# Constructing Expander Graphs: A Didactic Approach

(M.Sc. Thesis, Computer Science)

Faculty for Computer Science and Business Information Systems, Karlsruhe University of Applied Sciences

**Martin Redlof**

Reviewers:
Prof. Dr. Körner
Prof. Dr. Schaefer

25.09.2022

# Statement of Authorship

I hereby declare that I have produced this work by myself, and that I have labeled all material taken with or without modification from the work of others.

Karlsruhe, 25.09.2022

Martin Redlof

**Abstract**

We present a method for constructing certain expander graphs, as described in an article by N. Alon, O. Schwartz and A. Shapira [ASS08], with detailed preparations and explanations about all the mathematical and graph-theoretical concepts involved.

An extensive mathematical appendix may serve as refresher for those with undergraduate maths education and doubles as a reference for the necessary calculations in the main part.

# Contents

# Chapter 1

# Introduction

## 1.1 Problem

Expander graphs constitute a family of graphs with high yet sparse connectivity. When arbitrary edges are deleted, the resulting graph remains a single connected component for quite some time. Differently put, there is no way to cut through the graph's representation without slicing a significant number of edges.

There are practical applications for such graphs, e.g. in network planning, where it is a benefit to be able to tolerate single connections breaking down, and to re-route – this applies to traffic networks as well as the electric grid or internet connections.

However, expanders are also used in theoretical contexts, e.g. in I. Dinur's proof of the PCP theorem (PCP = probabilistically checkable proof) [Din05]. Further applications can be found in [HLW06].

## 1.2 Task Overview

The aim of this thesis is an easy-to-follow description of how to construct expander graphs, explaining all the necessary steps and providing the mathematical background that cannot be addressed in the confines of a research paper (or that might be expected from professional readers).

The thesis should be accessible to anyone interested with no more than undergraduate-level mathematics education.

## 1.3 Organization of this Thesis

The main body of this work comprises three parts:

1. An extensive preparatory chapter (2) introducing and illustrating concepts from graph theory as they are required later. This is by no means comprehensive and cannot substitute any textbook (like, for instance, [Nic18] (spectral graph theory) or [Sta17] on regular graphs).

2. The main construction chapter (3), modeled after an article [ASS08] by N. Alon, O. Schwartz and A. Shapira which introduces a concise and effective way to construct expander graphs with fixed degree (that is, the number of edges connecting a graph node to other nodes is not dependent on the graph's node count).

3. Several chapters (A to E) constituting a mathematical appendix, intended to refresh undergraduate maths knowledge and to serve as a reference for the calculations in the two chapters mentioned above.

   The appendix is designed incrementally, so that it can be perused from start to finish if so desired or needed.

Because the two main chapters frequently reference statements from the maths appendix, it can be helpful to glance over the contents of the latter, if only to familiarize oneself with notation, but perhaps also in order to ascertain the range of concepts used in the main part.

# Chapter 2

# Preliminaries

In order to approach the main subject, we first introduce some basic concepts from (spectral) graph theory. This is not intended in any way as a comprehensive overview, but as a selection of definitions and examples relevant to the particular topic of regular edge expanders. The interested reader may supplement the material with textbooks like [Nic18, Sta17].

We first state some basic properties of undirected graphs, particularly relating to their spectra. After this, we offer some additional properties of regular graphs, before we look at several examples.

We will refer to several results developed in the mathematical appendix: This chapter mainly relies on the linear algebra section A.3 of the first appendix chapter, and on the eigenvalue problem (chapter C). Readers familiar with determinants and permutations can ignore chapter B (situated in between), but the concept of determinants is in fact vital to the solution of the eigenvalue problem; we will calculate several determinants in the example section of this chapter.

## 2.1 Graph Vocabulary

We start by defining an *undirected graph*. We introduce *adjacency matrices*, which contain (at least for this work's purposes) all the information of a graph, provided that the vertices (a.k.a. nodes) are named with integer numbers.

The notion of *node degree* will allow us to define *regular graphs*; this special kind of graph will be examined in more detail because the expander graphs constructed in the main chapter 3 will all be regular.

We also present two ways to color graphs: vertex and edge coloring.

After that, we will introduce the concept of a graph's spectrum in a separate subsection.

### 2.1.1 Basic Definitions for Graphs

Undirected graphs are collections of vertices (a.k.a. nodes; we will use both terms interchangeably) that are connected in some specific way. For undirected graphs, those connections are symmetric (if node $j$ is connected to node $k$, then $k$ is also connected to $j$) and can be visualized as lines between the various nodes[1]. We start with multi-graphs and then specialize for simple graphs. In preparation, and to allow for a unified approach to describe the connections (*edges*), we first introduce the concept of a *half-edge*.

