the argument underlying the proof of Cheeger’s inequality to show that the bound $n^{-3/2}$ that follows directly from Proposition 3.2.33 can be improved to

$$
h(G_n) \ll n^{-2}.
$$

To get an upper bound for $\lambda_1(G_n)$, we use a specific ad-hoc test function $\varphi$ in the characterization (3.24). The goal is to have $\varphi$ be “almost” invariant under multiplication by $\tau$ 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 $\mathbb{Z}/n\mathbb{Z}$. 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 $\mathbb{Z}/n\mathbb{Z}$) may seem to have a good chance. Indeed, we have $\varphi(\sigma\sigma_n) = \varphi(\sigma\sigma_n^{-1}) = \varphi(\sigma)$ for all $x \in \mathbb{Z}/n\mathbb{Z}$, and therefore

$$
\langle (\text{Id} - M)\varphi, \varphi \rangle = \frac{1}{6|G_n|} \sum_{\sigma \in \mathfrak{S}_n} (\varphi(\sigma\tau) - \varphi(\sigma))^2.
$$

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 $\sigma$ such that

$$
\sigma^{-1}(1) \in \{1, 2\} \text{ or } \sigma^{-1}(2) \in \{1, 2\},
$$

and hence we get

$$
(3.37) \quad \langle (\text{Id} - M)\varphi, \varphi \rangle \leq \frac{2}{3n}.
$$

On the other hand, we have

$$
\|\varphi - \langle \varphi, 1 \rangle\|^2 = \frac{1}{|\mathfrak{S}_n|} \sum_{\sigma \in \mathfrak{S}_n} \varphi(\sigma)^2 - \left( \frac{1}{|\mathfrak{S}_n|} \sum_{\sigma \in \mathfrak{S}_n} \varphi(\sigma) \right)^2,
$$

70
which we evaluate by using the distribution of value of $\varphi$. 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

\[
\frac{1}{|\mathcal{S}_n|} \sum_{\sigma \in \mathcal{S}_n} \varphi(\sigma)^2 = \frac{1}{|\mathcal{S}_n|} \sum_{1 \leq j \leq n/2} j^2 |\{ \sigma \mid \varphi(\sigma) = j\}| = \frac{2}{n} \sum_{1 \leq j \leq n/2} j^2 \sim \frac{n^2}{12}
\]

for $n \to +\infty$. Similarly

\[
\frac{1}{|\mathcal{S}_n|} \sum_{\sigma \in \mathcal{S}_n} \varphi(\sigma) = \frac{2}{n} \sum_{1 \leq j \leq n/2} j \sim \frac{n}{4},
\]

and hence

\[
\|\varphi - \langle \varphi, 1 \rangle\|_2^2 \sim \frac{n^2}{48}
\]

in that case. 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

\[
\lambda_1(\Gamma_1) \leq \frac{\| (\text{Id} - M) \varphi, \varphi \|}{\| \varphi - \langle \varphi, 1 \rangle \|_2^2} \leq \frac{32}{n^3} + o(1)
\]

as $n \to +\infty$.

We now apply (3.28) to estimate $h(G_n)$ (in other words, Lemma 3.2.34 with the uniform probability measure on $[1, \lfloor n/2 \rfloor]$, 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

\[
h(G_n) \leq \frac{A}{B}
\]

with

\[
A = \frac{1}{2} \sum_{\sigma \in \mathcal{S}_n} |\varphi(\sigma) - \varphi(\sigma \tau)|, \quad B = \sum_{\sigma \in \mathcal{S}_n} |\varphi(\sigma) - t_0|.
\]

As we have seen, $|\varphi(\sigma) - \varphi(\sigma \tau)|$ is either 0 or 1, and therefore we get

\[
A \leq \frac{2|\mathcal{S}_n|}{3n},
\]

just as in (3.37). Proceeding for $B$ by summing according to the values of $\varphi$, in the manner used for the norm $\|\varphi - \langle \varphi, 1 \rangle\|_2^2$, we get

\[
B \sim \frac{n|\mathcal{S}_n|}{8},
\]

for $n \to +\infty$, and therefore

\[
h(G_n) \leq \frac{A}{B} \ll \frac{1}{n^2},
\]

as claimed. 

\[\square\]

**Exercise 3.4.9** (A comparison of Cayley graphs). Let $\tilde{G}_n = \mathcal{C}(\mathcal{S}_n, T_n)$ where $T_n$ is the set of all transpositions in $\mathcal{S}_n$.

1. Show that

\[
\lambda_1(\tilde{G}_n) = \frac{2}{n}.
\]

2. Deduce by comparison that

\[
\lambda_1(G_n) \gg n^{-3}.
\]
For the remainder of this book, the most important families of graphs will indeed arise from Cayley 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 \((\mathcal{C}(G_i, S_i))\). 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 \(K_i \trianglelefteq G\) with finite index \([G : K_i] \to +\infty\), and we consider the action graphs \(A(G/K_i, S)\).

Note that when the quotient maps \(G \to G_i = G/K_i\) are injective on \(S_i\), with image \(S_i\), 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 we could restrict without much loss to the first case.

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?

For the moment, we have only considered two examples of sequences of Cayley graphs: if we take either the cycles \(C_m\), \(m \geq 2\), which are 2-regular Cayley graphs of \(G_m = \mathbb{Z}/m\mathbb{Z}\), or the graphs \(G_n = \mathcal{C}(\mathbb{F}_n, S_n)\), we know that these are not expanders. But it turns out that, for many interesting sequences of “complicated” groups, the answer is positive, or conjectured to be so.

In the next chapter, we will prove the following case, which is a recent theorem of Bourgain and Gamburd [4]:

**Theorem 3.4.10 (Expansion in subgroups of \(SL_2(\mathbb{Z})\)).** Let \(S \subset SL_2(\mathbb{Z})\) be any finite symmetric subset and let \(G\) be the subgroup generated by \(S\). For prime numbers \(p\), let \(\Gamma_p = \mathcal{C}(SL_2(\mathbb{F}_p), S)\) be the Cayley action graph of finite quotient group \(SL_2(\mathbb{F}_p)\) with respect to the reduction modulo \(p\) of the set \(S\). Then \((\Gamma_p)_{p \geq p_0}\) is an expander family if and only if \(\Gamma_p\) is connected for all \(p \geq p_0\).

As we will also explain, a crucial step in the proof is a theorem of Helfgott, which was the first breakthrough towards expansion properties of such groups in this generality. We will state it precisely when needed, but it is helpful to state right now the following result, which turns out to be an immediate corollary.

**Theorem 3.4.11 (Esperantism for Cayley graphs of \(SL_2(\mathbb{F}_p)\)).** For any prime number \(p\), let \(S_p \subset SL_2(\mathbb{F}_p)\) be a symmetric generating set of \(SL_2(\mathbb{F}_p)\), and assume that
\[
|S_p| \leq k
\]
for some fixed \(k \geq 1\). Then the family \((\mathcal{C}(SL_2(\mathbb{F}_p), S_p))\) of Cayley graphs is an esperantist family, i.e., there exists \(c > 0\) and \(A \geq 0\) such that
\[
\lambda_1(\mathcal{C}(SL_2(\mathbb{F}_p), S_p)) \geq \frac{c}{(\log p)^A}.
\]

The condition that \(\Gamma_p\) be connected for all sufficiently large primes \(p\) that we use in Theorem 3.4.10 is obviously necessary for the theorem to hold. It is equivalent to another condition which can be checked very easily, and which holds (for instance) whenever \(S\) is chosen “randomly” (if \(|S| \geq 3\)). Even without such a criterion, we can certainly apply the theorem to concrete subsets of \(SL_2(\mathbb{Z})\) for which we understand the reduction modulo primes \(p\). The simplest case is probably given by the following corollary:

**Corollary 3.4.12.** Let \(k \geq 1\) be an integer, and let
\[
S = \left\{ \begin{pmatrix} 1 & \pm k \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ \pm k & 1 \end{pmatrix} \right\} \subset SL_2(\mathbb{Z}),
\]
and for \( p \) prime, let \( S_p \) denote the image of \( S \) modulo \( p \). Then the family of Cayley graphs \( \mathcal{C}(\text{SL}_2(\mathbb{F}_p), S_p) \) for \( p \nmid k \) is an expander family.

For \( k = 1 \) or \( k = 2 \), this result was known since the early 1980’s for by combining a spectral gap result of Selberg in hyperbolic geometry with a comparison principle of either Burger or Brooks (who proved two versions of this independently). However, for \( k \geq 3 \), this was a notorious open question until the results of Hefgott and Bourgain-Gamburd led to a general proof.

The restriction to subgroups of \( \text{SL}_2(\mathbb{Z}) \) and to reduction modulo primes in Theorem 3.4.10, and to subgroups of \( \text{SL}_2(\mathbb{F}_p) \) for Theorem 3.4.11, was necessary when these results were first proved. In the meantime, however, there has been immense progress in generalizing the methods involved, and hence extending the results. As a consequence, much stronger results of the same type are now known, the latest generalization of Theorem 3.4.10 being due to Salehi-Golsefidy and Varjú [41], while a remarkably general version of Theorem 3.4.11 was found by Pyber and Szabó [40]. We will briefly survey these generalizations in Section 5.7.

These developments go in the direction of showing that many families of “sufficiently complicated” groups form expanders, and are often strikingly (almost) just properties of the group structure, and do not seem to involve the details of the generating sets which are used. But this is not always the case, as the example of symmetric groups show.

We have already seen the Cayley graphs \( G_n = \mathcal{C}(S_n, S_n) \) which are not expanders, but on the other hand, in a remarkable work, Kassabov [25] succeeded in finding (effectively computable) generating sets \( T_n \) of \( S_n \), with bounded size, such that the Cayley graphs \( \mathcal{C}(S_n, T_n) \) do form an expander. Hence, for symmetric groups at least, the expansion property is not purely group-theoretical.

3.5. Existence of expanders

In this 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 [39, Lemma 1]. It turns out, in fact, that for many models of random graphs, there is large probability that they are expanders, in the sense that there is a positive lower bound for the Cheeger constant, valid with high probability.

Even more is true: the proof of this property turns out to be straightforward, in the sense that all estimates are done in the most trivial manner!

We will use a standard model to prove the result. To begin with, we construct bipartite expanders. Fix some integer \( k \geq 3 \). For any fixed \( n \geq 1 \) and any \( k \)-tuple \( \sigma = (\sigma_1, \ldots, \sigma_k) \) of permutations of \( \{1, \ldots, n\} \), we define a graph \( \Gamma_\sigma \) with vertex set

\[
V = \{(i, 0) \mid 1 \leq i \leq n \} \cup \{(i, 1) \mid 1 \leq i \leq n \} = V_0 \cup V_1,
\]

(independent of \( \sigma \)) and with edges joining \((i, 0)\) to \((\sigma_j(i), 1)\) for \(1 \leq j \leq k\): formally, we take

\[
E_\sigma = \{(i, \sigma_j(i)) \mid 1 \leq i \leq n, \ 1 \leq j \leq k\},
\]

and \( \text{ep}(i, \sigma_j(i)) = \{(i, 0), (\sigma_j(i), 1)\} \). These graphs are bipartite and \( k \)-regular, and they may have multiple edges.

We view these graphs as random graphs by thinking of the permutations \( \sigma_i \) as taken independently and uniformly at random in \( \mathfrak{S}_n \). Thus the probability that the graphs \( \Gamma_\sigma \) satisfy a property \( \mathcal{P}(\Gamma) \) of graphs, denoted \( \mathbf{P}(\Gamma_\sigma \text{ has } \mathcal{P}) \), is simply

\[
\mathbf{P}(\Gamma_\sigma \text{ has } \mathcal{P}) = \frac{1}{|\mathfrak{S}_n|^k} |\{\sigma \in \mathfrak{S}_n^k \mid \Gamma_\sigma \text{ has } \mathcal{P}\}|.
\]
Note that this depends on \( n \) and \( k \), and we will be interested in the limit when \( n \) goes to infinity for a fixed \( k \). Indeed, we have:

**Theorem 3.5.1.** Fix \( k \geq 3 \). There exists \( h_k > 0 \) such that
\[
\lim_{n \to +\infty} P(h(\Gamma_\sigma) < h_k) = 0.
\]

In particular, for \( n \) large enough, some \( \Gamma_\sigma \) satisfies \( h(\Gamma_\sigma) \geq h_k \).

**Remark 3.5.2.** 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 \( T_{d,k} \) with \( d \geq 3 \) fixed and \( k \to +\infty \), and attempt to add some edges connecting the leaves of the true to vertices “inside”, 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 put in randomly, to avoid dealing with two types of edges. This might naturally lead to the graphs of the type we consider here...

**Proof.** We use Lemma 3.1.16 (which is applicable to all graphs \( \Gamma_\sigma \)) to observe that
\[
P(h(\Gamma_\sigma) < h_k) \leq P(\hat{h}(\Gamma_\sigma) < 1 + 2h_k).
\]

A simple symmetry argument (left to the reader) also shows that, for any \( \delta > 0 \), we have
\[
P(\hat{h}(\Gamma_\sigma) < 1 + \delta) = P\left( \min_{1 \leq |W| \leq |V_0|/2} \frac{\partial_\sigma W}{|W|} < 1 + \delta \right)
\]
where we indicate with the subscript \( \sigma \) that the boundary is, of course, taken in the sense of \( \Gamma_\sigma \) (compare with (3.5: 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 \( \sigma \):
\[
p_n \leq \sum_{W \subset V, 1 \leq |W| \leq n/2} P(|\partial_\sigma W| < (1 + \delta)|W|).
\]

Again for symmetry reasons (which the reader should check), the probability \( P(|\partial_\sigma W| < \delta|W|) \) only depends on \( |W| \), because any subset of size \( \ell \) in \( V_0 \) is, for this model of random graphs, equivalent to \( \{1, \ldots, \ell\} \). Hence
\[
p_n \leq \sum_{1 \leq \ell \leq n/2} \binom{n}{\ell} P(|\partial_\sigma\{1, \ldots, \ell\}| < (1 + \delta)\ell).
\]

Since there are edges joining \((0, j)\) to \((1, \sigma_1(j))\) for all \( j \), and \( \sigma_1 \) is a permutation, we always have
\[|\partial_\sigma\{1, \ldots, \ell\}| \geq \ell.\]

Hence, if \( |\partial_\sigma\{1, \ldots, \ell\}| < (1 + \delta)\ell \) for some \( \sigma \), it is necessary that at most \( \delta \ell \) of the values \( \sigma_2(j) \) for \( 1 \leq j \leq \ell \) be outside the set
\[
I_\ell = \{\sigma_1(1), \ldots, \sigma_1(\ell)\},
\]

74
and similarly for $\sigma_3, \ldots, \sigma_k$, since the occurrence of any of these events would imply $|\partial_\sigma\{1, \ldots, \ell\}| \geq (1 + \delta)\ell$. These events are independent, since $\sigma_2, \ldots, \sigma_k$ are independently chosen in the symmetric group. We only consider $\sigma_2$ and $\sigma_3$, and find that the probability that they both occur is at most

$$\sum_{|E| = |\delta\ell|} P(\sigma_2(\{1, \ldots, \ell\}) \subset I_\ell \cup E)^2,$$

since for each of the two independent events the probability is the same. Again because of symmetry, this quantity is at most

$$\left(\frac{n}{|\delta\ell|}\right) \frac{1}{(n!)^2} \left\{ (\ell + |\delta\ell|)(\ell + |\delta\ell| - 1) \cdots (|\delta\ell| + 1)(n - \ell)! \right\}^2 = \left(\frac{n}{|\delta\ell|}\right) \left(\frac{\ell + |\delta\ell|}{\ell}\right)^2 \left(\frac{n}{\ell}\right)^{-2},$$

where the first factor is the number of possibilities for $E$, and the quantity within the square expresses the fact that if $\sigma$ is a permutation with $\sigma(\{1, \ldots, \ell\}) \subset I_\ell \cup E$, we have $\ell + |\delta\ell|$ possibilities for $\sigma(1)$, one less for $\sigma(2), \ldots, |\delta\ell| + 1$ for $\sigma(\ell)$, and then finally $(n - \ell)!$ possibilities for the remaining values of $\sigma$.

We are thus left with the estimate

$$p_n \leq \sum_{1 \leq \ell \leq n/2} \left(\frac{n}{\ell}\right) \left(\frac{n}{|\delta\ell|}\right) \left(\frac{\ell + |\delta\ell|}{\ell}\right)^2 \left(\frac{n}{\ell}\right)^{-2}.$$

□
CHAPTER 4

Applications

4.1. 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 [1]. We describe a variant using only embeddings
of the graph of a combinatorial nature which avoids discussing a “thickening” of the
vertices and edges. The idea is however intuitively the same.

Given a finite graph $\Gamma$, which we assume to be without loops or multiple edges, we
define a thick embedding of $\Gamma$ in $\mathbb{R}^3$ to be a pair $(\iota, j)$ such that $\iota$ is a map
\[ \iota : V \to \mathbb{R}^3 \]
and $j$ is a map $E \to C^1([0,1], \mathbb{R}^3)$, where $C^1([0,1], \mathbb{R}^3)$ is the space of $C^1$-functions from
$[0,1]$ to $\mathbb{R}^3$, with the following properties:

1. the map $\iota$ is injective;
2. 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;
3. the balls of center $\iota(x)$ and radius 1 are disjoint;
4. for any distinct edges $\alpha$ and $\alpha'$, if there exists $t$ and $t'$ in $[0,1]$ such that
   the functions $\varphi = j(\alpha)$ and $\varphi' = j(\alpha')$ satisfy $|\varphi(t) - \varphi(t')| \leq 1/2$, then $\alpha$ and $\alpha'$
have a common extremity $x$, and $\varphi(t)$ and $\varphi(t')$ are at distance $\leq 1/2$ of $\iota(x)$.

Intuitively, we view this data as a way of embedding the graph in $\mathbb{R}^3$: 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 4.1.1.** Let $\Gamma$ be a finite graph with maximal degree at most 6 and no
loops or multiple edge. Then $\Gamma$ 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 $\mathbb{R}^3$. For each new edge $\alpha$, we must ensure that the “main
part” of the path $\varphi = j(\alpha)$ is contained in $\mathbb{R}^3$ minus the finite union of the points at
distance $\leq 1$ of the previous paths. This is certainly possible since this space is pathwise
connected.

Given a thick embedding $\iota$, we define its radius $r(\iota)$ to be the infimum of those real
numbers $r \geq 0$ such that, for some $x \in \mathbb{R}^3$, the image of $\iota$ is contained in the ball of
radius $r$ centered at $x$. In particular, the smallest volume of a ball in $\mathbb{R}^3$ in which $\Gamma$
can be “drawn” using a thick embedding. The theorem of Barzdin and Kolmogorov is concerned with how large \( r(\ell) \) should be.

**Theorem 4.1.2 (Barzdin–Kolmogorov).** Let \( \Gamma \) be a finite graph with degree at most 6 at each vertex.

1. Let \( \iota \) be a thick embedding of \( \Gamma \). There exists a constant \( c > 0 \), independent of \( \Gamma \) and \( \iota \), such that the radius of \( \iota \) is at least \( c \sqrt{h(\Gamma)}|V| \).

2. There exists a constant \( c > 0 \), independent of \( \Gamma \), such that \( \Gamma \) admits a thick embedding with radius \( \leq c \sqrt{h(\Gamma)}|V| \).

In particular, if \( (\Gamma_n)_{n \geq 1} \) is a family of expander graphs with degree at most 6, the optimal order of magnitude of the radius of a thick embedding of \( \Gamma_n \), 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), since (2) is outside the scope of this book. We may certainly assume that \( |V| \geq 3 \). We assume first that \( \iota \) is such that the \( z \)-coordinate of all points \( \iota(x) \), for \( x \in V \), are distinct. Let \( \iota \) be a thick embedding of \( \Gamma \) and let \( x_0 \in \mathbb{R}^3 \) 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 \( \Gamma \) are such that the third coordinate of \( \iota(x) \) is \( \leq z \), and at least are such that the third coordinate if \( \leq z \) (\( z_0 \) is a median of the third coordinate function on \( V \)).

Let \( V_1 \subset V \) be the set of those vertices with third coordinate \( \leq z \). We have \( \frac{1}{2}|V| - 1 \leq |V_1| \leq \frac{1}{2}|V| \). By definition of the Cheeger constants, there are at least \( h(V)|V_1| \geq \frac{1}{4}h(V)|V| \) edges in \( \Gamma \) with one extremity in \( V_1 \) and one extremity in \( V_2 \). For each such edge \( \alpha \), with extremities \( x_1 \) and \( x_2 \), the definition of a thick embedding implies that the path \( j(\alpha) \) joining \( \iota(x_1) \) and \( \iota(x_2) \) intersects the horizontal plane with equation \( z = z_0 \). Let \( f(\alpha) \) be such intersection point. All the points of the form \( f(\alpha) \) are also in a disc in the plane \( z = z_0 \) with radius \( \leq r \). On the other hand, by the last condition in the definition of a thick embedding, if \( \alpha \) and \( \alpha' \) are distinct edges, then the balls with radius \( 1/2 \) and centers \( f(\alpha) \) and \( f(\alpha') \) are disjoint, unless \( \alpha \) and \( \alpha' \) have a common extremity. Since the maximal degree is 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

\[
\frac{1}{96}h(V)|V| \leq r^2.
\]

Consider now the general case where no assumption on the \( z \)-coordinates of the points \( \iota(x) \) is made. We can then perturb slightly \( \iota \) to ensure that this condition holds, and one convinces oneself then that the paths \( j(\alpha) \) can also be changed slightly to obtain a new thick embedding to which the previous argument applies.

**4.2. Error reduction in probabilistic algorithms**

In this section, which is based on [21, 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

\[
r : I \rightarrow [0, +\infty[.
\]
is a (deterministic) computer program which takes as input an element \( i \in I \), and (always) ends its run by giving an output in \( O \), denoted \( A[i] \), the “time” necessary for the computation being \( r(i) \) (for instance, \( A \) can be defined as a Lisp function [33], 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 with the number of binary digits needed to encode \( i \), and \( r(i) \) will be a function of \( \ell(i) \) only. For instance, if \( I = \mathbb{Z} \), 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

\[
r(i) \leq c_1 \ell(i)^A
\]

for some constants \( c_1 \geq 0 \) and \( A \geq 0 \) which are independent of \( i \in I \).

**Remark 4.2.1.** In principle, the running time should include the time needed to “output” the value \( A[i] \). However, we will consider algorithms with \( O = \{0, 1\} \) for which such a distinction is irrelevant.

**Example 4.2.2.** Given a subset \( M \subset \mathbb{Z} \), one can ask for a fast algorithm \( A_M \) which “recognizes” \( M \), i.e., which has input set \( I = \mathbb{Z} \), output \( O = \{0, 1\} \), runs in polynomial time (relative to the number of digits of \( i \)) and is such that

\[
A_M[i] = \begin{cases} 
1 & \text{if } i \in M, \\
0 & \text{if } i \notin M.
\end{cases}
\]

A natural number-theoretic example is the set \( M \) of prime numbers. In that case, the naive “trial-division” algorithm certainly has the right output, but is *not* fast: its running time satisfies \( r(i) \leq \sqrt{i} \approx 2^{\ell(i)/2} \).

Attempting to come up (without cheating) with a polynomial time algorithm to recognize the set of primes should convince the reader that this is not a trivial problem. 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

\[
f : I \rightarrow \{0, 1\}
\]

where \( I \) is given with a size function \( \ell \) 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 \( \hat{A} \) with input set

\[
\hat{I} = \bigcup_{i \in I} (\{i\} \times \Omega_i)
\]

such that (1) the auxiliary non-empty sets \( \Omega_i \) (the sets of “random bits”) are also given with a size function \( \ell(\omega) \) such that \( \ell(\omega) \leq c_2 \ell(i)^B \), where \( c_2 \geq 0 \) and \( B \geq 0 \) are fixed; (2) the algorithm \( \hat{A} \) runs in polynomial time relative to the size

\[
\ell(i, \omega) = \ell(i) + \ell(\omega), \quad \omega \in \Omega_i,
\]

and hence, for any \( i \in I \) and any choice of \( \omega \in \Omega_i \), \( \hat{A}[i, \omega] \) runs in polynomial time in terms of \( \ell(i) \); (3) for all \( i \in I \) such that \( f(i) = 1 \), we have

\[
\hat{A}[i, \omega] = 1
\]
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

\begin{equation}
\frac{1}{|\Omega_i|} |\{\omega \in \Omega_i \mid \tilde{A}[i, \omega] = 1\}| \leq p.
\end{equation}

The probability $p$ is called the error rate of the probabilistic algorithm. Intuitively, the idea is to attempt to compute $f(i)$ by selecting $\omega \in \Omega_i$ uniformly at random and running $\tilde{A}[i, \omega]$. 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 (4.1), this hint will be the right answer with probability at least $1 - p$.

**Example 4.2.3 (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 probability test. It turns out that a probabilistic polynomial time algorithm for this question, with error rate $< 1/2$, can be devised quite easily: this is the Solovay-Strassen test [45], which goes back to 1977.

This test starts with the definition of the Legendre symbol modulo an odd prime $p$, which is the function

\[ \begin{cases} & (\mathbb{Z}/p\mathbb{Z}) \to \{-1, 0, 1\} \\ & a \mapsto \left( \frac{a}{p} \right) \end{cases} \]

where

\[ \left( \frac{a}{p} \right) = \begin{cases} 0 & \text{if } a = 0 \\ 1 & \text{if there exists } y \in \mathbb{F}_p \text{ such that } x = y^2 \\ -1 & \text{otherwise.} \end{cases} \]

The first crucial ingredient of the test is the fact, due to Euler, that the Legendre symbol, for a fixed prime $p$, can be computed by means of the congruence

\begin{equation}
\left( \frac{a}{p} \right) \equiv a^{(p-1)/2} \pmod{p}.
\end{equation}

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 $a \in \mathbb{Z}/p\mathbb{Z}$).

As a next, the Legendre symbol is extended to the Jacobi symbol modulo an odd integer $n \geq 1$, which is defined by

\[ \left( \frac{n}{m} \right) = \prod_{i=1}^{k} \left( \frac{n}{p_i} \right)^{v_i} \]

if $n = p_1^{v_1} \cdots p_k^{v_k}$ is the factorization of $n$ into distinct prime powers 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 4.2.4 (Quadratic reciprocity law).** (1) For any odd positive integers $n$ and $m$ which are coprime, we have

\[ \left( \frac{n}{m} \right) \left( \frac{m}{n} \right) = (-1)^{(n-1)(m-1)/4}. \]

---

1 This assumption that this choice can be done efficiently, which is in practice another non-trivial problem...
(2) For any odd integer \( n \), we have
\[
\left( \frac{-1}{n} \right) = (-1)^{(n-1)/2},
\]
and
\[
\left( \frac{2}{n} \right) = (-1)^{(n^2-1)/8}.
\]

Note the following corollary, which is of great importance for this example:

**Corollary 4.2.5 (Computation of Jacobi symbols).** There is a deterministic polynomial time algorithm \( J \) with inputs \( I = \mathbb{Z} \times \{ \text{odd integers} \} \) and outputs \( O = \{-1, 0, 1\} \) such that \( J(m, n) \) 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 \(|n| \leq m/2\), and use quadratic reciprocity to “switch” \( n \) and \( m \), then reduce \( m \) modulo \( n \) and repeat. The “supplementary laws” computing the Jacobi symbols \( \left( \frac{-1}{n} \right) \) and \( \left( \frac{2}{n} \right) \) 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.)

We can now define the Solovay-Strassen primality test \( S \): we take as input set
\[
I = \bigcup_{i \text{ odd integer}} \{(i) \times (\mathbb{Z}/i\mathbb{Z})^\times\},
\]
and define \( S[i, a] \), for \( i \geq 1 \) an odd integer and \( a \in (\mathbb{Z}/i\mathbb{Z})^\times \), to be 1 if
\[
\left( \frac{a}{i} \right) \equiv a^{(i-1)/2} \pmod{i},
\]
and 0 otherwise. According to Corollary 4.2.5, 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.

First of all, if \( i \) is an odd prime number, then by Euler’s formula (4.2), we have \( S[i, a] = 1 \) for all \( a \in (\mathbb{Z}/i\mathbb{Z})^\times \), 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
\[
\frac{1}{|\mathbb{Z}/i\mathbb{Z})^\times|} \left| \left\{ a \in (\mathbb{Z}/i\mathbb{Z})^\times \mid \left( \frac{a}{i} \right) \equiv a^{(i-1)/2} \pmod{i} \right\} \right|.
\]

To do this, we claim that if \( i \) is not prime, the set
\[
B = \{ a \in (\mathbb{Z}/i\mathbb{Z})^\times \mid \left( \frac{a}{i} \right) \equiv a^{(i-1)/2} \pmod{i} \}
\]
is a proper subgroup of \((\mathbb{Z}/i\mathbb{Z})^\times\). If that is the case, then
\[
\frac{|B|}{|\mathbb{Z}/i\mathbb{Z})^\times|} = \frac{1}{[\mathbb{Z}/i\mathbb{Z})^\times : B]} \leq \frac{1}{2},
\]
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 \( B \neq (\mathbb{Z}/i\mathbb{Z})^\times \) if \( i \) is not prime (which is really the point, since Euler’s formula precisely means that \( B = (\mathbb{Z}/i\mathbb{Z})^\times \) when \( i \) is prime).

Because of the Chinese Remainder Theorem, we may assume that \( i = p^v \) is a power of a prime, with \( v \geq 2 \). We recall that, since \( p \) is odd, the group \((\mathbb{Z}/p^v\mathbb{Z})^\times\) is cyclic of order
p^v - p^{v-1}. Thus, if we take for \( a \) a generator, we have \( B = (\mathbb{Z}/p^v\mathbb{Z})^\times \) 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 not equal to 1), the Jacobi symbol is \(-1\) (because \( a \) is not a square modulo \( p \)), and since \( p^v - 1 \equiv p^{v-1} - 1 \) (mod \( (p^v - p^{v-1}) \)), which is non-zero, we have \( a^{(p^v-1)/2} \neq 1 \).

The way expanders come in to this discussion is to address the following question: given a probabilistic algorithm \( \tilde{A} \) 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 \( \tilde{A} \) 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 \( \varepsilon \), \( 0 < \varepsilon \leq p \), is fixed beforehand, taking \( t = \lceil \log \frac{1}{\varepsilon} \log \frac{1}{p} \rceil \) will do.

Formally, this means that we consider the algorithm \( \tilde{A}_t \) with inputs \( I_t = \bigcup_{i \in I} (\{i\} \times \Omega_t^i) \) and
\[
\tilde{A}_t[i, \omega_1, \ldots, \omega_t] = \begin{cases} 1 & \text{if } \tilde{A}_t[i, \omega_1] = \cdots = \tilde{A}_t[i, \omega_t] = 1 \\ 0 & \text{otherwise}, \end{cases}
\]
(which still uses a polynomially bounded amount of extra randomness), we still have \( \tilde{A}_t[i, \omega] = 1 \) for any \( i \in I \) such that \( f(i) = 1 \) (by definition of \( \tilde{A} \)), and the error rate for \( i \) with \( f(i) = 0 \) is given by
\[
\frac{1}{|\Omega_t^i|} |\{(\omega_1, \ldots, \omega_t) \in \Omega_t^i | \tilde{A}_t[i, \omega_j] = 1 \text{ for all } 1 \leq j \leq t\}| = \left( \frac{1}{|\Omega_i|} |\{(\omega_1, \ldots, \omega) \in \Omega_i | \tilde{A}[i, \omega] = 1\}| \right)^t \leq p^t.
\]
(this simple computation, with the \( t \)-th power arising, is the formal incarnation of the intuitive notion of applying the original algorithms with independent choices of random bits.)

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 \( \Omega_i \) is easy. However, neither of these is clear in practice. We consider the first problem only: where one run of \( \tilde{A} \) requires intuitively \( \ell(\omega) \leq c_2 \ell(i)^B \) random bits, we need \( t \) times as many for \( \tilde{A}_t \). 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 \( \tilde{A} \) as above, we define a new probabilistic algorithm \( E_A \) 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 \( \Gamma_i = (\Omega_i, E, ep) \), with vertex step \( \Omega_i \), which is connected, \( d \)-regular for some \( d \), non-bipartite, and has equidistribution radius \( g_i = g_r \).

**Step 2.** Pick uniformly at random an initial vertex \( \omega_0 \in \Omega_i \);
Step 3. Start a random walk on $\Gamma_i$ with initial vertex $\omega_0$, say $(X_n^{(\omega_0)})_{n \geq 0}$;

Step 4. Return 1 if

$$\tilde{A}[i, X_0^{(\omega_0)}] = \cdots = \tilde{A}[i, X_t^{(\omega_0)}] = 1,$$

and 0 otherwise.

In other words, intuitively, we replace the $t$ independent choices of uniformly distributed $\omega_j$'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 (suitably connected) graph with vertex set $\Omega_i$. The necessary amount of randomness is now that needed for the choice of the first vertex $\omega_0$, and for throwing $t$ times a $d$-sided dice to perform the random walk. (Formally, we can take $\Omega_i^{EA} = \Omega_i \times \{1, \ldots, d\}^t$, assuming we also fix with $\Gamma_i$, explicit bijections from $\{1, \ldots, d\}$ to the neighbors of any vertex $\omega \in \Omega_i$.)

If $d$ is small, this is significantly less randomness than required for independent trials of $\tilde{A}$.

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 $\tilde{A}$ when $t = 0$, it has error rate at most $p$. Is it significantly smaller? To answer this, note that if $f(i) = 0$ and

$$B = \{\omega \in \Omega_i \mid \tilde{A}[i, \omega] = 1\} \subset \Omega_i$$

is the set of random choices of $\omega$ for which the original probabilistic algorithm give the wrong answer, the failure rate for $EA$ is given by

$$P(X_0 \in B, X_1 \in B, \ldots, X_t \in B)$$

where $(X_n)$ is now a random walk on $\Gamma_i$ with uniformly distributed initial step $X_0$.

To bound this probability, we have the following general result of Ajtai-Komlós-Szemerédi and Agon-Feige-Wigderson-Zuckerman, which has independent interest:

**Proposition 4.2.6** (Decay of “confinement” probabilities). Let $\Gamma = (V, E, ep)$ be a finite connected non-bipartite graph, and let $B \subset V$ be a subset of vertices. If $(X_n)$ is the random walk on $\Gamma$ with uniformly distributed initial step $X_0$, we have

$$P(X_j \in B \text{ for all } j, 0 \leq j \leq t) \leq (\mu(B) + \varrho \Gamma)^t$$

for all $t \geq 0$, where $\mu(B) = |B|/|V|$ is the graph measure of $B$ and $\varrho \Gamma$ the equidistribution radius of $\Gamma$.

Before proving this, let’s see how it applies to the study of the algorithm $EA$. In that case, we apply the proposition for the graph $\Gamma_i$ of Step 1, with $B$ given by (4.3). By construction, we have $\mu(B) = |B|/|\Omega_i| \leq p$, so we obtain the upper bound

$$(p + \varrho_i)^t$$

for the error rate. This is however trivial unless $p + \varrho_i < 1$, and it provides exponential decay of the error rate (as a function of $t$), comparable with the independent model $\tilde{A}_t$, only if $p + \varrho_i < \varrho_0 < 1$ where $\varrho_0$ is independent of $i$. This certainly requires that the family $(\Gamma_i)$ be an expander family, but this is not enough if $p$ is relatively large (say $p = 1/2$ as in the Solovay-Strassen primality test).

There is a work-around for this, if one knows an explicit bound $\varrho$ on the equidistribution radius for the family $(\Gamma_i)$. In that case, one can first replace $\tilde{A}$ with $\tilde{A}_s$ for some $s$ such that $p^s + \varrho = \varrho_0 < 1$, and then construct the corresponding probabilistic algorithm $EA_s$. Since $s$ will be fixed, this provides then a procedure for reducing arbitrarily the error rate of $A$ 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 \( EA \) is only effective (i.e., runs in polynomial time) if we have also a deterministic polynomial-time algorithm to construct the graphs \( \Gamma_i \). Here a probabilistic construction, or a construction of a family where the expansion parameters are not explicitly known, will not be sufficient.

**Proof of Proposition 4.2.6.** Let \( \varpi \) be the probability to compute. By direct expansion, we have

\[
\varpi = \frac{1}{|V|} \sum_{x_0 \in B} \sum_{x_1, \ldots, x_t \in B} P(X_1 = x_1 \mid X_0 = x_0) P(X_2 = x_2 \mid X_0 = x_0, X_1 = x_1) \cdots P(X_t = x_t \mid (X_0, \ldots, X_{t-1}) = (x_0, \ldots, x_{t-1})),
\]

and the Markov property of the random walk (3.7) reduces this to

\[
\varpi = \frac{1}{|V|} \sum_{x_0 \in B} \sum_{x_1, \ldots, x_t \in B} P(x_0, x_1) \cdots P(x_{t-1}, x_t)
\]

where the transition probability \( P(x, y) \) is defined in (3.8). Now we claim that this can also be expressed, in terms of the Markov operator \( M \), as the inner product

\[
\varpi = \langle 1_B, (MP_B)^t 1 \rangle,
\]

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

\[
P_B \varphi = 1_B \varphi,
\]

(and \( 1 \) is 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

\[
(MP)^t \varphi(x) = \sum_{x_1, \ldots, x_t \in V} P(x, x_1) \cdots P(x_{t-1}, x_t) 1_B(x_1) \cdots 1_B(x_t) \varphi(x_t),
\]

which follows by induction from (3.13).

Now, since \( M \) and \( P_B \) are self-adjoint and \( 1_B = P_B 1 \), we get

\[
\varpi = \langle (P_B M)^t P_B 1, 1 \rangle.
\]

But since \( P_B \) is a projection, we have \( P_B^2 = P_B \), and hence

\[
(P_B M)^t P_B = (P_B M)(P_B M) \cdots (P_B M)P = (P_B M P_B)(P_B M P_B) \cdots (P_B M P_B) = (P_B M P_B)^t,
\]

which gives the upper-bound

\[
\varpi \leq \|P_B M P_B\|^t.
\]

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 \( \varpi_T + \mu(B) \).

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 \( \psi \in L^2_0(\Gamma) \) (recall that \( \Gamma \) is not bipartite by assumption, so \( -1 \) is not an eigenvalue of \( M \)). By linearity, we obtain

\[
P_B M P_B \varphi = m 1_B + P_B M \psi,
\]

but since \( \|P_B\| \leq 1 \), the definition of \( \varpi_T \) gives

\[
\|P_B M \psi\| \leq \|M \psi\| \leq \varpi \|\psi\|,
\]

83
while
\[ \| m \mathbf{1}_B \| = \| m |\mu(B) \| \leq \| P_B \varphi \| \| \mu(B) \| \leq \| \varphi \|, \]
and hence we get the inequality
\[ \| P_B M P_B \varphi \| \leq (\mu(B) + \varrho \Gamma) \| \varphi \| \]
for any \( \varphi \in L^2(\Gamma) \), which establishes that the norm of \( P_B M P_B \) is at most \( \mu(B) + \varrho \Gamma \). \( \square \)

4.3. Sieve methods

4.4. Arithmetic geometry
CHAPTER 5

Expanders from finite linear groups

5.1. Introduction

In this chapter, we set out to prove Theorem 3.4.10 of Bourgain and Gamburd on expansion of Cayley graphs of $\text{SL}_2(F_p)$. We recall the statement:

THEOREM 5.1.1 (Expansion in subgroups of $\text{SL}_2(\mathbb{Z})$). Let $S \subset \text{SL}_2(\mathbb{Z})$ be any finite symmetric subset and let $G$ be the subgroup generated by $S$. For prime numbers $p$, let $\Gamma_p = \mathcal{C}(\text{SL}_2(F_p), S)$ be the Cayley action graph of finite quotient group $\text{SL}_2(F_p)$ with respect to the reduction modulo $p$ of the set $S$. Then $(\Gamma_p)_{p \geq p_0}$ is an expander family if and only if $\Gamma_p$ is connected for all $p \geq p_0$.

Although the arguments 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 $\text{SL}_2(F_p)$ which underlie the whole argument. Originally, the work of Bourgain and Gamburd was spurred by Helfgott’s “growth” theorem (which implies Theorem 3.4.11) 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...

REMARK 5.1.2 (Effectivity). 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. In a first version of these lecture notes, we gave such an argument. This was used as the basis of the paper [28], which the interested reader may consult to see how this is done, and to gauge the quality of the results... We will simply quote later some explicit results from this paper.

5.2. Preliminaries and strategy

Theorem 5.1.1 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 5.2.1 (Counting cycles). Let $\Gamma = (V, E, \text{ep})$ be a finite non-empty connected graph with Markov operator $M$, and let $(X_n)_{n \geq 0}$ be a random walk on $\Gamma$. For any integer $m \geq 0$, we have

$$\text{Tr}(M^m) = \sum_{x \in V} P(X_m = x \mid X_0 = x).$$
In particular, if $\Gamma = \mathcal{C}(G, S)$ is a Cayley graph and the random walk starts at $X_0 = 1$, we have

\begin{equation}
\frac{1}{|G|} \text{Tr}(M^m) = P(X_m = 1),
\end{equation}

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 $(\varphi_i)$ be an orthonormal basis of eigenfunctions of $M$ of $L^2(\Gamma)$, with $M \varphi_i = \lambda_i \varphi_i$. Then the trace of $M^m$ is the same as the sum of the $\lambda_i^m$. By orthonormality, we can write

$$\sum_i \lambda_i^m = \sum_i \langle \varphi_i, \varphi_i \rangle \lambda_i^m = \sum_i \left( \frac{1}{N} \sum_{x \in V} \text{val}(x) |\varphi_i(x)|^2 \right) \lambda_i^m.$$ 

After exchanging the sums, the inner summand is the value of $\text{val}(x) \frac{1}{N} \sum_i \lambda_i^m \overline{\varphi_i(x)} \varphi_i(y)$ at $y = x$. But the expression

$$\frac{\text{val}(x)}{N} \sum_i \lambda_i^m \overline{\varphi_i(x)} \varphi_i$$

is the spectral expansion, in the basis $(\varphi_i)$, of the characteristic function $\delta_x$ of the single point $x \in V$. Since the $\varphi_i$ are eigenfunctions of $M$ with eigenvalue $\lambda_i$, linearity implies that

$$\frac{\text{val}(x)}{N} \sum_i \lambda_i^m |\varphi_i(x)|^2 = (M^m \delta_x)(x).$$

By the basic property of the Markov operator (Lemma 3.2.13), we get

$$(M^m \delta_x)(x) = P(X_m(x) = x),$$

where $(X_m(x))$ refers to a random walk started at $x$. By the Markov property, this is the same as $P(X_m = x \mid X_0 = x)$ for an arbitrary random walk, and hence

$$\sum_i \lambda_i^m = \sum_{x \in V} P(X_m = x \mid X_0 = x).$$

When $\Gamma$ is a Cayley graph, the probability $P(X_m = x \mid X_0 = x)$ is independent of the starting point $x$, by homogeneity. Hence, selecting $x = 1$, we get

$$P(X_m = 1) = \frac{1}{|G|} \sum_i \lambda_i^m$$

for a random walk starting at $X_0 = 1$. 

**Example 5.2.2.** 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

$$\frac{1}{|G|} \text{Tr}(M^m) = P(X_m = 1) = \frac{1}{|S|^m} |\{(s_1, \ldots, s_m) \in S^m \mid s_1 \cdots s_m = 1\}|,$$

which is the number of “relations” in $G$ of length $m$ when presenting the group using the generators from $S$. 

86
Exercise 5.2.3 (Local spectral sums). The fact that the probability $\mathbf{P}(X_m^x = x)$ is independent of $x$ in the case of a Cayley graph has the following interesting interpretation: for any choice of orthonormal basis $(\varphi_i)$ as above, for any fixed $x_0 \in G$, we have
\[
\sum_i \lambda_i^m = \sum_i |\varphi_i(x_0)|^2 \lambda_i^m
\]
for all $m \geq 0$.

Show that this is a consequence of the following “local” spectral property: for any $x_0 \in G$, any orthonormal basis of $L^2(G)$, and any eigenvalue $\lambda$, we have
\[
\sum_{\lambda_i = \lambda} |\varphi_i(x_0)|^2 = \dim(\ker(M - \lambda))
\]
(the sum running over the eigenfunctions with eigenvalue $\lambda$). In particular, if $\lambda$ is a simple eigenvalue, it has constant modulus 1.