#### Half-Edges, Node Degrees, Edges, Multi-Edges and Loops

We start with multiple definitions, after which we will elaborate with some examples.

**Definition 2.1**

- *A* node *(a.k.a.* vertex*) may be any object that could be understood as "connected to some other object" in a graph. Such connections are called edges (see below).*
- *An undirected* graph *is a pair of a set of vertices $V$ and a structure describing the undirected edges.*

[1] For directed graphs, one would use arrows.

- *An undirected* half-edge *is one half of an undirected connection (*edge*) in a graph. Its fixed end is incident with (read: connected to) exactly one node of a graph. Its open end can be combined with the open end of exactly one other half-edge when building the graph.*
- *The number of half-edges of a node $j$ is called the node's* degree*, denoted* $\deg(j)$.
- *An undirected* edge *is a fixed pair of undirected half-edges.*
- *A* loop *is an edge whose half-edges both are incident with the same node.*
- *An $r$-fold* multi-edge *exists between two different nodes if there are $r$ edges between them. $r$-fold* multi-loops *involving a single node are also possible.*

We should note that directed graphs differ in that there are two kinds of half-edges (emanating and incoming, as it were), and that directed edges would have to be combinations of one half-edge of each kind. We will, however only be dealing with undirected graphs in this work.

Before we draw some formal conclusions, we visualize a node with degree 4.

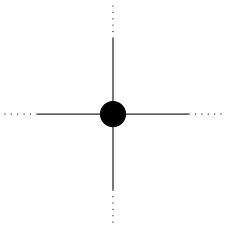

Figure 2.1: A degree-4 node with half-edges

Now, in a graph, there are only edges, no half-edges. When building a graph, all the half-edges must therefore be paired up by connecting each of them to another half-edge. An analogy for half-edges vs. edges would be the electrons vs. covalent molecular bonds in chemistry (ignoring the possibility of free radicals, that is).

This node might be connected to four other nodes, with a single edge each:

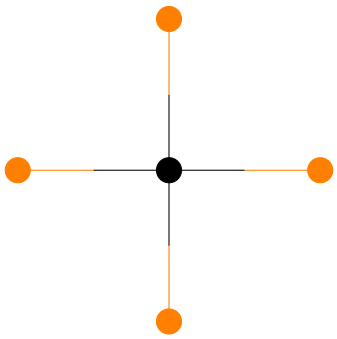

Figure 2.2: A degree-4 node, simply connected

Or it might be connected to only two other nodes, and have a loop; or have a double loop:

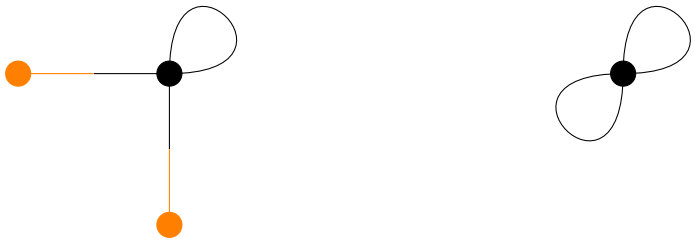

Figure 2.3: A degree-4 node with a loop (left), with a double loop (right)

Or it might be connected to two other nodes, but with a triple edge in one case; or to one, with a loop:

Figure 2.4: A degree-4 node with a triple edge (left), with a loop and double edge (right)

For all the above examples (not an exhaustive list), the black node has degree 4.
We collect some observations:

**Corollary 2.2** *For any node in a graph, its loops contribute two each to its degree. Edges to other nodes contribute one each.*

*For any graph, the degrees of its nodes sum up to an even number.*

Proof: Any loop is comprised of two half-edges belonging to the same node. Edges between two different nodes contribute one half-edge to each of those nodes.

Since a graph is comprised of nodes and edges, each edge contributes two half-edges to the total sum of degrees. There are no unconnected half-edges in a graph. ■

**Corollary 2.3** *In any graph, the number of nodes with odd degree is even.*

Proof: If there were an odd number of nodes with odd degrees, their respective degrees would some up to an odd number. In the overall graph, there would have to be one unconnected half-edge, which is not permitted. ■

### Undirected Graphs

In definition 2.1 (p. 9), we left the description of a graph's edges somewhat vague – this is because there are several ways to specify edges in a graph. For a directed graph with vertex set $V$, we could comprise the edges as a subset of $V \times V$; if $(j, k)$ belonged to that subset, the graph would contain a connection from node $j$ to $k$.