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 5.2.4. Let $\Gamma = \mathcal{C}(G, S)$ be a finite connected Cayley graph with Markov operator $M$, and let $(X_n)_{n \geq 0}$ be a random walk on $\Gamma$ with $X_0 = 1$ fixed. Let $\Lambda \subset [-1, 1]$ be the eigenvalues of $M$, with multiplicity $n(\lambda) \geq 1$ for $\lambda \in \Lambda$. For any subset $\Lambda_1 \subset \Lambda$ and for any integer $m \geq 0$, we have
\[
\frac{1}{|G|} \sum_{\lambda \in \Lambda_1} n(\lambda) \lambda^{2m} \leq \mathbf{P}(X_{2m} = 1),
\]
with equality if $\Lambda = \Lambda_1$.

This gives an upper bound for any eigenvalue $\lambda$, and in particular for the equidistribution radius $\rho_\Gamma$, 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 5.2.5 (Representation of $G$ on $L^2(G)$). Let $G$ be a finite group, $S \subset G$ a finite generating set.

1. The group $G$ acts by linear automorphisms on $L^2(G)$ by the left-regular representation
   \[ \text{reg}(g) \varphi(x) = \varphi(g^{-1}x) \]
   for all $x \in G$ and $g \in G$. This is a unitary representation
   \[ \text{reg} : G \rightarrow \text{U}(L^2(G)). \]

2. The regular representation commutes with the Markov operator of $\mathcal{C}(G, S)$, i.e.,
   \[ M(\text{reg}(g) \varphi) = \text{reg}(g)(M \varphi) \]
   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 $\text{reg}(g) \varphi = \varphi$ for all $g \in G$, if and only if $\lambda = 1$. 87
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

$$M(\text{reg}(g)\phi)(x) = \frac{1}{|S|} \sum_{s \in X} \phi(g^{-1}xs) = \text{reg}(M\phi)(x).$$

From, the stability of $\ker(M - \lambda)$ is easy: if $M\phi = \lambda\phi$, we have also

$$M(\text{reg}(g)\phi) = \text{reg}(g)(M\phi) = \text{reg}(g)(\lambda\phi) = \lambda \text{reg}(g)\phi,$$

i.e., $\text{reg}(g)\phi \in \ker(M - \lambda)$ for all $g \in G$, as claimed. Now if $\phi$ is any non-zero function invariant under the action of $G$, we get

$$\phi(x) = (\text{reg}(x)\phi)(1) = \phi(1)$$

for all $x$, which means that $\phi$ is constant. But then $M\phi = \phi$, so this only happens when $\lambda = 1$. \qed

The innocuous-looking corollary is the following:

**Corollary 5.2.6 (Bounding $\rho$ 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 $(X_n)_{n \geq 0}$ be a random walk on $\Gamma$ with $X_0 = 1$ fixed. Then

$$\rho \leq \left( \frac{|G|}{d(G)} P(X_{2m} = 1) \right)^{1/2m}$$

for all $m \geq 0$.

Proof. Indeed, $\rho$ is an eigenvalue $\neq 1$ of $M$, and hence the representation of $G$ on the $\rho$-eigenspace of $M$ is non-trivial. Thus the multiplicity of $\lambda$ as an eigenvalue is at least equal to $d(G)$, and we can apply the previous corollary with $\Lambda_1 = \{\rho\}$. \qed

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 5.4, to one such family, and we state here the corresponding result (for a proof, see Proposition B.2.1, (3)):

**Theorem 5.2.7 (Frobenius).** Let $\mathbf{F}_q$ be a finite field with $q$ elements. If $q$ is odd, we have

$$d(\text{SL}_2(\mathbf{F}_q)) = \frac{q - 1}{2}.$$

In particular, since $|\text{SL}_2(\mathbf{F}_q)| = q(q - 1)(q + 1)$, we have

$$d(\text{SL}_2(\mathbf{F}_q)) \sim |\text{SL}_2(\mathbf{F}_q)|^{1/3}$$

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 be pay close attention to detect it!) to a work of Huxley [23, §4], and was first applied by Sarnak and Xue [42] in contexts where other techniques to prove spectral gaps (specifically, Fourier expansions around cusps and estimates for Kloosterman sums) were not available. Another very relevant application is found in Gamburd’s thesis [15]. In fact, we will use again this property of $\text{SL}_2(\mathbf{F}_q)$ at a further step, in order to exploit Gowers’s notion of quasi-random groups (see Section 5.5).

88
Remark 5.2.8. Note that $d(G) \neq 1$ implies in particular that there can not be any surjective homomorphism $G \rightarrow \{\pm 1\} \subset U(C)$, and therefore that any Cayley graph of $G$ is non-bipartite.

The corollary leads to expander-quality bounds in the following situation: if we have

\[ P(X_{2m} = 1) \leq |G|^{-1+\varepsilon}, \]

for an a fixed $\varepsilon > 0$, when $m$ is relatively small, of size $m \leq \log |G|$, then we get

\[ \varepsilon_T \leq \exp\left(-\varepsilon \log |G| - \log d\right). \]

If, as happens for $G = SL_2(F_q)$, the group is such that $\log d > d \log |G|$ with a fixed $d > 0$, and if one can select $c > 0$ so that the bound (5.2) holds with $\varepsilon$ arbitrarily small, in particular $\varepsilon < d$, this leads to a uniform upper-bound for $\varepsilon_T$, namely

\[ \varepsilon_T \leq c = \exp\left(-\frac{d - \varepsilon}{2c}\right) < 1. \]

Here is an intermediate summary of this discussion:

**Proposition 5.2.9.** For $d > 0$, $\varepsilon > 0$ and $c > 0$, let $\mathcal{G}(d, \varepsilon, c)$ be the family of all finite connected Cayley graphs $\mathcal{C}(G, S)$ where $d(G) \geq |G|^d$ and

\[ P(X_{2m} = 1) \leq |G|^{-1+\varepsilon} \]

for some $m \leq \log |G|$. Then for any $\varepsilon < d$ and any graph $\Gamma \in \mathcal{G}(d, \varepsilon, c)$, the equidistribution radius satisfies

\[ \varepsilon_T \leq \exp\left(-\frac{d - \varepsilon}{2c}\right). \]

In particular, if $0 < \varepsilon < d$ and $c > 0$ are such that the family $\mathcal{G}(d, \varepsilon, c)$ 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 $P(X_{2m} = 1)$, 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 $P(X_{2m} = 1)$ converges to $1/|G|$ as $m \rightarrow +\infty$ (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 3.2.25).

It may be helpful to keep in mind the concrete interpretation of the probability $P(X_{2m} = 1)$: we want to prove that

\[ \frac{1}{|S|^{2m}}|\{(s_1, \ldots, s_{2m}) \in S^{2m} | s_1 \cdots s_{2m} = 1\}| \leq |G|^{-1+\varepsilon} \]

for $m$ of size $c \log |G|$.

In effect, the result of Bourgain-Gamburd can then be stated as follows:

**Theorem 5.2.10 (Bourgain-Gamburd).** Fix $S \subset SL_2(Z)$ a finite symmetric subset such that the projection of $S$ modulo $p$ generates $SL_2(F_p)$ for $p \geq p_0$. Then, for any $\varepsilon > 0$, there exists $c > 0$, depending on $S$ and $\varepsilon$, such that

\[ \mathcal{C}(SL_2(F_p), S) \in \mathcal{G}(1/7, \varepsilon, c) \]

for all $p \geq p_0$. 

89
5.3. The Bourgain-Gamburd argument

This section contains the great new ingredient discovered by Bourgain and Gamburd that 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 $\text{SL}_2(\mathbf{F}_p)$-valued symmetrically distributed random variable $X$ is not too concentrated, but also not very uniformly distributed on $\text{SL}_2(\mathbf{F}_p)$, 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 $X_{2k}, X_{4k}, \ldots, X_{2^j k}$, until the assumptions of the lemma fail. In that situation, it then seen that $m = 2^j k$ is of size $\ll \log |G|$, and that $P(X_m = 1)$ satisfies (5.2), 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 5.3.1 (Return probabilities).** Let $G$ be a finite group and let $X$ be a $G$-valued random variable which is symmetrically distributed, i.e.,

$$P(X = g) = P(X = g^{-1})$$

for all $g \in G$. The return probability $\text{rp}(X)$ is defined as

$$\text{rp}(X) = P(X_1 X_2 = 1),$$

where $(X_1, X_2)$ are independent random variables with the same distribution as $X$. Equivalently, we have

$$\text{rp}(X) = \sum_{g \in G} P(X = g)^2.$$

If $X, Y$ are two $G$-valued random variables, we denote

$$\text{rp}^+(X, Y) = \max(\text{rp}(X), \text{rp}(Y)).$$

**Example 5.3.2 (Uniform measures).** If $A \subset G$ is any subset and $X$ is uniformly distributed on $A$ (equivalently, if we consider the measure $\nu$ 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 $\text{rp}(X) = \frac{1}{|A|}$.

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 5.3.5 below, which it would be awkward to state now). This provides an approach to estimate $\text{rp}(X_1 X_2)$ in terms of $\text{rp}^+(X_1, X_2)$ for fairly general independent symmetric $G$-valued random variable $X_1$ and $X_2$. Only afterwards will we use special features of the groups $\text{SL}_2(\mathbf{F}_p)$.

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 $\text{rp}^+(X_1, X_2) = \text{rp}(X_1) = \text{rp}(X_2)$).
By definition, we have

\[ \text{rp}(X_1X_2) = \sum_{g \in G} \mathbf{P}(X_1X_2 = g)^2. \]

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 \( \text{rp}(X_1X_2) \) (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 \( I \geq 1 \), to be chosen later, and decompose \([\min \mathbf{P}(X = x), \max \mathbf{P}(X = x)] \subset [0,1] = \mathcal{J}_0 \cup \mathcal{J}_1 \cup \cdots \cup \mathcal{J}_I \)

where \( \mathcal{J}_i \) is, for \( 0 \leq i < I \), the dyadic interval

\[ \mathcal{J}_i = [2^{-i-1}, 2^{-i}], \]

and the final complementary interval is \( \mathcal{J}_I = [0,2^{-I}] \), to account for the small values. This gives corresponding partitions of \( G \) in subsets

\[ \begin{align*}
    A_{1,i} &= \{ x \in G \mid \nu_1(x) = \mathbf{P}(X_1 = x) \in \mathcal{J}_i \}, \\
    A_{2,i} &= \{ x \in G \mid \nu_2(x) = \mathbf{P}(X_2 = x) \in \mathcal{J}_i \},
\end{align*} \]

about which we note right now that, for \( 0 \leq i < I \), we have a rough size estimate

(5.3) \[ |A_{j,i}| \leq 2^{i+1}, \quad j = 1,2. \]

Using this decomposition into the formula above, and the fact that

\[ \mathbf{P}(X_1X_2 = g, X_1 \in A_{1,I} \text{ or } X_2 \in A_{2,I}) \leq \mathbf{P}(X_1 \in A_{1,I}) + \mathbf{P}(X_2 \in A_{2,I}) \leq \frac{|G|}{2^{I-1}}, \]

we obtain

\[ \begin{align*}
    \text{rp}(X_1X_2) &= \sum_{g \in G} \left( \sum_{0 \leq i,j \leq I} \mathbf{P}(X_1X_2 = g, X_1 \in A_{1,i}, X_2 \in A_{2,j}) \right)^2 \\
    &\leq 8|G|^3 2^{-2I} + 2 \sum_{g \in G} \left( \sum_{0 \leq i,j < I} \mathbf{P}(X_1X_2 = g, X_1 \in A_{1,i}, X_2 \in A_{2,j}) \right)^2 \\
    &\leq 2^{2I-2I}|G|^3 + 2I^2 \sum_{0 \leq i,j < I} \sum_{g \in G} \mathbf{P}(X_1X_2 = g, X_1 \in A_{1,i}, X_2 \in A_{2,j})^2
\end{align*} \]

91
by the Cauchy-Schwarz inequality. Furthermore, the inner sums in the second term, say \( B(A_{1,i}, A_{2,j}) \) are given by
\[
B(A_{1,i}, A_{2,j}) = \sum_{g \in G} \mathbf{P}(X_1X_2 = g, \ X_1 \in A_{1,i}, \ X_2 \in A_{2,j})^2
\]
\[
= \sum_{g \in G} \left( \sum_{(x,y) \in A_{1,i} \times A_{2,j}} \mathbf{P}(X_1 = x) \mathbf{P}(X_2 = y) \right)^2
\]
\[
= \sum_{x_1, x_2 \in A_{1,i}, y_1, y_2 \in A_{2,j}} \nu_1(x_1) \nu_1(x_2) \nu_2(y_1) \nu_2(y_2)
\]
\[
\leq 2^{-2i-2j} \left| \{ (x_1, x_2, y_1, y_2) \in A_{1,i}^2 \times A_{2,j}^2 \mid x_1y_1 = x_2y_2 \} \right|,
\]
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 5.3.3 (Multiplicative energy).** Let \( G \) be a finite group and \( A, B \) subsets of \( G \). The *multiplicative energy* \( E(A, B) \) is given by
\[
E(A, B) = \left| \{ (x_1, x_2, y_1, y_2) \in A^2 \times B^2 \mid x_1y_1 = x_2y_2 \} \right|,
\]
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
\[
\text{rp}(X_1X_2) \leq 2^{3-2i}|G|^3 + 2I^2 \sum_{0 \leq i, j < I} 2^{-2(i+j)} E(A_{1,i}, A_{2,j})
\]

We now want to insert, for comparison, the return probability \( \text{rp}^+(X_1, X_2) \) 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 5.3.4.** (1) For any finite group \( G \), and any subsets \( A, B \subset G \), we have
\[
E(A, B) \leq \min(|A|^2|B|, |A||B|^2).
\]

(2) With notation as above, for all \( i \) and \( j \), we have
\[
2^{-2(i+j)} E(A_{1,i}, A_{2,j}) \leq 4^i \text{rp}^+(X_1, X_2)e(A_{1,i}, A_{2,j}),
\]
and for all \( \alpha \geq 1 \), we have
\[
2^{-2(i+j)} E(A_{1,i}, A_{2,j}) \leq \alpha^{-1} \text{rp}^+(X_1, X_2),
\]
unless
\[
\frac{|A_{1,i}|}{2^i} \geq \frac{1}{2\sqrt{\alpha}}, \quad \frac{|A_{2,j}|}{2^j} \geq \frac{1}{2\sqrt{\alpha}}.
\]

(3) If \( |A_{1,i}| \geq \alpha^{-1}|G| \) for some \( \alpha \geq 1 \), then
\[
2^{-2(i+j)} E(A_{1,i}, A_{2,j}) \leq 2^i \alpha |G|^{-1}.
\]
\textbf{Proof.} (1) follows from the definition

\[ E(A, B) = \left| \{(x_1, x_2, y_1, y_2) \in A^2 \times B^2 \mid x_1 y_1 = x_2 y_2 \} \right|, \]

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 \(E(A, B) \leq |A|^2|B|\), and the second gives \(E(A, B) \leq |A||B|^2\).

(2) This is in some sense the crucial point of the whole argument, and yet it is surprisingly simple. We remark that

\[ E = \left| \{(A, B) \mid \right. \]

\(\text{for } (i, j) \in Q_\alpha, \) the second part of the lemma gives us

\[ 2^{-2(i+j)} E(A_1, i, A_2, j) \leq \alpha^{-2} \text{rp}(X_1, X_2), \]

and for \((i, j) \in \bar{Q}_\beta, \) the third part gives

\[ 2^{-2(i+j)} E(A_1, i, A_2, j) \leq 2^4 \beta |G|^{-1}, \]

by (5.3), which was the first goal.

If instead we assume that \(2^{-2(i+j)} E(A_1, i, A_2, j) > \alpha^{-1} \text{rp}(X_1, X_2), \) then we write simply

\[ 2^{-2(i+j)} |A_1, i|^2 |A_2, j| \geq 2^{-2(i+j)} E(A_1, i, A_2, j) \geq \alpha^{-1} \text{rp}(X_2) \geq \alpha^{-1} |A_2, j| \]

and get the first inequality of (5.5), the second being obtained symmetrically.

(3) This is also elementary: using (5.5) twice, we have

\[ 2^{-2(i+j)} E(A_1, i, A_2, j) \leq \frac{|A_1, i|^2 |A_2, j|^2}{2^{2i+2j}} \leq 2^{-i-3} \leq 2^4 |A_1, i|^{-1} \]

and hence the assumption leads directly to

\[ 2^{-2(i+j)} E(A_1, i, A_2, j) \leq 2^4 \alpha |G|^{-1}. \]

\[ \Box \]
and thus we have now shown that

\[
\rp(X_1X_2) \leq 2^{3-2I}|G|^3 + 2^4\beta I^4|G|^{-1} + 2\alpha^{-2}\rp(X)I^4 + 2^5\rp^+(X_1, X_2)I^2 \sum_{(i,j)\in P} e(A_1,i, A_2,j),
\]

where we just estimated the size of \(Q_\alpha\) and \(\tilde{Q}_\beta\) by \(I^2\). We now select

\[
I = \left\lceil \frac{2\log 2|\mathcal{G}|}{\log 2} \right\rceil \leq 3\log(3|\mathcal{G}|),
\]

and hence obtain a first basic inequality:

**Theorem 5.3.5 (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 \(\alpha\), \(\beta \geq 1\), we have

\[
(5.8) \quad \rp(X_1X_2) \ll \left(\frac{\log 3|G|}{|G|}\right)^4\beta \frac{1}{\alpha^2} + \rp^+(X_1, X_2)(\log 3|G|)^4 \left\{\sum_{(i,j)\in P} e(A_1,i, A_2,j)\right\}
\]

where the implied constant is absolute.

Intuitively, \(\beta\) will be a small power of \(|\mathcal{G}|\), and the first term here is then close to \(|\mathcal{G}|^{-1}\). 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 \(\rp(X_1)\) and \(\rp(X_2)\), with the parameters \(\alpha\) and \(\beta\) at our disposal, this 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 5.3.6 (Extreme of the normalized energy).** Let \(G\) be a finite group and \(A, B \subset G\) non-empty subsets. We have \(e(A, B) \leq 1\), with equality if and only if there exists a subgroup \(H \subset G\), and elements \(x, y \in G\) such that

\[
A = xH, \quad B = Hy.
\]

Moreover, if \(e(A, B) \geq \alpha^{-1}\) with \(\alpha \geq 1\), we have

\[
(5.9) \quad \alpha^{-2}|A| \leq |B| \leq \alpha^2|A|.
\]

**Proof.** Let \(a = |A|\) and \(b = |B|\) for simplicity. We already know that

\[
E(A, B) \leq \min(a^2b, ab^2),
\]

(by \((5.4))\), and to deduce \(e(A, B) \leq 1\), we observe that

\[
(5.10) \quad \min(a^2b, ab^2) \leq (ab)^{3/2},
\]

if need be by considering the cases \(a \leq b\) and \(b \leq a\) separely (for example, in the first case, we have

\[
\min(a^2b, ab^2) = a^2b = a^{3/2}a^{1/2}b \leq (ab)^{3/2}
\]

and the other case is symmetric.)

We now prove \((5.9)\). If \(e(A, B) \geq \alpha^{-1}\), with \(\alpha \geq 1\), and if \(b \leq a\), we deduce

\[
\alpha^{-1}(ab)^{3/2} \geq E(A, B) \geq \min(ab^2, a^2b) = ab,
\]

so that

\[
\alpha^{-2}a \leq b \leq a \leq \alpha^2a,
\]

94
which is (5.9) in that case. Of course, assuming $b \leq a$ 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

\[
\{(x_1, x_2, y_1, y_2) \in A^2 \times B^2 \mid x_1y_1 = x_2y_2\} = \{(h_1, h_2, h_3, h_4) \in H^4 \mid x_1h_3y = x_2h_4y\}
\]

which contains $|H|^3 = (|A||B|)^{3/2}$ elements.

For the converse, we note first that (5.9) 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 $|H| \leq a$. We will now prove that $b \leq |H|$: 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

\[|H| \leq a = b \leq |Hy_2| = |H|\]

so there must be equality $|H| = a = b$, and then there must also be equalities in the inclusions we used, i.e.,

\[x_1H = A, \quad Hy_2 = B,\]

which was our desired conclusion. \(\square\)

Comparing this statement with the inequality (5.8), we can see that each term $e(A_{i, i}, A_{j, j})$ 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 5.3.7 (Baby case).** Show that, with notation as in (5.8), there exists an (explicit) absolute constants $\delta > 0$ such that, if there exists $(i, j) \in P$ with $e(A_{i, i}, A_{j, j}) = 1$, there also exists a subgroup $H \subset G$ and $x \in G$ for which

\[P(X \in xH) \geq 4\alpha^{-1}.\]

Show that $H$ is a proper subgroup of $G$ unless

\[\text{rp}(X) \leq 4|G|^{-1}.\]

If we assume that $\alpha$ is a small power of $|G|$, this means that if there is a term in (5.8) with $e(A_{i, i}, A_{j, j}) = 1$, 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 (5.8) is less than 1 is not particularly useful,\(^1\) and we want a significantly better estimate on $e(A_{i, i}, A_{j, j})$. Precisely, we are looking for a structural understanding of pairs of sets $A, B \subset G$ such that $e(A, B) \geq \alpha^{-1}$ with $\alpha$ 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.

One can indeed do this for many groups, such as $\text{SL}_2(F_p)$. 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 $H \subset G$

\[^1\) Because it is not difficult to check that $\text{rp}(X_1X_2) \leq \text{rp}(X)$ anyway.
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 usually much more involved, we must classify approximate subgroups for certain families of finite groups \( G \). For the groups \( \text{SL}_2(\mathbb{F}_p) \), such a classification is equivalent with Helfgott’s theorem, which informally will show that all approximate subgroups of \( \text{SL}_2(\mathbb{F}_p) \) are essentially controlled by actual subgroups. As we will survey in Section 5.7, such a classification has now been established in much greater generality.

We start with defining approximate subgroups, following Tao (see [46, Def. 3.8]).

**Definition 5.3.8 (Approximate subgroup).** let \( G \) be a finite group and \( \alpha \geq 1 \). A non-empty subset \( H \subseteq G \) is an \( \alpha \)-approximate subgroup if \( 1 \in H \), \( H = H^{-1} \) and there exists a symmetric subset \( X \subseteq G \) of order at most \( \alpha \) such that

\[
H \cdot H \subseteq X \cdot H,
\]

which implies also \( H \cdot H \subseteq H \cdot X \). The **tripling constant** of \( H \) is defined by

\[
\text{trp}(H) = \frac{|H \cdot H \cdot H|}{|H|}.
\]

**Remark 5.3.9.** (1) Early works concerning approximate subgroups sometimes include further conditions. For instance, in [46, Def. 3.8], it is also asked that \( X \subseteq H \cdot H \), and in [47], it is further necessary that \( X \cdot H \subseteq H \cdot X \cdot X \). 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 (5.11) we have an immediate bound for the tripling constant:

\[
H \cdot H \cdot H \subseteq (X \cdot H) \cdot H \subseteq (X \cdot X) \cdot H
\]

leads to

\[
\text{trp}(H) \leq |X|^2 \leq \alpha^2.
\]

However, if one is concerned with explicit upper bounds, one may well know a better bound that this, as in the next theorem.

We can then state a generalization of Lemma 5.3.6 where the condition \( e(A, B) = 1 \) is relaxed.

**Theorem 5.3.10 (Sets with large multiplicative energy).** Let \( G \) be a finite group and \( \alpha \geq 1 \). If \( A \) and \( B \) are subsets of \( G \) such that \( e(A, B) \geq \alpha^{-1} \), there exist constants \( \beta_1, \beta_2, \beta_3 \geq 1 \), such that \( \beta_i \leq c_i \alpha^{c_2} \) for some constants \( c_1, c_2 > 0 \), and there exist a \( \beta_1 \)-approximate subgroup \( H \subseteq G \) and elements \( x, y \in G \) with

\[
|H| \leq \beta_2 |A| \leq \beta_2 \alpha^2 |B|,
\]

\[
|A \cap xH| \geq \frac{1}{\beta_3} |A|, \quad |B \cap Hy| \geq \frac{1}{\beta_3} |B|,
\]

\[
\text{trp}(H) \leq \beta_4.
\]

In fact, one can take \( c_1 = 2^{2424} \) and \( c_2 = 937 \), and more precisely

\[
\beta_1 \leq 2^{1851} \alpha^{720}, \quad \beta_2 \leq 2^{325} \alpha^{126}, \quad \beta_3 \leq 2^{2424} \alpha^{937}, \quad \beta_4 \leq 2^{930} \alpha^{360}.
\]

Except for the values of the constants, this is proved in [46, Th. 5.4, (i) implies (iv)] and quoted in [47, Th. 2.48]. For the values of the constant, see [28, Th. 2.1, Appendix A].
Exercise 5.3.11. (1) Show that a 1-approximate subgroup of $G$ is just a subgroup.

(2) Let $G = \mathbb{Z}/m\mathbb{Z}$ where $m \geq 1$ is an integer, and let $H \subset G$ 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 $r, s \in \mathbb{F}_p^\times$. Let

$$H = \left\{ \begin{pmatrix} r^n & x & y \\ 0 & s^n & z \\ 0 & 0 & (rs)^{-n} \end{pmatrix} \mid x, y, z \in \mathbb{F}_p, -k < n < k \right\} \subset \text{SL}_3(\mathbb{F}_p).$$

Show that $H$ is an $\alpha$-approximate subgroup of $\text{SL}_3(\mathbb{F}_p)$ for some $\alpha$ independent of $p$ and $k$.

We can now combine Theorem 5.3.10 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 5.3.12.** For $\delta > 0$, a finite group $G$ is $\delta$-flourishing if any symmetric subset $H \subset G$, containing 1, which generates $G$ and has tripling constant $\text{trp}(H) < |H|^{\delta}$ 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 5.6, can be stated as follows: there exists $\delta > 0$, an absolute constant, such that $\text{SL}_2(\mathbb{F}_p)$ is $\delta$-flourishing for all primes $p$ (in fact, we will prove this with $\delta = 1/1344$). The following exercise gives some useful basic insight on the nature of this property.

**Exercise 5.3.13.** (1) Show that there exists no $\delta > 0$ such that $\mathbb{Z}/p\mathbb{Z}$ is $\delta$-flourishing for all primes $p$.

(2) Show that if $G$ is $\delta$-flourishing, there exists some explicit $\delta_1 > 0$ (depending on $\delta$) such that any symmetric generating subset $H$ containing 1 of size $|H| \geq |G|^{1-\delta_1}$ satisfies $H \cdot H \cdot H = G$.

(3) Show that if $G$ is $\delta$-flourishing, there exists $A > 0$, depending only on $\delta$, such that

$$\text{diam}(\mathbb{C}(G, S)) \leq (\log |G|)^A$$

for any symmetric generating set $S$ of $G$.

We now establish a general form of $L^2$-flattening.

**Theorem 5.3.14 ($L^2$-flattening conditions).** Let $G$ be a finite group which is $\delta$-flourishing for some $\delta$ with $0 < \delta \leq 1$. Let $X_1, X_2$ be $G$-valued independent symmetric random variables. Let $0 < \gamma < 1$ be given, and assume that

$$P(X_1 \in xH) \leq |G|^{-\gamma}$$

for all proper subgroups $H \subset G$ and $x \in G$. 97
Then for any \( \varepsilon > 0 \), there exists \( \delta_1 > 0 \), depending only on \( \varepsilon, \delta \) and \( \gamma \), 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 5.3.15.** 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 (5.8) with \( \alpha = |G|^{|\delta_1-\varepsilon} \), for some \( \delta_1 > 0 \) to be chosen later, keeping \( \beta \) free for the moment. We obtain

\[
\mathrm{rp}(X_1X_2) \ll \left\{ |G|^{-1+\varepsilon} + \mathrm{rp}^+(X_1, X_2)|G|^{-2\delta_1} + \mathrm{rp}^+(X_1, X_2) \sum_{(i,j) \in P} e(A_{1,i}, A_{2,j}) \right\}
\]

where the implied constant depends only on \( \varepsilon \).

Let then

\[
R = R_\alpha = \{(i, j) \in P \mid e(A_{1,i}, A_{2,j}) \geq \alpha^{-1}\} \subset P,
\]

so that the contribution of those \((i, j) \in P\) which are not in \( R_\alpha \), 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 \( \varepsilon \).

This is of the right shape. We will now analyze the set \( R_\alpha \) and show that it is empty when \( \delta_1 \) is chosen small enough, and \( \beta \) is well-chosen. By Theorem 5.3.10, for each \((i, j) \in R\), there exists a \( \beta_4 \)-approximate subgroup \( H_{i,j} \) and elements \((x_i, y_j) \in A_{1,i} \times A_{2,j}\) such that

\[
|H_{i,j}| \leq \beta_2|A_{1,i}|, \quad |A_{1,i} \cap x_iH_{i,j}| \geq \beta_3^{-1}|A_{1,i}|, \quad |A_{2,j} \cap H_{i,j}y_j| \geq \beta_3^{-1}|A_{2,j}|,
\]

and with tripling constant bounded by \( \beta_4 \), where

\[
\beta_4 \leq c_1|G|^{c_2|\delta_1|}
\]

for some absolute constants \( c_1, c_2 > 0 \). We then note first that if \( H_{i,j} \) denotes the “ordinary” subgroup generated by \( H_{i,j} \), we have

\[
P(X \in x_iH_{i,j}) \geq P(X \in x_iH_{i,j})
\]

\[
\geq P(X \in A_{1,i} \cap x_iH_{i,j}) \geq \frac{1}{\beta_3} \frac{|A_{1,i}|}{2+1} \geq \frac{1}{4\beta_3\alpha} \geq \frac{1}{|G|^{(1+c_2|\delta_1|}},
\]

with an absolute implied constant, using the fact that elements of \( P \) are not in the set (5.6). If \( \delta_1 \) is small enough that

(5.14) \( (1+c_2)\delta_1 < \gamma \),

and if \( |G| \) is large enough, this is not compatible with (5.13), and hence, for such a choice of \( \delta_1 \), we deduce that each \( H_{i,j} \) (if any!) generates the group \( G \).

We next observe that \( H_{i,j} \) can not be extremely small, which will allow us to relate the tripling constant to the size of \( H_{i,j} \) instead of that of \( G \). Indeed, we have

\[
|H_{i,j}| \geq |x_iH_{i,j} \cap A_{1,i}| \geq \beta_3^{-1}|A_{1,i}|,
\]

on the one hand, and by applying (5.13) with \( H = 1 \), we can see that \( A_{1,i} \) is not too small, namely

\[
|A_{1,i}| \geq \frac{P(X \in A_{1,i})}{\max_{g \in G}P(X = g)} \geq |G|^\gamma \frac{P(X \in A_{1,i})}{2+1} \geq \frac{|G|^\gamma}{4\alpha}.
\]

98
using again the definition of $P$.

This gives the lower bound

$$|H_{i,j}| \geq \frac{|G|^{\gamma}}{4\alpha^2\beta_3} \gg |G|^{\gamma_1}$$

with $\gamma_1 = \gamma - \delta_1(1 + c_2) \geq 0$ by (5.14), for some absolute implied constant. This leads to control of the tripling constant, namely

$$\text{trp}(H_{i,j}) \leq \beta_4 \leq c_1|G|^{c_2\delta_1} \ll |H_{i,j}|^{c_2\delta_1 \gamma_1^{-1}}$$

where the implied constant depends on $\gamma$ and $\delta_1$.

Since we assumed that $G$ is $\delta$-flourishing, we see from Definition 5.3.12 that if $\delta_1$ is such that

$$\frac{c_2\delta_1}{\gamma_1} = \frac{c_2\delta_1}{\gamma_1} - (1 + c_2)\delta_1 < \delta,$$

and again if $|G|$ is large enough, the approximate subgroup $H_{i,j}$ must in fact be very large, specifically it must satisfy

$$H_{i,j} \cdot H_{i,j} \cdot H_{i,j} = G,$$

and in particular

$$|H_{i,j}| \geq \frac{|H_{i,j} \cdot H_{i,j} \cdot H_{i,j}|}{\beta_4} = \frac{|G|}{\beta_4}.$$

Then we get

$$|A_{1,i}| \geq |A_{1,i} \cap x_i H_{i,j}| \geq \frac{|H_{i,j}|}{\beta_2} \geq \frac{|G|}{\beta_2\beta_4} \gg |G|^{1-2c_2\delta_1}$$

where the implied constant is absolute. If we now select $\beta = M|G|^{2c_2\delta_1}$ for $M$ large enough, this implies that $(i, j) \in \hat{Q}_\beta$. Since $P$ is also in the complement of (5.7), 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 (5.14) and (5.15) are satisfied, we have

$$\text{rp}(X_1, X_2) \ll |G|^{-1+2c_2\delta_1+\varepsilon} + |G|^{-\delta_1} \text{rp}^+(X_1, X_2),$$

where the implied constant depends on $\varepsilon$, $\delta_1$ and $\gamma$. Fixing $\delta_1$ 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 5.3.16 (Uniformity can only increase).** Let $G$ be a finite group, $S$ a symmetric generating subset, and let $(X_n)$ be the corresponding random walk on $\mathbb{C}(G, S)$. For any $n \geq 1$ and $m \geq n$ we have $\text{rp}(X_m) \leq \text{rp}(X_n)$.

**Proof.** By the spectral interpretation (5.1) of the return probability, we have

$$\text{rp}(X_m) = P(X_{2m} = 1) = \frac{1}{|G|} \text{Tr}(M^{2m}), \quad \text{rp}(X_{2n}) = P(X_{2n} = 1) = \frac{1}{|G|} \text{Tr}(M^{2n})$$

were $M$ is the Markov operator. Since all eigenvalues of $M^2$ are non-negative and $\leq 1$, it follows that

$$\text{Tr}(M^{2m}) \leq \text{Tr}(M^{2n})$$

for $m \geq n$, as desired.

99
We can summarize the conclusion of all this section as follows, in the spirit of Proposition 5.2.9.

**Corollary 5.3.17 (The Bourgain-Gamburd expansion criterion).** Let \( c = (c,d,\delta,\gamma) \) be a tuple of positive real numbers, and let \( \mathcal{S}_2(c) \) be the family of all finite connected Cayley graphs \( \mathcal{C}(G,S) \) for which the following conditions hold:

1. We have \( d(G) \geq |G|^d \);
2. The group \( G \) is \( \delta \)-flourishing;
3. For the random walk \( (X_n) \) on \( G \) with \( X_0 = 1 \), we have that
   \[
P(X_{2k} \in xH) \leq |G|^{-\gamma}
   \]
   for some \( k \leq c \log |G| \) and all \( x \in G \) and proper subgroups \( H \subset G \).

Then, if \( c \) is such that \( \mathcal{S}_2(c) \) 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 5.2.9, the content of this corollary is therefore that these two conditions imply that, after \( \leq m \log |G| \) steps, for some large \( m \), the random walk will become “almost” uniform, allowing us to apply Corollary 5.2.6.

**Proof.** Let \( \Gamma = \mathcal{C}(G,S) \) be a graph in \( \mathcal{S}_2(c) \). We apply Theorem 5.3.14 with \( 0 < \varepsilon < d \), say \( \varepsilon = d/2 \). Let \( \delta_1 \) be such that the \( L^2 \)-flattening inequality holds for this value, so that

\[
rp(Y_1 Y_2) \ll \max \left( \frac{1}{|G|^{1-d/2}}, \frac{\rp(Y_1, Y_2)}{|G|^\delta_1} \right)
\]

for random variables \( Y_1, Y_2 \) which satisfy the assumptions of this theorem.

Let \( k = |c \log |G|| \) be given by (3). We apply the theorem to \( Y_1 = X_{2^j k} \) and \( Y_2 = X_{2^j (j+1) k} Y_1^{-1} \) for \( j \geq 0 \). These are indeed independent and symmetric random variables, and Conditions (2) and (3) imply that we can indeed apply Theorem 5.3.14 to these random variables for any \( j \geq 2 \). Since \( Y_1 \) and \( Y_2 \) are identically distributed, we have

\[
\rp(Y_1, Y_2) = \rp(Y_1) = \rp(X_{2^j k}).
\]

Thus, applying the theorem, we obtain by induction

\[
\rp(X_{2^j k}) \ll \rp(X_k) \frac{|G|^{-j\delta_1/2}}{|G|^{-j\delta_1/2}} \ll |G|^{-j\delta_1/2}
\]

when \( j \) is such that

\[
|G|^{1-d/2} > |G|^{j\delta_1/2},
\]

and for larger \( j \), we get

\[
\rp(X_{2^j k}) \ll |G|^{1-d/2},
\]

where the implied constants depend only on \((d, \gamma)\). In particular, we obtain this last inequality for

\[
j \ll \frac{1}{\delta_1}
\]

which, by the “cycle-counting” Corollary (5.2.6), gives

\[
\varrho_\Gamma \leq \left( |G|^{1-d} \rp(X_{2^j k}) \right)^{1/(2^j k)} \leq \exp(-cd)
\]

for some constant \( c > 0 \) which is independent of \( \Gamma \in \mathcal{S}_2(c) \). This proves the theorem. \( \square \)
5.4. Implementing the Bourgain-Gamburd argument

Theorem 5.1.1 will now be proved by applying the criterion of Corollary 5.3.17. Thus we will consider the groups $G_p = \text{SL}_2(\mathbb{F}_p)$ for $p$ prime, for which Condition (1) of the Bourgain-Gamburd criterion (which is purely a group-theoretic property) is known: this is Theorem 5.2.7 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 5.6. However, it is still purely a property of the groups $\text{SL}_2(\mathbb{F}_p)$.

Condition (3), on the other hand, involves the choice of generating sets. The symmetric generating sets $S_p$ in Theorem 5.1.1 are assumed to be obtained by reduction modulo $p$ of a fixed symmetric subset $S \subseteq \text{SL}_2(\mathbb{Z})$. What makes this situation special, and in particular makes it possible to check Condition (3) of the criterion, is the following special case: if $S \subseteq \text{SL}_2(\mathbb{Z})$ 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 $P(X_k = g)$ that we need to estimate at the level of this tree, where they are easier to analyze.

We begin with a classical proposition, whose idea goes back to Margulis. For the statement, recall that the norm of a matrix $g \in \text{GL}_n(\mathbb{C})$ is defined by
\[
\|g\| = \max_{v,w \neq 0} \frac{|\langle gv, w \rangle|}{\|v\| \|w\|}
\]
where $\langle \cdot, \cdot \rangle$ is the standard inner product on $\mathbb{C}^n$. This satisfies
\[
(5.16) \quad \|g_1 g_2\| \leq \|g_1\| \|g_2\|, \quad \max_{i,j} |g_{i,j}| \leq \|g\| \text{ for } g = (g_{i,j}),
\]
the latter because $g_{i,j} = \langle ge_i, e_j \rangle$ in terms of the canonical basis.

**Proposition 5.4.1** (Large girth for finite Cayley graphs). Let $S \subseteq \text{SL}_2(\mathbb{Z})$ be a symmetric set, and let $\Gamma = \mathcal{C}(G,S)$ be the corresponding Cayley graphs. Let $\tau > 0$ be defined by
\[
(5.17) \quad \tau^{-1} = \log \max_{s \in S} \|s\| > 0,
\]
which depends only on $S$.

1. For all primes $p$ and all $r < \tau \log(p/2)$, where $G_p = \text{SL}_2(\mathbb{F}_p)$, the subgraph $\Gamma_r$ induced by the ball of radius $r$ in $\Gamma$ 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 $\text{girth}(\mathcal{C}(G_p,S))$ is at least equal to $2\tau \log(p/2)$.

**Proof.** The main point is that if all coordinates of two matrices $g_1, g_2 \in \text{SL}_2(\mathbb{Z})$ 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 $m \leq r$ and $s_i \in S$. Using (5.16), we get
\[
\max_{i,j} |x_{i,j}| \leq \|x\| \leq \|s_1\| \cdots \|s_m\| \leq e^{m/\tau} \leq e^{r/\tau}.
\]

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 $x \equiv y \pmod{p}$ implies $x = y$, which is (1).
Then (2) follows because any embedding of a cycle $\gamma : C_m \rightarrow \mathcal{C}(G_p, S)$ such that $\gamma(0) = 1$ and such that
\[d(1, \gamma(i)) \leq m/2 < \tau \log(p/2)\]
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 $m = \text{girth}(\mathcal{C}(G_p, S))$ must satisfy $m/2 \geq \tau \log(p/2)$. □

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 5.4.2 (Decay of probabilities).** Let $S \subset SL_2(\mathbb{Z})$ 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 $G_p = SL_2(\mathbb{F}_p)$, 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 $x \in SL_2(\mathbb{F}_p)$, we have
\[P(X_n = x) \leq |G_p|^{-\gamma}\]
for
\[n \leq \tau \log(p/2),\]
where $\tau$ is defined in Proposition 5.4.1.

**Proof.** There exists $\tilde{x} \in G$ such that $\tilde{x}$ reduces to $x$ modulo $p$ and $\tilde{x}$ is at the same distance to 1 as $x$, and by Proposition 5.4.1, (2), we have
\[P(X_n = x) = P(\tilde{X}_n = \tilde{x}),\]
for $n \leq \tau \log(p/2)$, where $(\tilde{X}_n)$ is the random walk starting at 1 on the $|S|$-regular tree $\mathcal{C}(G, S)$. By Proposition B.1.5 from Appendix B, we have
\[P(\tilde{X}_n = \tilde{x}) \leq r^{-n} \quad \text{with} \quad r = \frac{|S|}{2\sqrt{|S| - 1}},\]
for all $n \geq 1$ and all $\tilde{x} \in G$. For $p$ large enough, this implies
\[P(X_n = x) \leq |G_p|^{-\gamma}\]
where $\gamma > 0$ depends only on $S$. □

In order to deal with cosets of other proper subgroups of $SL_2(\mathbb{F}_p)$, 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 Theorem B.2.2 in Appendix B, we see that if $p \geq 5$ and $H \subset SL_2(\mathbb{F}_p)$ 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
\[(x_1, x_2), [x_3, x_4]] = 1.\]

The first ones are “small”: if $H$ is of this type and (5.18) holds, we get
\[P(X \in xH) \leq 120|G_p|^{-\gamma}\]
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 [4]) at the level of the free group generated by $S$, although it is also possible to keep attention...
focused to the finite groups $\text{SL}_2(F_p)$ (indeed, for groups which are more complicated than $\text{SL}_2$, there exist proper “complicated” subgroups, and this second option is then most natural.) Precisely, we have the following ad-hoc lemma:

**Proposition 5.4.3.** Let $k \geq 2$ 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

\[(5.20) \quad [[x_1, x_2], [x_3, x_4]] = 1\]

for all $(x_1, x_2, x_3, x_4) \in W$. Then for any $m \geq 1$, we have

$$|\{x \in W \mid d_T(1, x) \leq m\}| \leq (4m + 1)(8m + 1) \leq 45m^2,$$

where $T$ is the $|S|$-regular tree $\mathfrak{C}(F_k, S)$, $S = \{a_i^{\pm 1}\}$.

**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 $y \in C_{F_k}(x)$, which is an infinite cyclic group by Proposition B.1.1, (3). Denoting a generator by $z$, we find

\[(5.21) \quad |\{y \in B_1(m) \mid [x, y] = 1\}| = |\{h \in Z \mid d_T(1, z^h) \leq m\}| \leq 2m + 1\]

since (Proposition B.1.1, (5)), we have $d_T(1, z^h) \geq |h|$.

Now we come back to a set $W$ verifying the assumption (5.20), which we assume to be not reduced to 1, and we denote $W_m = W \cap B_1(m)$, 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 $W_m \subset C_{F_k}(x) \cap B_1(m)$, and (5.21) 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 $d_T(1, [x_0, y]) \leq 4m$, 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 $\varphi(yy_1^{-1}) = yyy_1^{-1}$, where $\varphi(y) = x_0yx_0^{-1}$ denotes the inner automorphism of conjugation by $x_0$. Hence $y_1$ satisfies $[x_0, y_1] = b$ if and only if $\varphi(yy_1^{-1}) = yyy_1^{-1}$, 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

$$|\{y_1 \in B_1(m) \mid [x_0, y_1] = [x_0, y]\}| = |\{h \in Z \mid yz^h \in B_1(m)\}| \leq 4m + 1,$$

since

$$d_T(1, z^h) = d_T(y, yz^h) \leq d_T(1, y) + d_T(1, yz^h) \leq 2m$$

for $h \in Z$ such that $yz^h$ is in $B_1(m)$.