In undirected graphs, ordered pairs would not be such a good description because the connections in an undirected graph are symmetric. The symmetry could rather be expressed by specifying not tuples but multi-sets[2] ("bags") of size two. However, this would not suffice if there were multiple loops or multi-edges.

The most flexible method of expressing graph edges is by an edge function $e : V \times V \to \mathbb{N}_0$. We opt for this approach and specify our definitions for undirected graphs:

**Definition 2.4** *An* undirected multi-graph *is a pair $(V, e)$ of a set of vertices (a.k.a. nodes) identified via their integer labels, $V = \{1, \cdots, n\}$, together with a function $e : V \times V \to \mathbb{N}_0$ that specifies connections (edges) between vertices:*

*For $j, k \in V$, $e(j, k)$ returns the number of half-edges of node $j$ that connect it with node $k$.*

**Corollary 2.5** *Let $G = (V, e)$ be an undirected multi-graph with edge function $e$, and $j, k \in V$.*

*Then, $e(j, k)$ equals either the number of edges between nodes $j$ and $k$ if $k \neq j$, or twice the number of loops for node $j$ if $k = j$.*

*The edge function is symmetric: $e(k, j) = e(j, k)$.*

Proof: If $k \neq j$, the number of edges connecting $j$ and $k$ equals the number of half-edges at node $j$ contained in those edges (refer to definition 2.1, p. 9, for details). If $k = j$, the function counts all half-edges connecting $j$ with $j$, two of which make up one loop. For any such loop, both its half-edges are counted.

Since the connectivity in an undirected graph is symmetric, for $k \neq j$, there are as many half-edges at node $k$ connecting it with node $j$ as there are half-edges at $j$ in connections with $k$; hence the symmetry of $e$. ■

**Definition 2.6** *An* undirected simple graph *is an undirected multi-graph $(V, e)$ whose edge function $e$ only maps to $\{0, 1\}$.*

[2] ordinary sets would not allow the same node to appear twice, which it would in a loop

**Corollary 2.7** *If $G = (V, e)$ is an undirected simple graph, $G$ is loop-free and has no multiple edges.*

Proof: Edges between different nodes can only be 1-fold.

Since $e(j, j)$ is even for any undirected multi-graph (cf. corollary 2.5), the only even number in $\{0, 1\}$ is zero. ■

### Adjacency Matrices

Instead of the edge function, we may express the connectivity in a graph $G = (V, e)$ with a matrix, too, if we label the nodes from 1 to $n = |V|$:

**Definition 2.8** *For a graph $G = (V, e)$ with $n := |V|$, its* adjacency matrix *is a matrix of integer entries from $\mathbb{N}_0^{n \times n}$:*

$$\forall j, k \in V : A_{jk} := e(j, k)$$

**Corollary 2.9** *For an undirected graph $G = (V, e)$, its adjacency matrix $A$ is symmetric. If $G$ is a simple graph, the elements of $A$ are either 0 or 1, and the diagonal elements of $A$ are zero.*

Proof: Since $e(k, j) = e(j, k)$, $A_{kj} = A_{jk}$. By definition A.25 (p. 77), this means that $A$ is symmetric. In a simple undirected graph, definition 2.6 stipulates that $e$ returns either 0 or 1, and also that $e(j, j) = 0$ for all $j \in V$. ■

Example: We consider the simple graph $G_1$ with six vertices and connections as shown in figure 2.5. The adjacency matrix $A$ of $G_1$ (and, equivalently, its edge function $e$), is given by:

$$A(G_1) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

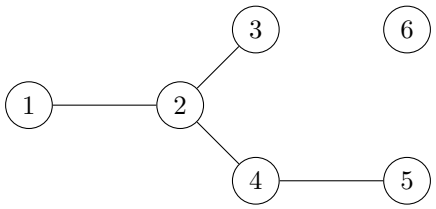


Figure 2.5: A simple graph ($G_1$)

In a multi-graph (as opposed to a simple graph), multiple connections between nodes are allowed; also, the graph may contain *loops*, i.e. edges that connect a vertex with itself. Such graphs can, for instance, be used to describe finite state machines in theoretical computer science.

Example: We consider the multi-graph $G_2$ with six vertices and connections as shown in figure 2.6 (p. 13). The adjacency matrix $A$ of $G_2$ is given by:

$$A(G_2) = \begin{pmatrix} 0 & 2 & 2 & 0 & 0 & 0 \\ 2 & 0 & 1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{pmatrix}$$

We observe that $G_2$'s adjacency matrix has two (even) non-zero diagonal entries – one for the single loop of node 5, and one for the double loop of node 6. Corollary 2.9 is valid for multi-graphs, too, as is definition 2.8 for the adjacency matrix: All undirected graphs (simple or not) have symmetric adjacency matrices.

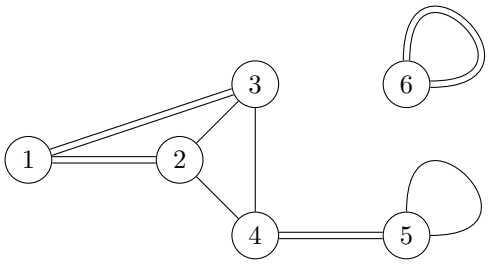


Figure 2.6: A multi-graph ($G_2$)

To finish off our general observations on adjacency matrices, we confirm that the adjacency matrix stays symmetric if we re-label the nodes, i.e. if we apply a permutation on the node numbers:

**Lemma 2.10** *For an undirected graph $G = (V, e)$ with adjacency matrix $A$, $n := |V|$, and a permutation of node labels $\sigma \in S_n$, the adjacency matrix $\tilde{A}$ of the graph with permuted labels is still symmetric.*

Proof: We use lemma B.23 (p. 87) and lemma B.22 (p. 87). The permuted adjacency matrix is given by $P_\sigma^T A P_\sigma$, and because the permutation matrix $P_\sigma$ is orthogonal, this also means that $\tilde{A} = P_\sigma^{-1} A P_\sigma$, so $\tilde{A}$ and $A$ are similar via an orthogonal matrix (cf. corollary A.29, p. 79). This means that the similar matrix $\tilde{A}$ is also symmetric, because $A$ is. ■

(This is just a technical reassurance, because since re-labeling the nodes does not change the fact that we are considering a graph, the relabeled graph is a bona fide undirected graph as well, and thus its adjacency matrix must of course be symmetric.)

### Node Degrees and Regular Graphs

We recall the definition 2.1 of a node's degree via its half-edges (p. 9), and our observations in corollary 2.5 (p. 11), and give an alternative definition of node degree via the adjacency matrix:

**Definition 2.11** *For a node $j \in V$ of an undirected graph $G = (V, e)$, the* degree *of $j$ is defined per*

$$\deg(j) := \sum_{k \in V} e(j,k) = \sum_{k \in V} e(k,j).$$

*If $A$ is the adjacency matrix of $G$, this means*

$$\deg(j) = \sum_{k \in V} A_{jk} = \sum_{k \in V} A_{kj}.$$

Thus, the degree of a node $j$ is just the sum of all the components in the adjacency matrix's $j$-th row (or column).

We observe that in our example $G_1$ (see above), there are several degrees: Nodes 1, 3 and 5 have degree 1, node 4 has degree 2, node 2 has degree 3, and node 6 has zero degree. In $G_2$, all nodes have degree 4. In fact, this makes $G_2$ a 4-regular graph:

**Definition 2.12** *An undirected (multi-)graph $G = (V, e)$ is called* regular with degree *$d$, or $d$-*regular*, if all its nodes have degree $d$.*

(More on regular graphs to follow in the next section.)

### Connected Components

Both of the above example graphs share another characteristic: they each consist of two connected components. In order to back this statement up formally, we introduce the notion of indirect connectedness:

**Definition 2.13** *For an undirected graph $G = (V, e)$, two nodes $j, k \in V$, $j \neq k$, are called* indirectly connected *if there is a path*

$$j = r_1 - r_2 - \cdots - r_m = k$$

*such that* $e(r_p, r_{p+1}) \neq 0$ *for all* $p \in \{1, \cdots, (m-1)\}$*; i.e. if one can find edges to travel between* $j$ *and* $k$*.*

*In addition, any node* $j \in V$ *is indirectly connected to itself.*

**Lemma 2.14** *The indirect connectedness is an equivalence relation.*

Proof: The relation is reflexive because any node is (by definition) indirectly connected to itself. It is symmetric because the graph is undirected. If a node $k$ can be reached from $j$, traveling along edges all the way, then, $j$ can also be reached (along the same route) from $k$. The relation is also transitive because if $j$ is indirectly connected to $r$, and $r$ to $k$, then any path from $j$ to $r$ can be extended by a path from $r$ to $k$ to yield a path from $j$ to $k$. ■

**Definition 2.15** *For an undirected graph* $G = (V, e)$*, its* connected components *are the equivalence classes of nodes under the relation of indirect connectedness.*

*If* $G$ *has exactly one connected component, it is called* connected*; otherwise, it is called* disconnected*.*

For both our two examples, those equivalence classes are the sets $\{1, 2, 3, 4, 5\}$ and $\{6\}$, respectively.