Hence we have $|W_m| \leq (4m + 1)(8m + 1)$ in that case, which proves the result. \qed

Using Corollary 5.4.2, we finally verify fully Condition (3) in Corollary 5.3.17:

**Corollary 5.4.4 (Decay of probabilities, II).** Let $S \subset \text{SL}_2(\mathbb{Z})$ 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 $G_p = \text{SL}_2(F_p)$, and let $(X_n)$ be the random walk on $\mathfrak{C}(G_p, S_p)$ with $X_0 = 1$.

There exist $c > 0$ and $\gamma > 0$ such that for any prime $p$ large enough, any $x \in \text{SL}_2(F_p)$ and any proper subgroup $H \subset \text{SL}_2(F_p)$, we have

\[(5.22) \quad P(X_n \in xH) \leq |G_p|^{-\gamma}\]

where

$$n \leq c \log |G_p|.$$
Proof. We start by noting that, for all $x \in G_p$ and $H \subset G_p$, we have
\[
P(X_{2n} \in H) = \sum_{g \in G_p} P(X_n = g \text{ and } X_n^{-1}X_{2n} \in H) \\
\geq \sum_{h \in H} P(X_n = xh) P(h^{-1}x^{-1}X_n \in H)
\]
because $X_n^{-1}X_{2n}$ is independent of, and has the same distribution as, $X_n$ (Proposition 3.2.8). Since
\[
P(h^{-1}x^{-1}X_n \in H) = P(X_n \in xH),
\]
this gives
\[
P(X_{2n} \in H) \geq P(X_n \in xH) \sum_{h \in H} P(X_n = xh) = P(X_n \in xH)^2,
\]
which means that it is enough to give an upper bound for $P(X_{2n} \in H)$ to get one for $P(X_n \in xH)$.

Consider first the case where $H$ is “big”, but (5.19) holds for $H$. Let $\tilde{H} \subset G$ be the pre-image of $H$ under reduction modulo $p$. If $2n \leq \tau \log(p/2)$, then as in the proof of Corollary 5.4.2, we get
\[
P(X_{2n} \in H) = P(\tilde{X}_{2n} \in \tilde{H}).
\]
Provided $n$ also satisfies the stronger condition $n \leq m = \frac{1}{16} \tau \log(p/2)$, any commutator
\[
[[x_1, x_2], [x_3, x_4]]
\]
with $x_i \in \tilde{H} \cap B_1(n)$ is an element at distance at most $\tau \log(p/2)$ from 1 in the tree $C(G, S)$, which reduces to the identity modulo $p$ by (5.19), and therefore must be itself equal to 1. In other words, we can apply Proposition 5.4.3 to $W = \tilde{H} \cap B_1(m)$ to deduce the upper bound
\[
|\tilde{H} \cap B_1(m)| \leq 45m^2.
\]
We now take
\[
n = \frac{1}{32} \lfloor \tau \log(p/2) \rfloor,
\]
and we derive
\[
P(X_{2n} \in H) \leq |\tilde{H} \cap B_1(m)| r^{-2n} \leq 45m^2 |G_p|^{-\gamma/16}
\]
where $\gamma$ is as in Corollary 5.4.2, and hence
\[
P(X_n \in xH) \leq |G_p|^{-\gamma/32}
\]
provided $p$ is large enough, which gives the conclusion in that case.

On the other hand, if $H$ is “small”, i.e., if $|H| \leq 120$, then for the same value of $n$ we get
\[
P(X_n \in xH) \leq 120 |G_p|^{-\gamma}
\]
by Corollary 5.4.2, and this gives again the desired result for $p$ large enough. \hfill \Box

We can now summarize what we have obtained concerning the Bourgain-Gamburd criterion (Corollary 5.3.17) in the situation of Corollary 5.4.4 for $G_p = SL_2(F_p)$. This will finally prove Theorem 5.1.1.

(1) We have
\[
d(G_p) = \frac{p-1}{2}
\]
for $p \geq 3$. In particular, $d(G_p) \geq |G_p|^d$ for any $d < 1/3$ provided $p$ is large enough.
(2) The next section will show that these groups are $\delta$-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

$$P(X_{2k} \in xH) \leq |G|^{-\gamma}$$

when

$$k \leq c \log |G_p|$$

for some $c > 0$, $\gamma > 0$ and $p$ large enough.

We conclude that, if $S \subset \text{SL}_2(\mathbb{Z})$ freely generates a free subgroup of rank $\geq 2$ and $\gamma$ is defined as above, then for $c$ defined by $c = (c,d,\delta,\gamma)$ for any $d < 1/3$, the family $\mathcal{G}_2(c)$ contains all but finitely many Cayley graphs of $\text{SL}_2(\mathbb{F}_p)$ with respect to $S$ modulo $p$. In particular, we then deduce from Corollary 5.3.17 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 5.1.1 when $S$ generates a free group.

We can now finally explain the classical tool which is used to reduce the full statement of Theorem 5.1.1 to the case where the given symmetric subset $S \subset \text{SL}_2(\mathbb{Z})$ generates a free group.

**Lemma 5.4.5.** Let $S \subset \text{SL}_2(\mathbb{Z})$ be any finite symmetric subset and let $G$ be the subgroup generated by $S$. If the reduction of $S$ modulo $p$ generates $\text{SL}_2(\mathbb{F}_p)$ 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 $\text{SL}_2(\mathbb{F}_p)$.

**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

$$\tilde{G} = G \cap \Gamma(2),$$

where

$$\Gamma(2) = \ker(\text{SL}_2(\mathbb{Z}) \to \text{SL}_2(\mathbb{Z}/2\mathbb{Z})).$$

By Proposition B.1.3, (3), the subgroup $\Gamma(2)$ of $\text{SL}_2(\mathbb{Z})$ is a free group of rank 2, hence the intersection $\tilde{G}$ is a free group (as a subgroup of the free group $G_2$, see Proposition B.1.1, (1)). It can not be of rank 1 (or 0) because it is of finite index in $G$, and hence (by the assumption on $G$) still surjects to $\text{SL}_2(\mathbb{F}_p)$ 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 $\tilde{G}$, and the subgroup $G_1$ they generate.

Using this, for a given $S \subset \text{SL}_2(\mathbb{Z})$, 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 $\text{SL}_2(\mathbb{F}_p)$ with respect to $S$ and to $S_1$.

For $p$ large enough so that $G_p = \text{SL}_2(\mathbb{F}_p)$ is generated both by $S$ modulo $p$ and $S_1$ modulo $p$, we have

$$d(x, y) \leq C d_1(x, y)$$

where $d_1(\cdot, \cdot)$ is the distance in the Cayley graph $\Gamma_1 = \mathcal{C}(G_p, S_1 (\text{mod } p))$, and $d(\cdot, \cdot)$ the distance in $\Gamma_2 = \mathcal{C}(G_p, S (\text{mod } p))$ and $C$ is the maximum of the word length of $s_1$, $s_2$. 

105
with respect to $S$. Hence, by Lemma 3.1.17 (applied to $\Gamma_1$ and $\Gamma_2$ with $f$ the identity), the expansion constants satisfy

$$h(\mathcal{C}(G_p, S \mod p)) = h(\Gamma_2) \geq w^{-1}h(\mathcal{C}(G_p, S_1 \mod p))$$

with

$$w = 4 \sum_{j=1}^{\lfloor C \rfloor} |S|^j.$$

In particular, if Theorem 5.1.1 holds for $G_1$, it will also hold for $G$.

**Example 5.4.6 (The Lubotzky group).** The group $L$ generated by

$$S = \left\{ \begin{pmatrix} 1 & \pm 3 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ \pm 3 & 1 \end{pmatrix} \right\} \subset SL_2(\mathbb{Z})$$

is a free group on 2 generators (see Proposition B.1.3 in Appendix B). For all $p \neq 3$, the reduction of $S$ modulo $p$ generates $SL_2(\mathbb{F}_p)$, by Proposition B.2.1, (2).

## 5.5. Quasi-random groups

Our next goal is the prove that the groups $SL_2(\mathbb{F}_p)$ satisfy Condition (2) of the Bourgain-Gamburd criterion: they are $\delta$-flourishing for some $\delta > 0$ independent of $p$. The first step, however, will be to prove a weaker property (which would also follow from $\delta$-flourishing) which turns out to be more accessible and useful to simplify the final arguments. This result is due to Gowers and Nikolov-Pyber:

**Theorem 5.5.1 (Gowers; Nikolov-Pyber).** Let $p \geq 3$ be prime, and let $H \subset SL_2(\mathbb{F}_p)$ be an arbitrary subset such that

$$|H| \geq 2|SL_2(\mathbb{F}_p)|^{8/9}.$$ Then we have $H \cdot H \cdot H = SL_2(\mathbb{F}_p)$.

This should be compared with Exercise 5.3.13, (2), which shows that if $SL_2(\mathbb{F}_p)$ is to be $\delta$-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 5.5.2.** Let $G$ be a finite group, and let $A, B, C \subset G$ be any subsets of $G$ such that

$$\frac{|A||B||C|}{|G|^3} \geq \frac{1}{d(G)}.$$ Then we have $A \cdot B \cdot C = G$.

In particular, if $|A| \geq (|G|/d(G))^{1/3}$, we have $A \cdot A \cdot A = G$.

Applying the theorem of Frobenius (Theorem 5.2.7), we see that Theorem 5.5.1 is a corollary of this general fact.

In some sense, Theorem 5.5.2 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 5.5.2 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 5.5.3 (Gowers).** Let $G$ be a finite group, and let $A$, $B$, $C \subset G$ be any subsets of $G$ such that

$$\frac{|A||B||C|}{|G|^3} \geq \frac{1}{d(G)}.$$

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

$$\frac{|A_1||B_1||C_1|}{|G|^3} = \frac{|A||B||C|}{|G|^3} \geq \frac{1}{d(G)},$$

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 $x = abc \in A \cdot B \cdot C$.

**Proof of Proposition 5.5.3.** The idea is to consider the function defined by

$$\varphi(x) = \frac{|C \cap gB|}{|G|} = \mu(C \cap gB),$$

(where $\mu$ is the uniform probability measure on $G$) and to show that it is non-zero on a “large” set by estimating its variance

$$V = \frac{1}{|G|} \sum_{g \in G} \left( \varphi(g) - \langle \varphi, 1 \rangle \right)^2.$$

Indeed, by positivity, the set $X = \{x \mid \varphi(x) = 0\}$ satisfies

$$\frac{|X|}{|G|} \langle (\varphi, 1) \rangle^2 \leq V,$$

hence any set $A$ which is large enough in the sense that

$$(5.23) \quad \mu(A) > \frac{V}{\langle (\varphi, 1) \rangle^2}$$

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 \( \delta_C \) 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 \( \varphi_0 = \varphi - \langle \varphi, 1 \rangle = T\tilde{\delta}_C \) where

\[
\tilde{\delta}_C = \delta_C - \mu(C),
\]

so that \( \tilde{\delta}_C \) is in the subspace \( L^2_0(G) \) of functions of average 0. We note that \( T \) acts on \( L^2_0(G) \). We see then that we can write

\[
V = \| \varphi_0 \|^2 = \| T\tilde{\delta}_C \|^2 = \langle T^*T\tilde{\delta}_C, \tilde{\delta}_C \rangle \leq \lambda^2 \| \tilde{\delta}_C \|^2,
\]

where \( \lambda^2 \geq 0 \) is the largest eigenvalue of the self-adjoint (non-negative) operator \( T^*T \) acting on \( L^2_0(G) \). The norm of \( \tilde{\delta}_C \) is easy to compute: we have

\[
\| \tilde{\delta}_C \|^2 \leq \| \delta_C \|^2 = \mu(C)(1 - \mu(C)) < \mu(C).
\]

We are therefore reduced to estimating from above the eigenvalue \( \lambda^2 \). This allows us to bring in representation theory, because \( T \), and \( T_2 = T^*T \), commute with the left-regular representation \( \text{reg} \) defined in Proposition 5.2.5, i.e., we have

\[
\text{reg}(x)(T_2\psi) = T_2(\text{reg}(x)\psi)
\]

for all \( \psi \in L^2(G) \) and \( x \in G \). As a consequence (and exactly as in Proposition 5.2.5, (2)), the \( \lambda^2 \)-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 5.2.6, we now use this multiplicity and positivity to deduce that

\[
\lambda^2 d(G) \leq \text{Tr}(T_2) = \frac{1}{|G|^2} \sum_{x,y \in B} \text{Tr}(\text{reg}'(x^{-1}y)) = \mu(B),
\]

where \( \text{reg}'(g) \) is the right-regular representation operator

\[
\text{reg}'(g)\psi(x) = \psi(xg),
\]

which has trace 0 if \( g \neq 1 \), and \( |G| \) otherwise.\(^2\) We have therefore obtained

\[
V \leq \lambda^2 \| \tilde{\delta}_C \|^2 < \frac{\mu(B)\mu(C)}{d(G)},
\]

and the condition (5.23) which ensures that \( C \cap AB \neq \emptyset \) is implied by

\[
\mu(A) \geq \frac{\mu(B)\mu(C)}{d(G)} \frac{1}{\langle \varphi, 1 \rangle^2} = \frac{1}{\mu(B)\mu(C)d(G)},
\]

which is the statement of Proposition 5.5.3. \(\square\)

\(^2\) Compute the trace using the basis of \( L^2(G) \) of characteristic functions.
5.6. Growth of generating subsets of $\text{SL}_2(\mathbb{F}_p)$

Finally, in this section, we will prove Helfgott’s growth theorem [20]:

**Theorem 5.6.1 (Helfgott).** There exists a constant $\delta > 0$ such that for any prime $p$ and any subset $H \subset \text{SL}_2(\mathbb{F}_p)$ which generates $\text{SL}_2(\mathbb{F}_p)$, we have either

$$|H \cdot H \cdot H| \geq |H|^{1+\delta},$$

or

$$H \cdot H \cdot H = \text{SL}_2(\mathbb{F}_p).$$

**Remark 5.6.2.** (1) In [28, Th. 1.2], we show that one can take $\delta = 1/3024$.

(2) Exercise 5.6.6 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 $H \subset \text{SL}_2(\mathbb{F}_p)$ “grows” significantly under product, in the sense that

$$\text{trp}(H) \geq |H|^\delta,$$

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 $\text{SL}_2(\mathbb{F}_p)$.

The argument presented here is basically the one sketched by Pyber and Szabó in [40, §1.1] (which is expanded in their paper to cover much more general situations). It is closely related to the one of Breuillard, Green and Tao [6], and many ingredients are already visible in Helfgott’s original argument [20].

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 5.6.1 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, or it is already so large that the triple product is all of $\text{SL}_2(\mathbb{F}_p)$.

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 5.6.3 (Non-concentration example).** Let $p$ be a prime number and let $U = \left\{ \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \mid t \in \mathbb{F}_p \right\} \subset \text{SL}_2(\mathbb{F}_p)$.

For any symmetric generating set $H$ of $\text{SL}_2(\mathbb{F}_p)$, we have

$$|H \cap U| \leq 2 |H^{(5)}|^{1/3}.$$
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

$$|H^{(5)}| \geq \frac{1}{2} p^3,$$

which shows that such a set is, in fact, rather fast-growing! (Since $|H^{(5)}| > \frac{1}{2} |SL_2(F_p)|$, the ten-fold product set will be all of $SL_2(F_p)$).

**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 because $H$ generates $SL_2(F_p)$, it can not be entirely contained in the subgroup

$$B = \left\{ \begin{pmatrix} a & b \\ 0 & a^{-1} \end{pmatrix} \mid a \in F_p^*, b \in F_p \right\}$$

of upper-triangular matrices. We fix then some $h \in H$ which is not in $B$, say

$$h = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with $c \neq 0$. Since $H$ is symmetric, we also have $h^{-1} \in H$.

Now consider $U^* = U - 1$, and define a map

$$\psi : \left\{ \begin{pmatrix} u_1 & u_2 \\ 0 & 1 \end{pmatrix} \right\} \times \begin{pmatrix} u_1 & u_2 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} u_1 & u_2 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} u_1 & u_2 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} u_1 & u_2 \\ 0 & 1 \end{pmatrix} \to SL_2(F_p)$$

$$u_1 u_2 h h^{-1} u_3$$

If we restrict $\psi$ to $(U^* \cap H)^3$, we see that $\psi((U^* \cap H))^3 \subset H \cdot H \cdot H \cdot H \cdot H$. So we can estimate the size of $U^* \cap H$ by summing according to the values of $\psi$, 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 $x \in SL_2(F_p)$, the inverse image $\psi^{-1}(x)$ is either empty or a single point (note that this is intuitively reasonable because the size of the domain of $\psi$ is about the same as the size of $SL_2(F_p)$.) Using this, we get

$$|U^* \cap H|^3 \leq |\psi((U^* \cap H)^3)| \leq |H^{(5)}|,$$

and then we add again the element $1 \in U \cap H$ to get

$$|U \cap H| = 1 + |U^* \cap H| \leq 1 + |H^{(5)}|^{1/3} \leq 2 \cdot |H^{(5)}|^{1/3},$$

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

$$\psi(u_1, u_2, u_3) = \begin{pmatrix} 1 - t_1 t_2 c^2 - t_2 a c & * \\ -t_2 c^2 & * \end{pmatrix},$$

and in order for this to be a fixed matrix $x \in SL_2(F_p)$, 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 $\psi$ using $U^*$ in order to ensure this...) Thus $t_1$ (i.e. $u_1$) is also uniquely determined, and finally

$$u_3 = (u_1 h u_2 h^{-1})^{-1} x$$
We now start the proof of Helfgott’s Theorem, following [40]. The point of view towards SL₂(𝔽ₚ) is that, for the most part, it consists of elements which are diagonalizable with distinct eigenvalues (though not necessarily with eigenvalues in the field 𝔽ₚ itself). Such elements produce a certain amount of extra structure: they come in “packages”, 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 “packages”.

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 5.6.3 would suggest strongly that we look for such a relation, since we obtain the growth of the five-fold product set, not of 𝐻 · 𝐻 · 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 5.6.4.** Let 𝐻 be a subset of a group 𝐺, and let 𝑛 ⩾ 0 be an integer. We define the 𝑛-fold symmetric product set

\[ 𝐻^{(𝑛)} = \{ x ∈ 𝐺 \mid x = a_1 \cdots a_𝑛 \text{ for some } a_i ∈ 𝐻 ∪ 𝐻^{-1} ∪ \{1\} \} \]

Note the immediate relations

\[ (𝐻^{(𝑛)})^{-1} = 𝐻^{(𝑛)}, \quad (𝐻^{(𝑛)})^{(𝑚)} = 𝐻^{(𝑛𝑚)}, \quad 𝐻^{(𝑛+𝑚)} = 𝐻^{(𝑛)} · 𝐻^{(𝑚)} \]

for 𝑛, 𝑚 ⩾ 0.

**Proposition 5.6.5 (Ruzsa).** Let 𝐺 be a finite group, and let 𝐻 ⊂ 𝐺 be a symmetric non-empty subset.

1. Denoting \( α_n = |H^{(n)}|/|H| \), we have

\[ α_n \leq α_{n-2} = \text{trp}(H)^{n-2} \]

for all \( n \geq 3 \).

2. We have \( \text{trp}(H^{(2)}) \leq \text{trp}(H)^4 \) and for \( k \geq 3 \), we have

\[ \text{trp}(H^{(k)}) \leq \text{trp}(H)^{3k-3}. \]

**Proof.** The first part is proved by induction on \( n \geq 3 \), with the initial case \( n = 3 \) being tautological. For the induction step, assuming (5.25) for some \( n \geq 3 \), we use the triangle inequality for the Ruzsa distance

\[ d(A, B) = \log \left( \frac{|A \cdot B^{-1}|}{\sqrt{|A||B|}} \right) \leq d(A, C) + d(C, B) \]

between non-empty subsets of 𝐺 (Lemma A.1.2 in Appendix A). We denote \( h_n = |H^{(n)}| \) for simplicity, and then write

\[ α_{n+1} = \frac{h_{n+1}}{h_1} = \frac{|H^{(n-1)} \cdot H^{(2)}|}{h_1}, \]

is uniquely determined, if it exists...
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

$$
\alpha_{n+1} = h_1^{-1} h_{n-1}^{1/2} h_2^{1/2} \exp(d(H^{(n-1)}, H^{(2)})) \\
\leq h_1^{-1} h_{n-1}^{1/2} h_2^{1/2} \exp(d(H^{(n-1)}, H^{(1)}) + d(H^{(1)}, H^{(2)})) \\
= h_1^{-1} h_{n-1}^{1/2} h_2^{1/2} (h_n h_{n-1}^{1/2} h_1^{1/2} (h_3 h_1^{1/2} h_2^{-1/2})) \\
= \frac{h_{n-1} h_3}{h_1} = \alpha_n \alpha_3 \leq \alpha_3^{n-1},
$$

using the induction assumption, and completing the proof of (5.25).

For (2), we have

$$
\text{trp}(H^{(k)}) = h_3 k = \frac{\alpha_3 k}{\alpha_k}.
$$

Since $\alpha_k \geq \alpha_3$ for $k \geq 3$, we obtain $\text{trp}(H^{(k)}) \leq \alpha_3^{k-3}$ for $k \geq 3$ by (1), while for $k \geq 2$, we simply use $\alpha_2 \geq 1$ to get $\text{trp}(H^{(2)}) \leq \alpha_3^4$. \hfill $\square$

**EXERCISE 5.6.6.** 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, $G = \text{SL}_2(\mathbb{F}_p)$ and $B$ the subgroup of upper-triangular matrices. Denote

$$
w = \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix} \not\in B,
$$

and define $H = B \cup \{w, w^{-1}\}$.

(1) Show that there exists no $A \geq 0$ such that

$$
\text{trp}(H) \leq \left(\frac{|H^{(2)}|}{|H|}\right)^A
$$

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 to show that Helfgott’s Theorem holds when $|H|$ is small, in the following sense:

**LEMMA 5.6.7 (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 $|H^{(3)}| \geq 2^{1/2} |H|$.

**PROOF.** The argument is in fact a bit similar to that in Proposition 5.6.3. 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

$$
i : \begin{cases} 
H \\
h
\end{cases} \rightarrow \begin{cases} 
G \\
hx
\end{cases}.
$$

The image of this map is contained in $H^{(4)}$ and it is disjoint with $H$ since $x \not\in H^{(2)}$. Hence $H^{(4)}$, which contains $H$ and the image of $i$, satisfies

$$
|H^{(4)}| \geq 2|H|.
$$

By Ruzsa’s Lemma, we obtain

$$
\text{trp}(H) \geq \left(\frac{|H^{(4)}|}{|H|}\right)^{1/2} \geq 2^{1/2}.
$$

112
This lemma and Proposition 5.6.5, as well as Theorem 5.5.1, show that in order to prove Helfgott’s Theorem, it is enough to exhibit a real number \( x \geq 2 \), an integer \( m \geq 3 \) and \( \delta > 0 \), all being absolute constant, such that for all primes \( p > x \) and all symmetric generating subsets \( H \subset SL_2(\mathbb{F}_p) \), we have either

\[
|H| \geq 2|SL_2(\mathbb{F}_p)|^{8/9},
\]

or

\[
|H^{(m)}| \geq |H|^{1+\delta}.
\]

Indeed, if \( p \leq x \), the exercise gives the result and otherwise we get \( H^{(3)} = SL_2(\mathbb{F}_p) \) in the first case, by Theorem 5.5.1, and

\[
trp(H) = \frac{|H^{(3)}|}{|H|} \geq \left( \frac{|H^{(m)}|}{|H|} \right)^{\frac{1}{m-2}} \geq |H|^{\delta/(m-2)}
\]

in the second, by Ruzsa’s Lemma.

We now introduce the crucial definitions for the Pyber-Szabó argument. 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-adhoc proof) is to introduce the infinite groups

\[
SL_2(\mathbb{F}_p)
\]

in addition to \( SL_2(\mathbb{F}_p) \). We will write \( G_p \) (or simply \( G \)) for the infinite group, and \( G_p \) (or \( G \)) for the finite group. We recall that there is a Frobenius automorphism \( \sigma \) defined on \( G_p \) by

\[
x^\sigma = \begin{pmatrix} a^p & b^p \\ c^p & d^p \end{pmatrix}
\]

for \( x = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \), which has the property that

\[
G_p = \{ x \in G_p \mid x^\sigma = x \}.
\]

**Definition 5.6.8 (Regular semisimple elements; maximal tori).** Fix a prime number \( p \) and let \( G = SL_2(\mathbb{F}_p) \), \( G = SL_2(\overline{\mathbb{F}}_p) \).

(1) An element \( x \in G \) is semisimple if it is diagonalizable (in some basis). It is regular semisimple if in addition the eigenvalues \( \alpha, \beta \) of \( x \), which are elements of \( \overline{\mathbb{F}}_p \), are distinct. For any subset \( H \subset G \), we write \( H_{\text{reg}} \) for the set of the regular semisimple elements in \( H \).

(2) A maximal torus \( T \) in \( G \) is the centralizer of a regular semisimple element \( x \), i.e., a subgroup of the form

\[
T = C_G(x)
\]

for some \( x \in G_{\text{reg}} \). A maximal torus \( T \) in \( G \) is a subgroup of the form \( T = T \cap G \), where \( T \subset G \) is a maximal torus of the infinite group \( G \) which is \( \sigma \)-invariant, i.e., such that \( x^\sigma \in T \) for all \( x \in T \).

Here are the basic properties of regular semisimple elements and their centralizers:

**Proposition 5.6.9.** As above, fix a prime number \( p \) and let \( G = SL_2(\mathbb{F}_p) \), \( G = SL_2(\overline{\mathbb{F}}_p) \).

(1) A regular semisimple element \( x \in G \) is contained in a unique maximal torus \( T \), namely its centralizer \( T = C_G(x) \). In particular, if \( T_1 \neq T_2 \) are two maximal tori, we have

\[
T_{1,\text{reg}} \cap T_{2,\text{reg}} = \emptyset.
\]
(2) For any maximal torus $T \subset G$ and any $x \in G$, the subgroup $xTx^{-1}$ is a maximal torus, and similarly if $T \subset G$ is a maximal torus and $x \in G$, the conjugate $xTx^{-1}$ is a maximal torus of $G$.

(3) If $T \subset G$ is a maximal torus, we have $|T_{nreg}| = |T - T_{reg}| = 2$.

(4) For any maximal torus $T$, the normalizer $N_G(T)$ contains $T$ as a subgroup of index 2. Similarly, for any maximal torus $T \subset G$, $N_G(T)$ contains $T$ as a subgroup of index 2, and in particular $2(p - 1) \leq |N_G(T)| \leq 2(p + 1)$.

(5) The conjugacy class $Cl(g)$ of a regular semisimple element $g \in G$ is the set of all $x \in G$ such that $Tr(x) = Tr(g)$. The set of elements in $G$ which are not regular semisimple is the set of all $x \in G$ such that $Tr(x)^2 = 4$.

To a large extent, this is “just” linear algebra, but we provide the proof. Many readers will probably want to skip it.

**Proof.** (4): this is done, in the case of $T$, in Remark B.2.3 in Appendix B.

Finally, (a variant of) the following concept was introduced under different names and guises by Helfgott, Pyber-Szabó, and Breuillard-Green-Tao. We chose the name from the last team.

**Definition 5.6.10 (A set involved with a torus).** Let $p$ be a prime number, $H \subset SL_2(F_p)$ a finite set and $T \subset SL_2(F_p)$ a maximal torus. Then $H$ is involved with $T$, or $T$ with $H$, if and only if $T$ is $\sigma$-invariant and $H$ contains a regular semisimple element of $T$ with non-zero trace, i.e., $H \cap T_{sreg} \neq \emptyset$ where the superscript “$sreg$” restricts to regular semisimple elements with non-zero trace.

**Remark 5.6.11.** The twist in this definition, compared with the one in [40] or [6], is that 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 5.6.14), and lead to a shorter proof, with better explicit constants. However, this restriction is not really essential in the greater scheme of things, and it would probably not be a good idea to do something similar for more general groups.

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 insist on finding some which have 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 $p \geq 7$ (though using any $H^{(k)}$, with $k$ fixed, would not compromise the proof, except for the values of the constants.)

**Lemma 5.6.12 (Helfgott).** Let $p \geq 7$ be a prime number and let $H \subset SL_2(F_p)$ 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. We will state a general version in Section 5.7, but since we only need this special case, we 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 $N = G - G_{reg}$ of elements which are not regular semisimple is invariant under conjugation, and (as observed in Proposition 5.6.9) is the set of all matrices with trace equal to 2 or $-2$. It is precisely the union of the two central elements $\pm 1$ 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 $\varepsilon \in \mathbb{F}_p^\times$ 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.2.5 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 $H^{(3)}_{reg}$ is empty and $p \geq 7$, and will derive a contradiction.

We distinguish two cases. In the first case, we assume that $H$ contains one element of trace $\pm 2$ which is not $\pm 1$. 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 $SL_2(\mathbb{F}_p)$. Then $H^{(3)}$ contains $ug$, $u^2g$, $u^{-1}g$, $u^{-2}g$, which have traces, respectively, equal to $\text{Tr}(g) + c$, $\text{Tr}(g) + 2c$, $\text{Tr}(g) - c$, $\text{Tr}(g) - 2c$. 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 $(u')^jg$ for $j \in \{-2, -1, 0, 1, 2\}$ is

$$\{\text{Tr}(g) + 2c, -\text{Tr}(g) - c, \text{Tr}(g), -\text{Tr}(g) + c, \text{Tr}(g) - 2c\},$$

and one can check that for $p \geq 5$, one of these is not 0, $-2$ or 2, although some could coincide (for instance, if $\text{Tr}(g) = 2$, 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 $p \geq 5$).

In the second case, all elements of $H$ except $\pm 1$ have trace 0. We split in two subcases, but depending on properties of $\mathbb{F}_p$. 

115
The first one is when $−1$ is not a square in $\mathbb{F}_p$. Conjugating again, we can assume that $g_0 ∈ H$. Because $H$ generates $\text{SL}_2(\mathbb{F}_p)$, we claim that there must exist a matrix
\[
g = \begin{pmatrix} a & b \\ c & −a \end{pmatrix}
\]
in $H$ with (i) $a ≠ 0$; (ii) $b ≠ c$. Indeed if all elements $≠ ±1$ of $H$ are of the form
\[
g = \begin{pmatrix} 0 & −c^{-1} \\ c & 0 \end{pmatrix},
\]
we can find such an element with $c ≠ ±1$ (i.e., $g ≠ ±g_0$), since otherwise $H$ is not a generating set; then the trace of $g_0g$ is $c + c^{-1}$, which is not in $\{-2, 0, 2\}$ (non-zero because $−1$ is not a square in our first subcase), so $H^{(2)}_{\text{reg}} ≠$, which we excluded. So all elements of $H$, except for $±1$ and $±g_0$ are of the type
\[
g = \begin{pmatrix} a & b \\ c & −a \end{pmatrix},
\]
with $a ≠ 0$. 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 $ε = −1$ (see (B.2)), again contradicting the assumption that $H$ is a generating set.

Now we argue with $g$ as above. We have
\[
g_0g = \begin{pmatrix} c & −a \\ −a & −b \end{pmatrix} ∈ H^{(2)},
\]
with non-zero trace $t = c−b$. Moreover, if $t = 2$, i.e., $c = b+2$, the condition $\det(g_0g) = 1$ 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 $a ≠ 0$, it follows in both cases that $−1$ is a square in $\mathbb{F}_p$, which contradicts our assumption in the first subcase.

Now we come to the second subcase when $−1 = z^2$ is a square in $\mathbb{F}_p$. We can then diagonalize $g_0$ over $\mathbb{F}_p$, 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 $g_0g’ ∈ H^{(2)}$ is $2za$. But we can find $g’$ with $a ≠ 0$, since otherwise $H$ would again not be a generating set, being contained in the normalizer of a split maximal torus, (B.1), and so this trace is non-zero.

The condition $2za = ±2$ would give $za = ±1$, 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 $b ≠ 0$, 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 \( g_0g'g'' \) gives
\[
bc - 2, \quad bcz
\]
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… \( \square \)

**Example 5.6.13.** If \( p = 5 \), one can check (e.g., with MAGMA [34]) that the set
\[
H = \left\{ 1, \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} \pm 1, \begin{pmatrix} 2 & 2 \\ 0 & -2 \end{pmatrix} \pm 1, \begin{pmatrix} 2 & 0 \\ 2 & -2 \end{pmatrix} \pm 1 \right\} \subset \text{SL}_2(\mathbb{F}_5)
\]
is a generating set of \( \text{SL}_2(\mathbb{F}_5) \) 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 5.6.14 (Involving dichotomy).** (1) For all prime number \( p \), all subsets \( H \subset \text{SL}_2(\mathbb{F}_p) \) and all maximal tori \( T \subset \text{SL}_2(\bar{\mathbb{F}}_p) \), if \( T \) and \( H \) are not involved, we have
\[
|H \cap T| \leq 4.
\]
(2) There exists \( \delta > 0 \) such that if \( p \geq 3 \) and \( H \subset \text{SL}_2(\mathbb{F}_p) = G \) is a symmetric generating set containing \( 1 \), we have
\[
|T_{\text{reg}} \cap H^{(2)}| \gg \alpha^{-4}|H|^{1/3}
\]
for any maximal torus \( T \subset \text{SL}_2(\mathbb{F}_p) \) which is involved with \( T \), where \( \alpha = \text{trp}(H) \), unless
\[
\alpha \gg |H|^{\delta},
\]
where the implied constants are absolute.

Here (1) is obvious, since \( |T - T_{\text{reg}}| \leq 2 \) (part (3) in Proposition 5.6.9) and there are also at most two elements of trace \( 0 \) in \( T \) (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 (\( \sigma \)-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 (5.27) should be interpreted as
\[
|T_{\text{reg}} \cap H^{(k)}| \gg \alpha^{-m}|H|^{1/3}
\]
for some \( k \geq 1 \) and \( m \geq 0 \) 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 \( \delta \) 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 5.6.1.** We can assume that \( p \geq 7 \), which means that Lemma 5.6.12 is applicable. We will show that for some \( \delta > 0 \), we have
\[
\text{trp}(H) \gg |H|^{\delta}
\]
with an absolute implied constant, unless \( H^{(3)} = \text{SL}_2(F_p) \). Then using Lemma 5.6.7, 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 \( H_1 = H^{(2)} \), so that (by Lemma 5.6.12) there exists at least one maximal torus \( T \) involved with \( L = H_1^{(2)} = H^{(4)} \).

If, among all maximal tori involved with \( L \), none satisfies (5.27), we obtain directly from Proposition 5.6.14 (applied to \( H_1 \) instead of \( H \)) the lower bound
\[
\text{trp}(H_1) \gg |H_1|^\delta \gg |H|^\delta,
\]
and since \( \text{trp}(H_1) \leq \alpha^4 \) by Ruzsa’s Lemma, we get
\[
\alpha \gg |H|^{\delta/4}
\]
which gives (5.29) after renaming \( \delta \).

Otherwise, we distinguish two cases.

Case (1). There exists a maximal torus \( T \) involved with \( L \) such that, for any \( g \in G \), the torus \( gTg^{-1} \) is involved with \( L \).

As we can guess from (5.27) and (5.26), 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 \( T = T \cap G \), we note that the maximal tori \( gTg^{-1} = (gTg^{-1}) \cap G \) are distinct for \( g \) taken among representatives of \( G/N_G(T) \). Then we have the inequalities
\[
|L^{(2)}| \geq \sum_{g \in G/N_G(T)} |L^{(2)} \cap gT_{\text{reg}}g^{-1}| \gg |G|/|N_G(T)|
\]
where \( \beta = \text{trp}(L) \), since each \( gTg^{-1} \) is involved with \( L \) and distinct regular semisimple elements lie in distinct maximal tori (and we are in a case where (5.27) holds for all tori involved with \( L \)).

Now we unwind this inequality in terms of \( H \) and \( \alpha = \text{trp}(H) \). We have \( L^{(2)} = H^{(8)} \), so
\[
|H| \geq \alpha^{-6} |L^{(2)}| \gg \alpha^{-6} \beta^{-4} (p - 1)^2 |L|^{1/3} \gg \alpha^{-6} \beta^{-4} (p - 1)^2 |H|^{1/3}
\]
by Ruzsa’s Lemma. Furthermore, we have
\[
\beta = \text{trp}(L) = \text{trp}(H^{(4)}) \leq \alpha^{10}
\]
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
\[
|H| \geq 2 |G|^{8/9}
\]
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 \( T \) involved with \( L = H^{(4)} \) and a conjugate \( gTg^{-1} \), 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 5.6.15 below). There is a first clever observation (the idea of which goes back to work of Glibichuk and Konyagin [16] on the “sum-product phenomenon”): one can assume, possibly after changing \( T \) and \( g \), that \( g \) is in \( H \).
Indeed, to check this claim, we start with $T$ and $h$ as above. Since $H$ is a generating set, we can write
\[ g = h_1 \cdots h_m \]
for some $m \geq 1$ and some elements $h_i \in H$. Now let $i \leq m$ be the smallest index such that the maximal torus
\[ T' = (h_{i+1} \cdots h_m)T(h_{i+1} \cdots h_m)^{-1} \]
is involved with $L$. Taking $i = m$ means that $T$ is involved with $L$, which is the case, and therefore the index $i$ exists. Moreover $i \neq 0$, again by definition. It follows that
\[ (h_i h_{i+1} \cdots h_m)T(h_i h_{i+1} \cdots h_m)^{-1} \]
is not involved with $L$. But this means that we can replace $(T, g)$ with $(T', h_i)$, 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 $S = gTg^{-1} = hTh^{-1}$ 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 $H \cap S$ grows inside $S \cap G$.

**Lemma 5.6.15.** Let $G$ be a finite group, $K \subset G$ a subgroup, and $H \subset G$ an arbitrary symmetric subset. For any $n \geq 1$, we have
\[ \frac{|H^{(n+1)}|}{|H|} \geq \frac{|H^{(n)} \cap K|}{|H^{(2)} \cap K|}. \]

**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 $g_0^{-1} \in K \cap H^{(2)}$ if $g \in xK \cap H$, and hence
\[ |xK \cap H| \leq |K \cap H^{(2)}|. \]

This gives the lower bound
\[ |X| \geq \frac{|H|}{|K \cap H^{(2)}|}. \]

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
\[ |xK \cap H^{(n+1)}| \geq |K \cap H^{(n)}| \]
for any $xK \in X$, and (cosets being disjoint)
\[ |H^{(n+1)}| \geq |X||K \cap H^{(n)}|, \]
which gives the result when combined with the lower bound for $|X|$. \qed

Apply Lemma 5.6.15 with $(H, K) = (\tilde{H}, hTh^{-1} \cap G)$ and $n = 5$. This gives
\[ \frac{|	ilde{H}^{(6)}|}{|	ilde{H}|} \geq \frac{|	ilde{H}^{(5)} \cap S|}{|	ilde{H}^{(2)} \cap S|}. \]
But since $L = \tilde{H}(2)$ and $S$ are not involved (by construction), we have $|\tilde{H}(2) \cap S| \leq 2$, by the easy part of the Key Proposition 5.6.14, and therefore

$$\frac{|\tilde{H}(6)|}{|H|} \geq \frac{1}{2} |\tilde{H}(5) \cap S|.$$ 

However, $L$ and $T$ are involved, and moreover

$$h(H(8) \cap T)h^{-1} \subset H(10) \cap S = \tilde{H}(5) \cap S,$$

so that

$$|\tilde{H}(5) \cap S| \geq |H(8) \cap T| = |L(2) \cap T| \gg \tilde{\alpha}^{-4}|L|^{1/3}$$

where $\tilde{\alpha} = \text{trp}(L)$, by the Key Proposition 5.6.14 (again, because (5.27) holds for all tori involved with $L$).

Thus

$$\frac{|\tilde{H}(6)|}{|H|} \gg \tilde{\alpha}^{-4}|H|^{1/3},$$

which translates to

$$\alpha^{10}|H| \gg \alpha^{-36}|H|^{4/3},$$

by Ruzsa’s Lemma. This gives a fairly strong bound for $\alpha$, namely

$$\alpha = \text{trp}(H) \gg |H|^{1/138}. \quad (5.31)$$

To summarize, we have obtained three possible lower bounds of the right kind for $\alpha$, and one of them holds if $H(3) \neq \text{SL}_2(F_p)$. All imply (5.29), 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 $T_{\text{sreg}} \cap H(k)$ (for some fixed integer $k$) is in fact equivalent with a simpler-looking upper-bound involving, instead of $T$, the conjugacy class of a regular semisimple element for which $T$ is the centralizer. This is an “approximate” version of the orbit-stabilizer theorem in group theory.

**Proposition 5.6.16 (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

$$|K \cap H(2)| \geq \frac{|H|}{|H \cdot x|},$$

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 $H$

$$\phi : \begin{cases} H \rightarrow X \\ h \mapsto h \cdot x \end{cases}$$

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

$$|\phi^{-1}(y)| \leq |K \cap H(2)|$$

120
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

$$|H| \leq |H \cdot x||K \cap H^{(2)}|,$$

as claimed. \hfill \square

As a corollary, let $T = T \cap G$ be a maximal torus in $G$. Fixing any $g \in T_{\text{reg}}$, we have $T = C_G(g)$, the stabilizer of $g$ in $G$ for its conjugacy action on itself. We find therefore that

$$|T \cap H^{(2)}| \geq \frac{|H|}{|\{hgh^{-1} \mid h \in H\}|}$$

for any symmetric subset $H$. If we now assume that $T$ and $H$ are involved, we can select $g$ in $T_{\text{reg}} \cap H$ in this inequality, and the denominator on the right becomes

$$|\{hgh^{-1} \mid h \in H\}| \leq |H^{(3)} \cap \text{Cl}(g)| \leq |H^{(3)} \cap \text{Cl}(g)|$$

where $\text{Cl}(g)$ (resp. $\text{Cl}(g)$) is the conjugacy class of $g$ in $G$ (resp. in $G$). 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 5.6.17** (Larsen-Pink non-concentration inequality). Let $p \geq 3$ be a prime number and let $g \in \text{SL}_2(F_p) = G$. 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

$$|\text{Cl}(g) \cap H| \ll \alpha^{2/3}|H|^{2/3}$$

where $\alpha = \text{trp}(H)$ is the tripling constant of $H$, unless

$$\alpha \gg |H|^\delta.$$

The implied constants are absolute.

Applying this result to $H^{(3)}$, with tripling constant bounded by $\alpha^6$ (by Ruzsa’s Lemma), we obtain by (5.32) the lower bound

$$|T \cap H^{(2)}| \geq \frac{|H|}{|H^{(3)} \cap \text{Cl}(g)|} \gg \alpha^{-4}|H|^{1/3},$$

unless $\alpha = \text{trp}(H) \gg |H|^{\delta/6}$. Since there are at most two elements of $T \cap H^{(2)}$ which are not regular, the first bound gives

$$|T_{\text{reg}} \cap H^{(2)}| \gg \alpha^{-14/3}|H|^{1/3},$$

unless $\alpha^{-4}|H|^{1/3} \ll 1$. This means that Proposition 5.6.14 is proved.

**Remark 5.6.18.** The original Larsen-Pink inequalities proved in [29, §4] concern the intersection of a set like $\text{Cl}(g)$, defined by algebraic equations in an algebraic group (like $\text{SL}_n(k)$ where $k$ is an algebraically closed field), with a finite subgroup of $\text{SL}_n(k)$. It was observed first by Hrushovski [22] that these inequalities extend in a natural way to approximate groups. In Section 5.7 we will explain a general case in this setting.