### Graph Coloring

For the purpose of the main chapter on expander graphs, we will also need the notion of *edge coloring*:

**Definition 2.16** *An undirected graph* $G = (V, e)$ *is called* $k$-edge-colorable *(with* $k \in \mathbb{N}$*) if any edge can be colored in one of* $k$ *colors such that no node is incident with two or more half-edges of the same color.*

**Corollary 2.17** *A* $k$*-edge-colorable graph contains no loops.*

Proof: A loop edge connects two half-edges of a single node – both of which would have to be of the same edge color, and thus violate definition 2.16. ■

**Corollary 2.18** *An edge-colorable* $d$*-regular graph cannot be less than* $d$*-edge-colorable.*

Proof: Any node such a graph is incident with $d$ edges, none of which are loops. We need at least $d$ different colors to distinguish between the incident edges. ■

Also, instead of (or in addition to) coloring the edges, we may color a graph's nodes:

**Definition 2.19** *An undirected graph* $G = (V, e)$ *is called* $k$-(vertex/node-)colorable *(with* $k \in \mathbb{N}$*) if any node can be colored in one of* $k$ *colors such that no two nodes of the same color are connected by an edge.*

*We call 2-colorable graphs* bipartite*.*

**Corollary 2.20** *A* $k$*-colorable graph contains no loops.*

Proof: A loop connects a node to itself. ■

**Corollary 2.21** *A simple graph* $G = (V, e)$*,* $n := |V|$*, is always* $n$*-colorable.*

Proof: $G$ is simple, and therefore loop-free (cf. corollary 2.7, p. 12). Regardless of any edges, we may assign an individual color to any of its $n$ nodes. ■

## 2.1.2 The Graph Spectrum

We recall from corollary 2.9 (p. 12) that undirected graphs have symmetric adjacency matrices. The *Spectrum* of $A$, i.e. the set of eigenvalues, therefore is made up entirely of real numbers (cf. lemma C.9, p. 102). We will use this fact later in the main construction chapter 3, but we can observe already that this will enable us to order all the eigenvalues with respect to their size (this would not be possible if there were eigenvalues with non-zero imaginary parts):

Before we calculate the spectra of our two example graphs, we state an important fact (that is developed in the appendix):

**Lemma 2.22** *Two graphs that differ only by a permutation of node numbers share the same spectrum of eigenvalues ("they are co-spectral").*

Proof: We already pointed out above (lemma 2.10, p. 13) that graphs with permuted labels have similar adjacency matrices. But then, lemma C.8 (p. 102) states that both matrices will also have the same spectrum of eigenvalues. Thus, the spectrum is truly a property of the connectivity of a graph. ■

We would also like to recall that all eigenspaces of a symmetric matrix are orthogonal on each other (and all eigenspaces can have a full orthogonal basis), as per corollary C.11 and theorem C.10 (pp. 103, 102, respectively). We will exploit this orthogonality in the coming section 2.2 on regular graphs.

### A Particular Determinant

In preparation for several eigenvalue calculations, we now prove a general formula for the determinant of a certain matrix. After that, we revisit our two example graphs.

**Lemma 2.23** *For $n \in \mathbb{N}$ and $a, b \in \mathbb{R}$, the matrix $M_n \in \mathbb{R}^{n\times n}$ with diagonal elements $a$ and off-diagonal elements $b$ (everywhere), i.e.,*

$$(M_n)_{jk} = \begin{cases} a, & j = k \\ b, & otherwise \end{cases}$$

*has determinant*

$$\det M_n = \det \begin{pmatrix} a & b & \cdots & \cdots & b \\ b & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & b \\ b & \cdots & \cdots & b & a \end{pmatrix} = (a-b)^{n-1} \cdot [a + (n-1)b]$$

Proof: First, we observe that every row of $M_n$ contains one component $a$ and $(n-1)$ elements $b$ (this is true for the degenerate case $n = 1$, too). We therefore add columns $2, \cdots, n$ to the first column, which will not change the determinant (according to corollary B.26, p. 88). This sums up all the components of each row in the first column:

$$\det M_n = \det \begin{pmatrix} [a + (n-1)b] & b & \cdots & \cdots & b \\ \vdots & a & \ddots & & \vdots \\ \vdots & b & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & b \\ [a + (n-1)b] & b & \cdots & b & a \end{pmatrix}$$

Now, if the first column of that matrix were zero, then the whole determinant would vanish, according to corollary B.32 (p. 89), which is in agreement with our statement, where the square bracket expression appears as a factor. In any case, we may use corollary B.25 (p. 88) to extract the factor and leave a column consisting of ones in index 1:

$$\det M_n = [a + (n-1)b] \cdot \det \begin{pmatrix} 1 & b & \cdots & \cdots & b \\ \vdots & a & \ddots & & \vdots \\ \vdots & b & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & b \\ 1 & b & \cdots & b & a \end{pmatrix}$$

In the next step, we use corollary B.26 again and add that column, scaled with $(-b)$ to all the other columns. This eliminates all off-diagonal components $b$, and leaves $(n-1)$ terms $(a-b)$ on

the diagonal.

$$\det M_n = [a+(n-1)b] \cdot \det \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & (a-b) & \ddots & & \vdots \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 1 & 0 & \cdots & 0 & (a-b) \end{pmatrix}$$

(If $b$ were zero, the last step would have amounted to no change, but the matrix would have looked like this after the previous step already.)

This remaining matrix is lower-triangular, and we can use lemma B.35 (p. 90) to calculate its determinant as the product of all its diagonal elements, which is just $(a-b)^{n-1}$. Together with the factor in square brackets, this yields the expression in the statement. ■

We now proceed to calculate our example spectra.

**Spectrum of $G_1$**

According to corollary C.3 (p. 100), we have to solve the equation $\det(A(G_1) - \lambda \mathbb{1}_6) = 0$ for $\lambda$:

$$0 = \det \begin{pmatrix} -\lambda & 1 & 0 & 0 & 0 & 0 \\ 1 & -\lambda & 1 & 1 & 0 & 0 \\ 0 & 1 & -\lambda & 0 & 0 & 0 \\ 0 & 1 & 0 & -\lambda & 1 & 0 \\ 0 & 0 & 0 & 1 & -\lambda & 0 \\ 0 & 0 & 0 & 0 & 0 & -\lambda \end{pmatrix} = (-\lambda) \det \begin{pmatrix} -\lambda & 1 & 0 & 0 & 0 \\ 1 & -\lambda & 1 & 1 & 0 \\ 0 & 1 & -\lambda & 0 & 0 \\ 0 & 1 & 0 & -\lambda & 1 \\ 0 & 0 & 0 & 1 & -\lambda \end{pmatrix},$$

where we used Laplace's expansion along column 6 in the second equality. We now expand along the first column:

$$\cdots = (-\lambda) \left[ (-\lambda) \det \begin{pmatrix} -\lambda & 1 & 1 & 0 \\ 1 & -\lambda & 0 & 0 \\ 1 & 0 & -\lambda & 1 \\ 0 & 0 & 1 & -\lambda \end{pmatrix} - \det \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & -\lambda & 0 & 0 \\ 1 & 0 & -\lambda & 1 \\ 0 & 0 & 1 & -\lambda \end{pmatrix} \right]$$

In the right-hand matrix, we can expand along the first row, to get a single non-zero contribution from a $3 \times 3$ matrix. This latter matrix is block-diagonal, and as per lemma B.48 (p. 98), we may calculate its determinant as a product of determinants of the two blocks, of which one is just $(-\lambda)$. The other block is two by two, which we cover as an example after the proof of theorem B.42 (p. 95); thus, the right-hand $4 \times 4$ determinant from above is $(-\lambda)(\lambda^2 - 1)$.

We expand the left-hand $4 \times 4$ determinant along its fourth column and get as overall result:

$$\cdots = (-\lambda) \left[ (-\lambda) \left( - \det \begin{pmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 0 \\ 0 & 0 & 1 \end{pmatrix} + (-\lambda) \det \begin{pmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 0 \\ 1 & 0 & -\lambda \end{pmatrix} \right) - \left[ (-\lambda)(\lambda^2 - 1) \right] \right]$$

Both remaining determinants can be expanded along the third row; this yields $(\lambda^2 - 1)$ for the left-hand one and $\lambda + (-\lambda)(\lambda^2 - 1)$ for the right-hand one. In total:

$$0 = (-\lambda) \left[ (-\lambda) \left( -(\lambda^2 - 1) - \lambda(\lambda - \lambda(\lambda^2 - 1)) \right) - \left[ (-\lambda)(\lambda^2 - 1) \right] \right]$$

We can factor out another $(-\lambda)$ to get:

$$0 = (-\lambda)^2 \left[ -(\lambda^2 - 1) - \lambda(\lambda - \lambda(\lambda^2 - 1)) - (\lambda^2 - 1) \right] = \lambda^2 \left[ -\lambda^2 + \lambda^2(\lambda^2 - 1) - 2(\lambda^2 - 1) \right]$$

We collect the terms in the square brackets:

$$0 = \lambda^2 \left[ \lambda^4 - 4\lambda^2 + 2 \right] \quad \Leftrightarrow \quad \lambda^2 = 0 \vee \left[ (\lambda^2)^2 - 4(\lambda^2) + 2 \right] = 0$$

The second equation has the solutions $\lambda^2 = 2 \pm \sqrt{2}$.