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 [29, §4]. We wish to consider a map like

\[ \phi \left\{ \begin{array}{c} \text{Cl}(g) \times \text{Cl}(g) \times \text{Cl}(g) \\ (x_1, x_2, x_3) \end{array} \rightarrow \text{G} \times \text{G} \right. \]

\[ \left( x_1g_1x_2, x_1g_2x_3 \right) \]

(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 \((x_1, x_2, x_3) \in (\text{Cl}(g) \cap H)^3\), 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 \(\phi\) are all finite with size bounded independently of \((y_1, y_2)\) ∈ \(\text{G} \times \text{G}\), say of size at most \(c_1 \geq 1\). The hope behind (ii) is that \(\text{Cl}(g)^3\) and \(\text{G}^2\) have the same dimension,\(^3\) 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 \(|\text{Cl}(g) \cap H|\) by summing according to the values of \(\phi\): denoting \(Z = (\text{Cl}(g) \cap H)^3\) and \(W = \phi(Z) = \phi((\text{Cl}(g) \cap H)^3)\), we have

\[ |\text{Cl}(g) \cap H|^3 = |Z| = \sum_{(y_1, y_2) \in W} |\phi^{-1}(y_1, y_2) \cap Z| \]

which – under our optimistic assumption – leads to the estimate

\[ |\text{Cl}(g) \cap H|^3 \leq c_1|W| \leq c_1|H^{(k)}|^2 \leq c_1\alpha^{2(k-2)}|H|, \]

which has the form we want.

As it turns out, the assumption on \(\phi\) 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 \(x_1g_1x_2, x_1g_2x_3\) 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 \(\phi\) in that case, namely

\[ \phi(x_1, x_2, x_3) = (x_1x_2, x_1x_3) \in (H^{(2)})^2 \]

if \((x_1, x_2, x_3) \in (\text{Cl}(g) \cap H)^3\).

We are therefore led to analyze ruthlessly the fibers of the map \(\phi\). The resulting computations were explained to the author by R. Pink, and start with an easy observation:

**Lemma 5.6.19.** Let \(k\) be any field, and let \(G = \text{SL}_2(k)\). Let \(C \subset G\) be a conjugacy class, and define

\[ \phi \left\{ \begin{array}{c} C^3 \\ (x_1, x_2, x_3) \end{array} \rightarrow \text{G}^2 \right. \]

\[ \left( x_1x_2, x_1x_3 \right) \]

Then for any \((y_1, y_2) \in G \times G\), we have a bijection

\[ \begin{cases} C \cap y_1C^{-1} \cap y_2C^{-1} \\ x_1 \end{cases} \rightarrow \phi^{-1}(y_1, y_2) \]

\[ \left( x_1, x_2, x_3 \right) \]

In particular, if \(k = \mathbb{F}_p\) and \(C\) is a regular semisimple conjugacy class, we have a bijection

\[ \phi^{-1}(y_1, y_2) \rightarrow C \cap y_1C \cap y_2C. \]

\(^3\) Here dimension can be understood intuitively for readers who do not know the rigorous definition in algebraic geometry.
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 $\SL_2(k)^3$ 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.

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 $g$. \hfill \Box

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 5.6.20 (Pink). Let $k$ be an algebraically closed field of characteristic not equal to 2, and let $g \in \SL_2(k)$ be a regular semisimple element, $C$ 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 $B = xB_0 x^{-1}$ of the subgroup
   \[ B_0 = \left\{ \begin{pmatrix} a & b \\ 0 & a^{-1} \end{pmatrix} \right\} \subset \SL_2(k) \]

and an element $t \in B \cap C$ such that
   \[ y_1, y_2 \in U \cup t^2 U \]

where
   \[ U = xU_0 x^{-1}, \quad U_0 = \left\{ \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \right\} \subset B_0. \]

In that case, we have $X \subset C \cap B$.

3. The trace of $g$ is 0.

Remark 5.6.21. This result is the one place where it is really useful for us to work with $\SL_2(\mathbb{F})$. 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 be handled. 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 5.6.17, assuming $\Tr(g) \neq 0$. We construct the map

\[ \phi \left\{ \begin{array}{ccc} \Cl(g) \times \Cl(g) \times \Cl(g) & \longrightarrow & G \times G \\ (x_1, x_2, x_3) & \mapsto & (x_1 x_2, x_1 x_3) \end{array} \right. \]

and write (as before)

\[ Z = (\Cl(g) \cap H)^3, \quad W = \phi(Z) = \phi((\Cl(g) \cap H)^3) \]
to obtain
\[(5.36) \quad |\text{Cl}(g) \cap H|^3 = \sum_{(y_1, y_2) \in W} |\phi^{-1}(y_1, y_2) \cap Z| = S_0 + S_1 + S_2,
\]
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 5.6.20 hold. Precisely, we do not put into \(W\) 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 5.6.22.** With notation as above, we have
\[
S_0 \ll |H^{(2)}|^2 \ll \alpha^2|H|^2, \quad S_1 \ll |H^{(3)}|^2 \ll \alpha^2|H|^2,
\]
\[
S_2 \ll \alpha^{34/3}|H|^{5/3},
\]
where the implied constants are absolute.

Assuming this, we get immediately
\[
|\text{Cl}(g) \cap H| \ll \alpha^{2/3}|H|^{2/3} + \alpha^{34/9}|H|^{5/9}
\]
from (5.36). Now either the second term is smaller than the first, and we get (5.33) or
\[
\alpha^{34/9}|H|^{5/9} \gg \alpha^{2/3}|H|^{2/3}
\]
which gives
\[
\alpha \gg |H|^{1/28},
\]
the second alternative (5.34) of Theorem 5.6.17, 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 \(|W_0| \leq |W| \leq |H^{(2)}|^2\).

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
\[
S_{1,1} \leq \sum_{y_2 \in H^{(2)}} |\phi^{-1}(1, y_2) \cap Z|.
\]

But using Lemma 5.6.19, we have
\[
|\phi^{-1}(1, y_2) \cap Z| = |\{(x_1, x_1^{-1}, x_1^{-1} y_2) \in (\text{Cl}(g) \cap H^3)\}| \leq |H^{(3)}|
\]
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 \(x_1^{-1} = x_1^{-1} y_2 y_2^{-1} \in H^{(3)}\) for any such triple if \(y_2 \in H^{(2)}\). Therefore
\[
S_{1,1} \leq |H^{(2)}||H^{(3)}| \leq |H^{(3)}|^2,
\]
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 \(B_0\). Hence, up to a factor 2, the choice of \(y_1\) fixes that of the relevant conjugate \(B\) for which Case (2) applies. Next we observe that \(C_B = \text{Cl}(g) \cap B\) 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
and $\alpha$ is such that $\alpha + \alpha^{-1} = \text{Tr}(g)$. Given $y_1 \in H^{(2)}$ and $B$ containing $y_1$, we have by (5.35)
\[ y_2 \in (H^{(2)} \cap U) \cup (H^{(2)} \cap t^2 U) \]
for some $t \in C_B$. We note that $t^2 U$ is itself conjugate to $C_{\alpha^2}$ 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
\[ x_1 \in C_B \cap H, \]
we see that we should attempt to bound from above
\[ H^{(k)} \cap C_\gamma \]
for a fixed $k$ and a fixed $\gamma \in F_p^\times$, as this will lead us to estimates for the number of possibilities for $y_2$ as well as $x_1$. Using Lemma 5.6.23 below, we get indeed
\[ |\{y_2 \mid (y_1, y_2) \in W_2\}| \leq 8\text{trp}(H^{(2)}/H^{(2)})^{1/3} \ll \alpha^{25/3}|H|^{1/3}, \]
(the factor 8 accounts for the two possible choices of $B$ 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. \qed

**Lemma 5.6.23.** For any prime $p \geq 3$, any $\gamma \in F_p^\times$, any $x \in \text{SL}_2(F_p)$ and any symmetric generating set $H$ of $\text{SL}_2(F_p)$ containing 1, we have
\[ |H \cap xC_\gamma x^{-1}| = |H \cap x\left\{ \begin{pmatrix} \gamma & t \\ 0 & \gamma^{-1} \end{pmatrix} \mid t \in F_p \right\} x^{-1}| \leq 2\alpha^2|H|^{1/3} \]
where $\alpha = \text{trp}(H)$.

**Proof.** We first need to deal with the fact that $x$ and $\gamma$ are not necessarily in $\text{SL}_2(F_p)$ (since we used Lemma 5.6.20, which refers to algebraically closed field – we will see in the proof that it brings helpful simplifications). We have $xC_\gamma x^{-1} \cap \text{SL}_2(F_p) \subset xB_0 x^{-1} \cap \text{SL}_2(F_p)$, and there are three possibilities for the latter: either $xB_0 x^{-1} \cap \text{SL}_2(F_p) = 1$, or $xB_0 x^{-1} \cap \text{SL}_2(F_p) = T$ is a non-split maximal torus of $\text{SL}_2(F_p)$, or $xB_0 x^{-1} \cap \text{SL}_2(F_p) = B$ is an $\text{SL}_2(F_p)$-conjugate of the group $B_0 = B_0 \cap \text{SL}_2(F_p)$ of upper-triangular matrices (see Lemma B.2.6 in Appendix B). In this last case, we can assume that $x \in \text{SL}_2(F_p)$ and $\gamma \in F_p$. In the first, of course, there is nothing to do. And as for the second, note that $\gamma$ and $\gamma^{-1}$ are the eigenvalues of any element in $\text{SL}_2(F_p) \cap xC_\gamma x^{-1}$, and there are at most two elements in a maximal torus with given eigenvalues. A fortiori, we have $|H \cap xC_\gamma x^{-1}| \leq 2 \leq 2\alpha^2|H|^{1/3}$ in that case.

Thus we are left with the situation where $x \in \text{SL}_2(F_p)$. Using $\text{SL}_2(F_p)$-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
\[ |H \cap C_\gamma| \leq |H^{(2)} \cap C_1| = |H^{(2)} \cap U_0|, \]

125
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 5.6.3. Appplying the latter to $H^{(2)}$, we get

$$|H^{(2)} \cap U_0| \leq 2|H^{(8)}|^{1/3} \leq 2\alpha^2|H|^{1/3}$$

by Ruzsa’s Lemma (the statement of the proposition suggests $H^{(10)}$ instead, but looking
at the proof, we see that we can define (5.24) using an auxiliary element $h \in H$ instead
of $H^{(2)}$, so that the image is in $H^{(8)}$.)

We must still prove Lemma 5.6.20.

**Proof of Lemma 5.6.20.** This computation is based on a list of simple checks. We
can assume that the regular semisimple element $g$

$$g = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha^{-1} \end{pmatrix}$$

where $\alpha^4 \neq 1$, because $\alpha = \pm 1$ implies that $g$ is not regular semisimple, and $\alpha$ a fourth
root of unity implies that $\text{Tr}(g) = 0$, 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 $y_1 \in \text{SL}_2(k)$ and $x \in \text{SL}_2(k)$, 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 $y_1$...

It is a fact that the list of conjugacy classes in $\text{SL}_2(k)$ is well-known (see Proposi-
tion B.2.5 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:

$$y_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad y_1 = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix},$$

$$y_1 = \begin{pmatrix} \beta & 0 \\ 0 & \beta^{-1} \end{pmatrix}, \quad \beta \neq \pm 1, \beta \neq \alpha^\pm$$

$$y_1 = \begin{pmatrix} \alpha^2 & 0 \\ 0 & \alpha^{-2} \end{pmatrix}.$$

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):

$$\begin{pmatrix} \alpha & a \\ 0 & \alpha^{-1} \end{pmatrix} \text{ or } \begin{pmatrix} \alpha^{-1} & a \\ 0 & \alpha \end{pmatrix},$$

$$\begin{pmatrix} a & (-a^2 + at - 1)/(2t) \\ 2t & t - a \end{pmatrix}.$$
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 (5.37) but without assuming $\beta \neq \alpha^\pm 2$. For

$$x = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \in C,$$

we compute

$$y_1x = \begin{pmatrix} \beta a \\ \beta b \\ \beta^{-1} c \\ \beta^{-1}d \end{pmatrix}.$$

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

$$a = \frac{t}{\beta + 1}, \quad d = \frac{\beta t}{\beta + 1}.$$

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 $\beta$ is not one of $\alpha^2, \alpha^{-2}$ (the third case), and then $c$ and $d$ are non-zero, and we can parameterize the solutions as in (5.39), or else (the fourth case) $c$ or $d$ must be zero, and then we get upper or lower-triangular matrices, as described in (5.40).

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 $\alpha, \alpha^{-1}$ (as a set), see (5.38). We compute the trace of $y_2x$, and find that is

$$ax_3 + x_1\alpha + x_4\alpha^{-1}, \quad \text{or} \quad ax_3 + x_1\alpha^{-1} + x_4\alpha.$$

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 $1, \alpha^2$ and $\alpha^{-2}$, so that $y_2$ is upper-triangular with diagonal coefficients $(1, 1), (\alpha^2, \alpha^{-2})$ or $(\alpha^{-2}, \alpha^2)$, and this is one of the instances of Case (2) of Lemma 5.6.20.
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 (5.39), and compute the trace of $y_2x$ as a function of $a$, which gives

$$\text{Tr}(y_2x) = -\frac{x_3}{2t}a^2 + \left( x_1 - x_4 + \frac{x_3}{2} \right)a + (x_4 + 2x_2)t.$$ 

The equation $\text{Tr}(y_2x) = t$ 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 5.6.20.

All in all, going through the remaining situations, we finish the proof... □

5.7. A survey of generalizations
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 [46], which we follow closely (keeping track of the constants). For simplicity, we work only with symmetric sets, usually containing the identity.

The presentation is highly condensed, and there might well be minor computational mistakes remaining.

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 $AB = \{ab \mid a \in A, b \in B\}$, $A^n = \{a_1 \cdots a_n \mid a_i \in A\}$, and we use again the notation $A^{(n)}$ for $n$-fold product sets. 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 5.3.3), 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

$$d(A, B) = \log \left( \frac{|A \cdot B^{-1}|}{\sqrt{|A||B|}} \right).$$

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

$$d(A, B) \leq d(A, C) + d(C, B).$$

**Proof.** Spelling out the meaning of this inequality, we see that it is equivalent with

$$|A \cdot B^{-1}| \leq \frac{|A \cdot C^{-1}| |C \cdot B^{-1}|}{|C|},$$

which one proves by constructing an injective map

$$i : C \times (A \cdot B^{-1}) \rightarrow (A \cdot C^{-1}) \times (C \cdot B^{-1}),$$

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

$$i(c, x) = (a(x)c^{-1}, cb(x)^{-1})$$

129
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.
\[\square\]

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 \(d(A, B) \leq \log \alpha\), we write
   \[
   A \bullet_{\alpha} B ,
   \]
2. If \(A\) and \(B\) are sets with \(|B| \leq \alpha|A|\), we write
   \[
   B \bullet_{\alpha} A ,
   \]
   and in particular if \(|X| \leq \alpha\), we write
   \[
   X \bullet_{\alpha} 1 ,
   \]
3. If \(A\) and \(B\) are sets with \(e(A, B) \geq 1/\alpha\), we write
   \[
   A \bullet_{\alpha} B ,
   \]
4. If \(A \subset B\), we write
   \[
   A \rightarrow_{\alpha} B .
   \]

The following rules are easy to check (in addition to some more obvious ones which we do not spell out):

1. From
   \[
   A \bullet_{\alpha} B
   \]
   we can get
   \[
   A \rightarrow_{\alpha^2} B , \quad B \rightarrow_{\alpha^2} A .
   \]
2. (Ruzsa’s triangle inequality, [46, Lemma 3.2]) From
   \[
   A \bullet_{\alpha_1} B \bullet_{\alpha_2} C
   \]
   we get
   \[
   A \bullet_{\alpha_1 \alpha_2} C .
   \]
3. From
   \[
   C \bullet_{\alpha_1} B \bullet_{\alpha_2} A
   \]
   we get
   \[
   C \bullet_{\alpha_1 \alpha_2} A .
   \]
(4) (“Unfolding edges”) From
\[
\begin{align*}
B & \xrightarrow{\alpha} A \\
& \underset{\beta}{\text{}}
\end{align*}
\]
we get
\[
AB^{-1} \xrightarrow{\sqrt{\alpha \beta}} A
\]
(note that by the second point in this list, we only need to have
\[
B \xrightarrow{\beta} A
\]
to obtain the full statement with \(\alpha = \beta^2\), which is usually qualitatively equivalent.)

(5) (“Folding”) From
\[
\begin{align*}
AB^{-1} & \xrightarrow{\alpha} A \xrightarrow{\beta} B \\
& \underset{\alpha \beta^{1/2}}{\text{}}
\end{align*}
\]
we get
\[
A \xrightarrow{\alpha \beta^{1/2}} B
\]

Note that the relation \(A \xrightarrow{\alpha} B\) is purely a matter of the size of \(A\) and \(B\), while the other arrow types depend on structural relations involving the sets (for \(A \xrightarrow{\alpha} B\)) and product sets (for \(A \xrightarrow{\alpha} B\) or \(A \xrightarrow{\alpha} B\)).

### A.3. Statements and “Proofs”

The main result that we use in Chapter 5 is Theorem A.3.6 below. We present the arguments leading to the proof, following [46, §3.4.5].

**Theorem A.3.1** (Ruzsa covering lemma; Tao, Lemma 3.6). If
\[
AB \xrightarrow{\alpha} A
\]
there exists a set \(X\) which satisfies
\[
X \xrightarrow{\alpha} 1, \quad B \xrightarrow{\alpha} A^{-1}AX
\]
and symmetrically, if
\[
BA \xrightarrow{\alpha} A
\]
there exists \(Y\) with
\[
Y \xrightarrow{\alpha} 1, \quad B \xrightarrow{\alpha} XAA^{-1}
\]

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 5.6.5 in Chapter 5, but we state it again in our diagrammatic language.

**Theorem A.3.2** (Ruzsa). If \(A\) is symmetric and
\[
A^{(3)} \xrightarrow{\alpha} A
\]
then we have
\[
A^{(n)} \xrightarrow{\alpha^{n-2}} A
\]
for all \(n \geq 3\). In particular, we get
\[
A^{(7)} \xrightarrow{\alpha^5} A
\]
In [38, Th. 1.6] or [46, Lemma 3.4], one finds 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 much better constants.

**Theorem A.3.3 (Tao, Th. 3.9 and Cor. 3.10).** Let $A = A^{-1}$ with $1 \in A$ and

$A^{(3)} \overset{\alpha}{\longrightarrow} A$.

Then $H = A^{(3)}$ is a $(2\alpha^{44})$-approximate subgroup containing $A$.

**Proof.** We have first

$H \overset{\alpha}{\longrightarrow} A, \quad A \overset{\alpha}{\longrightarrow} H$.

Then by Ruzsa’s result, we get

$AH^{(2)} = A^{(7)} \overset{\alpha^5}{\longrightarrow} A$,

and by the Ruzsa covering lemma there exists $X$ with

$X \overset{\alpha}{\longrightarrow} H^{(2)}, \quad X \overset{\alpha^5}{\longrightarrow} 1$,

such that

$H^{(2)} \overset{\alpha}{\longrightarrow} A^{(2)}X \overset{\alpha}{\longrightarrow} A^{(3)}X = HX$.

Taking $X_1 = X \cup X^{-1}$, we get

$X_1 \overset{\alpha}{\longrightarrow} H^{(2)}, \quad X_1 \overset{2\alpha^5}{\longrightarrow} 1$,

and

$H^{(2)} \overset{\alpha}{\longrightarrow}HX, \quad H^{(2)} \overset{\alpha}{\longrightarrow} XH$,

which are the properties defining a $(2\alpha^5)$-approximate subgroup. □

**Theorem A.3.4 (Tao, Th. 4.6, (i) implies (ii)).** Let $A$ and $B$ with

$A \overset{\alpha}{\longrightarrow} B^{-1}$

Then there exists a $\gamma$-approximate subgroup $H$ and a set $X$ with

$X \overset{\gamma_1}{\longrightarrow} 1, \quad A \overset{\gamma}{\longrightarrow} XH, \quad B \overset{\gamma_2}{\longrightarrow} HX, \quad H \overset{\gamma_2}{\longrightarrow} A$,

where

$\gamma \leq 2^{21}\alpha^{80}, \quad \gamma_1 \leq 2^{28}\alpha^{104}, \quad \gamma_2 \leq 8\alpha^{14}$.

Furthermore, one can ensure that

\begin{equation}
(A.1) \quad H^{(3)} \overset{2^{10}\alpha^{48}}{\longrightarrow} H.
\end{equation}

**Proof.** From

$A \overset{1}{\longrightarrow} A, \quad A \overset{\alpha^2}{\longrightarrow} A$,

we get first

$AA^{-1} \overset{\alpha^2}{\longrightarrow} A$.

By [46, Prop. 4.5], we find a set $S$ with $1 \in S$ and $S = S^{-1}$ such that

$A \overset{2\alpha^2}{\longrightarrow} S, \quad AS^{(n)}A^{-1} \overset{2^{n}\alpha^{4n+2}}{\longrightarrow} A$.

\footnote{The property $1 \in S$ is not explicitly stated in [46], but follows from the explicit definition used by Tao, namely $S = \{x \in G \mid |A \cap Ax| > (2\alpha^2)^{-1}|A|\}$.}
for all $n \geq 1$. In particular, we get

$$AS^{-1} = AS \bullet_{2\alpha^6} \rightarrow A, \quad S \bullet_{2\alpha^6} \rightarrow A.$$  

Thus

$$AS^{-1} \bullet_{2\alpha^6} \rightarrow A \bullet_{2\alpha^6} \rightarrow S,$$  

which gives

$$A \bullet_\beta S$$  

by folding, with $\beta = 2\sqrt{2}\alpha^7$.

In addition, we have

$$S^{(3)} \bullet_{8\alpha^{14}} \rightarrow A \bullet_{2\alpha^2} \rightarrow S,$$  

and Theorem A.3.3 says that $H = S^{(3)}$ is a $\gamma$-approximate subgroup containing $S$, with $\gamma = 2(16\alpha^{16})^5 = 2^{21}\alpha^{80}$, and (as we see)

$$H \bullet_{8\alpha^{14}} \rightarrow A.$$  

Moreover, we have

$$H^{(3)} = S^{(9)} \overset{\gamma}{\longrightarrow} AS^{(9)} A^{-1} \bullet_{2^9\alpha^{38}} \rightarrow A \bullet_{2\alpha^2} \rightarrow S,$$  

which gives (A.1).