Thus, the (ordered) spectrum of $G_1$ is:

$$-\sqrt{2+\sqrt{2}},\ -\sqrt{2-\sqrt{2}},\ 0,\ 0,\ \sqrt{2-\sqrt{2}},\ \sqrt{2+\sqrt{2}}$$

**Spectrum of $G_2$**

We solve $\det(A(G_2) - \lambda \mathbb{1}_6) = 0$ for $\lambda$:

$$0 = \det \begin{pmatrix} -\lambda & 2 & 2 & 0 & 0 & 0 \\ 2 & -\lambda & 1 & 1 & 0 & 0 \\ 2 & 1 & -\lambda & 1 & 0 & 0 \\ 0 & 1 & 1 & -\lambda & 2 & 0 \\ 0 & 0 & 0 & 2 & 2-\lambda & 0 \\ 0 & 0 & 0 & 0 & 0 & 4-\lambda \end{pmatrix} = (4-\lambda) \det \begin{pmatrix} -\lambda & 2 & 2 & 0 & 0 \\ 2 & -\lambda & 1 & 1 & 0 \\ 2 & 1 & -\lambda & 1 & 0 \\ 0 & 1 & 1 & -\lambda & 2 \\ 0 & 0 & 0 & 2 & 2-\lambda \end{pmatrix},$$

where we again expanded along the sixth column. This was possible because node 6 is (as in $G_1$) isolated from the rest of the graph – we will elaborate on that after this calculation, and again in the coming section on regular graphs.

We now employ another technique that is possible for regular graphs (which we will formally validate in section 2.2, too). Since every node in a $d$-regular graph has degree $d$, all the rows (and columns) sum up to $d$. In the determinant for the eigenvalue problem, the corresponding sum also contains a negative $\lambda$ and therefore equals $(d - \lambda)$. Thus, if we add columns 1 to 4 to the fifth one, we get a value of $(4 - \lambda)$ for every one of its components, and can extract this factor as per corollary B.25 (p. 88):

$$\cdots = (4-\lambda) \det \begin{pmatrix} -\lambda & 2 & 2 & 0 & 4-\lambda \\ 2 & -\lambda & 1 & 1 & 4-\lambda \\ 2 & 1 & -\lambda & 1 & 4-\lambda \\ 0 & 1 & 1 & -\lambda & 4-\lambda \\ 0 & 0 & 0 & 2 & 4-\lambda \end{pmatrix} = (4-\lambda)^2 \det \begin{pmatrix} -\lambda & 2 & 2 & 0 & 1 \\ 2 & -\lambda & 1 & 1 & 1 \\ 2 & 1 & -\lambda & 1 & 1 \\ 0 & 1 & 1 & -\lambda & 1 \\ 0 & 0 & 0 & 2 & 1 \end{pmatrix}$$

Before the next expansion, we add row 4 to row 1 (scaled with a factor of $(-2)$):

$$\cdots = (4-\lambda)^2 \det \begin{pmatrix} -\lambda & 0 & 0 & 2\lambda & -1 \\ 2 & -\lambda & 1 & 1 & 1 \\ 2 & 1 & -\lambda & 1 & 1 \\ 0 & 1 & 1 & -\lambda & 1 \\ 0 & 0 & 0 & 2 & 1 \end{pmatrix}$$

We expand along the fifth row:

$$\cdots = (4-\lambda)^2 \left[ (-2) \det \begin{pmatrix} -\lambda & 0 & 0 & -1 \\ 2 & -\lambda & 1 & 1 \\ 2 & 1 & -\lambda & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} + \det \begin{pmatrix} -\lambda & 0 & 0 & 2\lambda \\ 2 & -\lambda & 1 & 1 \\ 2 & 1 & -\lambda & 1 \\ 0 & 1 & 1 & -\lambda \end{pmatrix} \right]$$