Now from

$$AH = AS^{(3)} \bullet_{8\alpha^{14}} \rightarrow A \bullet_{2\alpha^2} \rightarrow S \bullet_{1} \rightarrow H,$$  

we see by the Ruzsa covering lemma that there exists $Y$ with

$$Y \overset{\gamma}{\longrightarrow} A, \quad Y \overset{16\alpha^{16}}{\longrightarrow} 1, \quad A \overset{\gamma}{\longrightarrow} YHH.$$  

By definition of an approximate subgroup, there exists $Z$ with

$$Z \overset{\gamma}{\longrightarrow} 1, \quad HH \overset{\gamma}{\longrightarrow} ZH,$$  

and hence

$$A \overset{\gamma}{\longrightarrow} (YZ)H.$$  

Now we go towards $B$. First we have

$$AH^{-1} = AS^{(3)} \bullet_{8\alpha^{14}} \rightarrow A \bullet_{\alpha^2} \rightarrow H$$  

which, again by folding, gives

$$A \bullet_{\alpha_1} H$$  

with $\alpha_1 = 8\sqrt{2}\alpha^{15}$. Hence we can write

$$H \overset{\alpha_1}{\longrightarrow} A \overset{\alpha}{\longrightarrow} B^{-1},$$  

and so

$$H \overset{\alpha\alpha_1}{\longrightarrow} B^{-1}.$$  

In addition, we have

$$H \overset{8\alpha^{14}}{\longrightarrow} A \overset{\alpha^2}{\longrightarrow} B^{-1},$$  

133
and therefore we get
\[ H \xrightarrow{8\alpha_{16}} B^{-1}, \]
from which it follows by unfolding that
\[ B^{-1}H^{-1} = B^{-1}H \xrightarrow{32\alpha_{20}} B^{-1} \xrightarrow{\alpha_{2}} A \xrightarrow{2\alpha_{2}} H. \]

Once more by the Ruzsa covering lemma, we find \( Y_1 \) with
\[ Y_1 \xrightarrow{\gamma_1} B^{-1}, \quad Y_1 \xrightarrow{32\alpha_{20}} 1, \quad B^{-1} \xrightarrow{\gamma_1} Y_1HH \xrightarrow{\gamma_1} (Y_1Z)H. \]

Now we need only take \( X = (Y_1Z \cup YZ) \), so that
\[ X \xrightarrow{\gamma_1} 1 \]
with \( \gamma_1 = \gamma(64\alpha_{24} + 16\alpha_{16}) \), in order to conclude. Since
\[ \gamma_1 \leq 2^{28} \alpha_{104}, \]
we are done.

The next result is a version of the Balog-Gowers-Szemerédi Lemma.

**Theorem A.3.5** (Balog-Gowers-Szemerédi; Tao, Th. 5.2). Let \( A \) and \( B \) with
\[ A \xrightarrow{\alpha} B. \]

Then there exist \( A_1, B_1 \) with
\[ A_1 \xrightarrow{\gamma_1} A, \quad B_1 \xrightarrow{\gamma_1} B, \]
as well as
\[ A \xrightarrow{8\sqrt{2}\alpha} A_1, \quad B \xrightarrow{8\alpha} B_1, \]
and
\[ A_1 \xrightarrow{\alpha_1} B_1^{-1}, \]
where \( \alpha_1 = 2^{2019} \alpha_{9}. \)

This is not entirely spelled out in [46], but only the last two or three inequalities in the proof need to be made explicit to obtain this value of \( \alpha_1. \)

**Theorem A.3.6** (Tao, Th. 5.4; (i) implies (iv)). Let \( A \) and \( B \) with
\[ A \xrightarrow{\beta} B. \]

Then there exist a \( \beta \)-approximate subgroup \( H \) and \( x, y \in G \), such that
\[ H \xrightarrow{\beta_2} A, \quad A \xrightarrow{\beta_1} A \cap xH, \quad B \xrightarrow{\beta_1} B \cap Hy, \]
where
\[ \beta \leq 2^{1621} \alpha_{720}, \quad \beta_1 \leq 2^{2112} \alpha_{937}, \quad \beta_2 \leq 2^{283} \alpha_{126}. \]

Moreover, one can ensure that
\[ H^{(3)} \xrightarrow{\beta_3} H \]
where \( \beta_3 = 2^{810} \alpha_{360}. \)
PROOF. By the Balog-Gowers-Szemerédi Theorem, we get $A_1, B_1$ with

$$A_1 \rightarrow A, \quad B_1 \rightarrow B,$$

as well as

$$A \xrightarrow{8\sqrt{2}\alpha} A_1, \quad B \xrightarrow{8\alpha} B_1,$$

and

$$A_1 \xrightarrow{\alpha_1} B_1^{-1}$$

where $\alpha_1 = 2^{20}\alpha^9$. Applying Theorem A.3.4 to $A_1$ and $B_1$, we get a $\beta$-approximate subgroup $H$ and a set $X$ with

$$H \xrightarrow{2\alpha_1^4} A_1 \xrightarrow{1} A$$

and

$$X \xrightarrow{\gamma} 1, \quad A_1 \rightarrow XH, \quad B_1 \rightarrow HX,$$

where

$$\beta = 2^{21}\alpha_1^{80} = 2^{1621}\alpha^{720}, \quad \gamma = 2^{28}\alpha_1^{104} = 2^{2108}\alpha^{936},$$

and moreover

$$H^{(3)} \xrightarrow{\beta_3} H$$

where $\beta_3 = 2^{10}\alpha_1^{40} = 2^{810}\alpha^{360}$.

Applying the pigeonhole principle, we find $x$ such that

$$A \xrightarrow{8\sqrt{2}\alpha} A_1 \xrightarrow{\gamma} A_1 \cap xH \rightarrow A \cap xH$$

and $y$ with

$$B \xrightarrow{8\alpha} B_1 \xrightarrow{\gamma} B_1 \cap Hy \rightarrow B \cap Hy.$$ 

This gives what we want with

$$\beta_1 \leq 8\sqrt{2}\alpha \gamma \leq 2^{2112}\alpha^{937}, \quad \beta_2 = 8\alpha_1^{14} = 2^{283}\alpha^{126}.$$ 

$\square$
APPENDIX B

Some group theory

B.1. Free groups

In Chapter 5, we use a few basic facts about free groups, which we quickly summarize here.

**Proposition B.1.1.** Let $G$ be a free group on a set $S$ of $k \geq 1$ 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 $S = \{a_1^\pm, \ldots, a_k^\pm\}$ of free generators of $G$ and their inverses and let $\Gamma = C_G(G, S)$. If $x \neq 1$, then $d_{\Gamma}(1, x^n) \geq |n|$ for all $n \in \mathbb{Z}$.

**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., [24, Cor. 2.9].

(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 $n$. 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 $\geq 2$. 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 $S$. Let $r \geq 0$ 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

$$x^n = yx_1^ny^{-1}$$

for $n \in \mathbb{Z}$. Assume $n \geq 0$; 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

$$d_{\Gamma}(1, x^n) = 2r + nd_{\Gamma}(1, x_1) \geq n.$$  

For $n \leq 0$, 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., [24, §1.4, Cor. 1.2.2 and Prob. 2]).

We use the following example in Chapter 5.
Proposition B.1.3 (The Lubotzky group is free). (1) For \( k \geq 2 \), let
\[
    s_1 = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}, \quad s_2 = \begin{pmatrix} 1 & 0 \\ k & 1 \end{pmatrix} \subseteq \text{SL}_2(\mathbb{Z}).
\]

Then \( s_1, s_2 \) generate a free subgroup of rank 2 in \( \text{SL}_2(\mathbb{Z}) \). Moreover, for all \( p \nmid k \), the image of \( \{ s_1, s_2 \} \) modulo \( p \) generate \( \text{SL}_2(\mathbb{F}_p) \).

(2) In particular the Lubotzky group of Example 5.4.6 is free of rank 2 and the Cayley graphs \( \mathcal{G}(\text{SL}_2(\mathbb{F}_p), \{ s_1^{\pm 1}, s_2^{\pm 1} \}) \) are connected for all primes \( p \neq 3 \).

(3) For \( k = 2 \), the subgroup generated by \( s_1 \) and \( s_2 \) is the finite-index subgroup
\[
    \Gamma(2) = \ker(\text{SL}_2(\mathbb{Z}) \to \text{SL}_2(\mathbb{Z}/2\mathbb{Z})) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \pmod{2} \right\}
\]
of \( \text{SL}_2(\mathbb{Z}) \).

Proof. This is one of the simplest examples of the so-called “ping-pong technique” to produce free groups (see, e.g., [19, §II.B] for a more general discussion).

Let \( G \) be the subgroup generated by \( s_1 \) and \( s_2 \). We consider the action of \( G \) on \( \mathbb{R}^2 \) by left-multiplication, and introduce the two subsets
\[
    X_1 = \{(x, y) \in \mathbb{R}^2 \mid |x| > |y|\}, \quad X_2 = \{(x, y) \in \mathbb{R}^2 \mid |y| > |x|\}
\]
of \( \mathbb{R}^2 \). If \( (x, y) \in X_2 \) and \( m \in \mathbb{Z} \) we have
\[
    s_1^m(x, y) = (x + kmy, y)
\]
which, if \( m \neq 0 \), is in \( X_1 \) since
\[
    |x + kmy| \geq |kmy| - |x| = k|m||y| - |x| \geq (k - 1)|m||y| \geq |y|
\]
(note that \( k \geq 2 \) is important here). Thus we have \( s_1^m(X_2) \subseteq X_1 \) for all \( m \neq 0 \), and one checks similarly that \( s_2^m(X_1) \subseteq X_2 \) for \( m \neq 0 \).

Now we can start playing ping-pong to show that \( G \) is freely generated by \( s_1 \) and \( s_2 \). First, let \( k \geq 1 \) be odd and let \( n_1, \ldots, n_k \) be non-zero integers; consider the element
\[
    g = s_1^{n_1}s_2^{n_2}s_1^{n_3}\cdots s_1^{n_k} \in G.
\]
We must show that \( g \neq 1 \) (since it is a non-trivial reduced word in the generators); but indeed \( g \) must be non-trivial since
\[
    g(X_2) = (s_1^{n_1}s_2^{n_2}s_1^{n_3}\cdots s_1^{n_k})(X_2) \subseteq (s_1^{n_1}s_2^{n_2}s_1^{n_3}\cdots s_2^{n_k-1})(X_1) \subseteq \cdots \subseteq X_1,
\]
and \( X_1 \) and \( X_2 \) are disjoint! Similarly, if a word begins and ends with a power of \( s_2 \), it is non-trivial in \( G \), and if
\[
    g = s_1^{n_1}s_2^{n_2}s_1^{n_3}\cdots s_1^{n_k-1}s_2^{n_k},
\]
the conjugate element
\[
    s_2^{-n_k}gs_2^{n_k} = s_2^{-n_k}s_1^{n_1}s_2^{n_2}s_1^{n_3}\cdots s_1^{n_k-1}s_2^{2n_k}
\]
is non-trivial by the previous case, which means that \( g \) itself is, and the same for words beginning with \( s_2 \) and ending with \( s_1 \).

This proves the first part of (1). The second part, concerning the group generated by \( s_1 \pmod{p} \) and \( s_2 \pmod{p} \), follows from Proposition B.2.1, (2), below: \( \text{SL}_2(\mathbb{F}_p) \) is generated by
\[
    t_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad t_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},
\]
and for \( p \nmid k \) we have \( t_1 = s_1^n, t_2 = s_2^n \in \text{SL}_2(\mathbb{F}_p) \), where \( n \) is the inverse of \( k \) in \( \mathbb{Z}/p\mathbb{Z} \). Thus \( s_1 \) and \( s_2 \) also generate \( \text{SL}_2(\mathbb{F}_p) \) in that case.
Of course, part (2) here is just the special case $k = 3$.

As for (3), it is clear that the group $G_2$ generated by $s_1$ and $s_2$, in the case $k = 2$, satisfies

$$G_2 \subset \Gamma(2).$$

The converse inclusion follows by an easy variant of the “row-column operations” proof that the matrices

$$(\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}),$$

generate $\text{SL}_2(\mathbb{Z})$. \hfill \Box

**Remark B.1.4.** For $k = 1$, as just recalled, the group generated by

$$s_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad s_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},$$

is simply $\text{SL}_2(\mathbb{Z})$. This is not a free group: e.g., we have

$$t = s_1^{-1}s_2s_1^{-1} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

which satisfies $t^2 = 1$.

We also used some facts concerning random walks on regular trees, or equivalently on Cayley graphs of free groups. These are due to Kesten.

**Proposition B.1.5 (Kesten).** Let $k \geq 1$ be an integer, $S = \{a_1^{\pm 1}, \ldots, a_k^{\pm 1}\}$ such that $a_1, \ldots, a_k$ generate a free group $F_k$ of rank $k$, and let $T_k = \mathcal{C}(F_k, S)$ be the corresponding $(2k)$-regular tree. Let $(X_n)$ be the random walk on $T_k$ with $X_0 = 1$.

For all $n \geq 0$ and all $x \in T_k$, we have

$$P(X_n = x) \leq r^{-n}, \quad \text{where} \quad r = \frac{k}{\sqrt{2k - 1}} = \frac{|S|}{2\sqrt{|S| - 1}}.$$  

Note that $r > 1$ as soon as $k \geq 2$, but $r = 1$ for $k = 1$, which means that the proposition is trivial in that case.

**Proof.** Let $L^2(F_k)$ be the space of square-integrable functions on $F_k$ with the inner-product

$$\langle \varphi_1, \varphi_2 \rangle = \sum_{x \in F_k} \varphi_1(x)\overline{\varphi_2(x)},$$

(note that we can not normalize this since $F_k$ is infinite), and let $M$ be the Markov operator

$$\begin{cases} L^2(F_k) \rightarrow L^2(F_k) \\
\varphi \mapsto M\varphi
\end{cases}$$

with

$$M\varphi(x) = \frac{1}{|S|} \sum_{s \in S} \varphi(xs)$$

where $S = \{a_1^{\pm 1}, \ldots, a_k^{\pm 1}\}$. Just as in the case of finite graphs, one checks that $M$ is self-adjoint and has norm $\leq 1$. However, because the constant function 1 is not in $L^2(F_k)$, it is not clear that $\|M\| = 1$ – and in fact, this is not the case. We denote by $\rho$ the spectral radius of $M$, the largest absolute value of an eigenvalue of $M$ (without restricting to functions of average 0, since this does not make sense).
Still as in the case of finite graphs, we find
\[ P(X_n = x) = E(\delta_x(X_n)) = E((M^n\delta_x)(X_0)) = (M^n\delta_x)(1) \]
where \( \delta_x \) is the characteristic function of \( x \). But
\[ \varphi(1) = \langle \varphi, \delta_1 \rangle \]
for any \( \varphi \in L^2(F_k) \) (this uses the fact that the inner product is not normalized), and hence
\[ (M^n\delta_x)(1) = \langle M^n\delta_x, \delta_1 \rangle \leq \varrho^n \|\delta_x\|\|\delta_1\| = \varrho^n \]
since \( \|\delta_x\| = \|\delta_1\| = 1 \).

Thus it is now enough to prove that \( \varrho = r^{-1} \). Indeed, this is the main content of Kesten’s theorem. \( \square \)

**B.2. Properties of \( SL_2 \)**

We gather here the properties of \( SL_2(k) \), for \( k \) arbitrarily closed, and of the finite groups \( SL_2(F_p) \) which are used in Chapter 5. All of these are very standard, and go back to the earliest investigations of finite linear groups.

**Proposition B.2.1.**

1. For \( p \) prime, we have
\[ |SL_2(F_p)| = p(p^2 - 1). \]

2. For \( p \) prime, the group \( SL_2(F_p) \) is generated by the elements
\[ \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}. \]

3. For \( p \geq 3 \) prime, the invariant \( d(SL_2(F_p)) \) is equal to \((p - 1)/2\). In particular, \( SL_2(F_p) \) is perfect if \( p \geq 5 \).

We also use the knowledge of maximal (proper) subgroups of \( SL_2(F_p) \). To state the corresponding result, we first recall some terminology which is also used in Section 5.6.

A **split maximal torus** \( T \) in \( SL_2(F_p) \) is a conjugate of the subgroup
\[ T_s = \left\{ \begin{pmatrix} x & 0 \\ 0 & x^{-1} \end{pmatrix} \mid x \in F_p^\times \right\} \]
of diagonal matrices (equivalently, \( T \) is the subgroup of all elements which are diagonal with respect to a fixed basis of \( F_p^2 \)). A **non-split maximal torus** in \( SL_2(F_p) \) is a conjugate of the subgroup
\[ T_{ns} = \left\{ \begin{pmatrix} a & b \\ \varepsilon b & a \end{pmatrix} \in SL_2(F_p) \mid a^2 - \varepsilon b^2 = 1 \right\} \]
where \( \varepsilon \in F_p^\times \) is a fixed element which is not a square.

**Theorem B.2.2** (Dickson). Let \( p \geq 5 \) be a prime number and let \( H \subset SL_2(F_p) \) be a proper subgroup, not contained in any other proper subgroups. Then one of the following possibilities holds:

1. The group \( H/\{\pm 1\} \) is isomorphic to one of the groups \( A_4, S_4, A_5 \).
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. Most book treatments prove a similar statement for the simple group
\[ \text{PSL}_2(F_p) = \text{SL}_2(F_p)/\{\pm 1\}, \]
but it is easy to reduce to that case. First, a maximal subgroup \( H \) as above must contain \( \{\pm 1\} \), since \( H \subset H\{\pm 1\} \). Then, denoting by \( \pi \) the projection \( \text{SL}_2(F_p) \to \text{PSL}_2(F_p) \), we have
\[ H \subset \pi^{-1}(H'), \]
and by maximality, either \( \pi^{-1}(H') = \text{SL}_2(F_p) \), or else \( H = \pi^{-1}(H') \). But the first case can not occur, since it would mean that \( \pi \) restricted to \( H \) is an isomorphism \( \phi : H \cong \text{PSL}_2(F_p) \), and then \( H \) would be normal in \( \text{SL}_2(F_p) \) (being of index 2) and one would obtain a surjective homomorphism
\[ \text{SL}_2(F_p) \to \text{SL}_2(F_p)/H \cong \mathbb{Z}/2\mathbb{Z}, \]
contradicting the fact that \( \text{SL}_2(F_p) \) is perfect (or that \( d(\text{SL}_2(F_p)) \geq 2 \)). \( \square \)

Remark B.2.3. It can be useful to see concretely the shape of a normalizer of a maximal torus: it is conjugate either to
(B.1) \[ N(T_s) = \left\{ \begin{pmatrix} x & 0 \\ 0 & x^{-1} \end{pmatrix} \right\} \cup \left\{ \begin{pmatrix} 0 & x \\ -x^{-1} & 0 \end{pmatrix} \right\} \]
in the split case or to
(B.2) \[ N(T_{ns}) = \left\{ \begin{pmatrix} a & b \\ \varepsilon b & a \end{pmatrix} \right\} \cup \left\{ \begin{pmatrix} \alpha & \beta \\ -\varepsilon \beta & -\alpha \end{pmatrix} \right\} \]
in the non-split case, with \( a, b \) and \( \alpha, \beta \) constrained by \( a^2 - \varepsilon b^2 = 1 \) and \( -\alpha^2 + \varepsilon \beta^2 = 1 \), respectively. We can see in particular that \( N(T)/T \cong \mathbb{Z}/2\mathbb{Z} \) in all cases (this is easy to see in the split case, and may require a small computation in the non-split case.)

We use in Chapter 5 the following immediate corollary:

Corollary B.2.4. Let \( p \geq 5 \) be a prime number and let \( H \subset \text{SL}_2(F_p) \) be a proper subgroup, not contained in any other proper subgroups. Then either \( |H| \leq 120 \) or we have
\[ [[x_1, x_2], [x_3, x_4]] = 1 \]
for any \( x_i \in H \).

In our proof of the non-concentration inequalities, we also use the classification of conjugacy classes of \( \text{SL}_2(F_p) \) and \( \text{SL}_2(F_p) \).

Proposition B.2.5. (1) Let \( p \geq 3 \) be a prime, and let \( \varepsilon \in F_p^\times \) be a fixed element which is not a square. There are \( p + 4 \) conjugacy classes in \( \text{SL}_2(F_p) \), which have representatives of the following forms:
\[
\begin{aligned}
1, & \quad \begin{pmatrix} x & 0 \\ 0 & x^{-1} \end{pmatrix}, \quad x \in F_p^\times - \{\pm 1\}, \\
& \quad \begin{pmatrix} a & b \\ \varepsilon b & a \end{pmatrix}, \quad b \neq 0, \quad a^2 - \varepsilon^2 = 1, \\
& \quad \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}, \\
& \quad \begin{pmatrix} 1 & \varepsilon \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} -1 & \varepsilon \\ 0 & -1 \end{pmatrix}.
\end{aligned}
\]
(2) Let $k$ be an algebraically closed field of characteristic $p \geq 3$. Representatives of the conjugacy classes in $\text{SL}_2(k)$ are given by the elements

$$1, -1, \begin{pmatrix} x & 0 \\ 0 & x^{-1} \end{pmatrix}, x \in k^\times - \{\pm 1\},$$

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}.$$

Here is a last useful fact used in Section 5.6:

**Lemma B.2.6.** Let $p \geq 3$ be a prime number, $G = \text{SL}_2(\mathbb{F}_p)$ and $\bar{G} = \text{SL}_2(\overline{\mathbb{F}}_p)$. Let $B \subset G$ be the subgroup of upper-triangular matrices, and let $x \in G$. Then $xBx^{-1} \cap G$ is either trivial, or a non-split maximal torus in $G$, or a $G$-conjugate of $B \cap G$.

**Proof.** The group $B_1 = xBx^{-1}$ can be described as the set of $g \in G$ such that $gv$ is a multiple of $v$, for some fixed non-zero vector $v \in \mathbb{F}_p^2$. If $g \in G \cap B_1$, then all Galois-conjugates of $v$ are eigenvectors of $g$ (since the latter is fixed under Galois), and so there are at most two of them, up to scaling. If $v^\sigma$ is a multiple of $v$, then after replacing $v$ by some non-zero multiple, we can assume that $v \in \mathbb{F}_p^2$. In that case $G \cap B_1$ is the set of matrices which are upper-triangular in a fixed basis with $v$ as first element, and thus it is $G$-conjugate to $G \cap B$. If $w = v^\sigma$ is linearly independent to $v$, and $w^\sigma = v$, then having $v$ and $w$ as eigenvectors means precisely that elements of $G \cap B_1$ are in the unique maximal torus in $B_1$, which is then a non-split maximal torus of $G$. \qed
Bibliography