Both these $4 \times 4$ determinants can be expanded along their respective first rows:

$$\begin{aligned} \cdots &= (4-\lambda)^2 \left[ (-2) \left( (-\lambda) \det \begin{pmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 1 \\ 1 & 1 & 1 \end{pmatrix} + \det \begin{pmatrix} 2 & -\lambda & 1 \\ 2 & 1 & -\lambda \\ 0 & 1 & 1 \end{pmatrix} \right) \right. \\ &\quad \left. + \left( (-\lambda) \det \begin{pmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 1 \\ 1 & 1 & -\lambda \end{pmatrix} - 2\lambda \det \begin{pmatrix} 2 & -\lambda & 1 \\ 2 & 1 & -\lambda \\ 0 & 1 & 1 \end{pmatrix} \right) \right] \\ &= (4-\lambda)^2 \left[ 2\lambda \det \begin{pmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 1 \\ 1 & 1 & 1 \end{pmatrix} - 2(\lambda+1) \det \begin{pmatrix} 2 & -\lambda & 1 \\ 2 & 1 & -\lambda \\ 0 & 1 & 1 \end{pmatrix} \right. \\ &\quad \left. - \lambda \det \begin{pmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 1 \\ 1 & 1 & -\lambda \end{pmatrix} \right] \end{aligned}$$

For the third determinant, we can use our formula 2.23 (p. 15) with $a = (-\lambda)$ and $b = 1$ and obtain: $(-\lambda - 1)^2(2 - \lambda) = (\lambda + 1)^2(2 - \lambda)$.

In the first determinant, we add the negative of column 3 each of the other two columns; and in the second, we add the negative of row 3 to each of the other two rows:

$$\begin{aligned} 0 \;=\; & (4-\lambda)^2 \left[ 2\lambda \det \begin{pmatrix} -(\lambda+1) & 0 & 1 \\ 0 & -(\lambda+1) & 1 \\ 0 & 0 & 1 \end{pmatrix} - 2(\lambda+1) \det \begin{pmatrix} 2 & -(\lambda+1) & 0 \\ 2 & 0 & -(\lambda+1) \\ 0 & 1 & 1 \end{pmatrix} \right] \\ & -(4-\lambda)^2 \left[ \lambda(\lambda+1)^2(2-\lambda) \right] \end{aligned}$$

The first of these determinants evaluates to $(\lambda+1)^2$ (upper triangular matrix). The second determinant can be expanded along its first column; this yields $4(\lambda+1)$. Thus, we have:

$$\begin{aligned} 0 &= (4-\lambda)^2\left[(2\lambda)(\lambda+1)^2 - 8(\lambda+1)^2 - \lambda(\lambda+1)^2(2-\lambda)\right] \\ &= (4-\lambda)^2(\lambda+1)^2\left[(2\lambda) - 8 - \lambda(2-\lambda)\right] \\ &= (4-\lambda)^2(\lambda+1)^2\left[\lambda^2 - 8\right], \end{aligned}$$

and the spectrum of $G_2$ therefore is:

$$-2\sqrt{2},\ -1,\ -1,\ 2\sqrt{2},\ 4,\ 4$$

**Graphs with Several Connected Components**

Before we expand on regular graphs, we would like to point out one aspect that applies to all undirected graphs, whether regular or not:

**Claim 2.24** *If an undirected graph $G = (V, e)$ has $k$ connected components, its spectrum may be calculated separately for each of those components. The overall spectrum is the combination of the $k$ sub-spectra.*

Proof: We recall definition 2.15 (p. 14). Since the spectrum of $G$ is invariant under re-labeling of the nodes (cf. lemma 2.22, p. 15), we may choose a labeling where each connected component corresponds to a contiguous slice of the numbers $\{1, \cdots, |V|\}$. This implies an ordering of the components (e.g. by their smallest node label), such that component 1 has node labels $\{1, \cdots, j_1\}$, component 2 has $\{(j_1+1), \cdots, j_2\}$, etc., and component $k$ has $\{(j_{k-1}+1), \cdots, j_k\}$, where $j_k = |V|$.

Since there are no edges connecting different connected components of $G$, the edge function $e$ can only yield non-zero values if both its arguments belong to the same slice of node labels. But this implies that the adjacency matrix $A(G)$ is block-diagonal; and the same holds for $(A(G) - \lambda\mathbb{1}_n)$.

Taking the determinant of the latter matrix (i.e. solving the eigenvalue problem for $A(G)$), we may use lemma B.48 (p. 98) recursively, and find that the determinant equals the product of all the determinants of the diagonal blocks. This means that the characteristic polynomial of $A(G)$ separates into the $k$ polynomial factors obtained from the block matrices. Therefore it is irrelevant if we solve for the roots of the combined polynomial or if we combine the roots of the separate $k$ polynomials. ■

(This is no real surprise, because each of the connected components of $G$ may also be viewed as a separate graph itself – apart from labeling concerns, this is just a matter of perspective.)

## 2.2 Regular Graphs

Before we give some further examples of regular graphs (cf. definition 2.12, p. 13) in the next section, we establish some general characteristics, including their spectra, and introduce the notion of edge expanders.

### 2.2.1 Basic Properties of Regular Graphs

First, we show a few example graphs (without node labels) for some values for $d$.

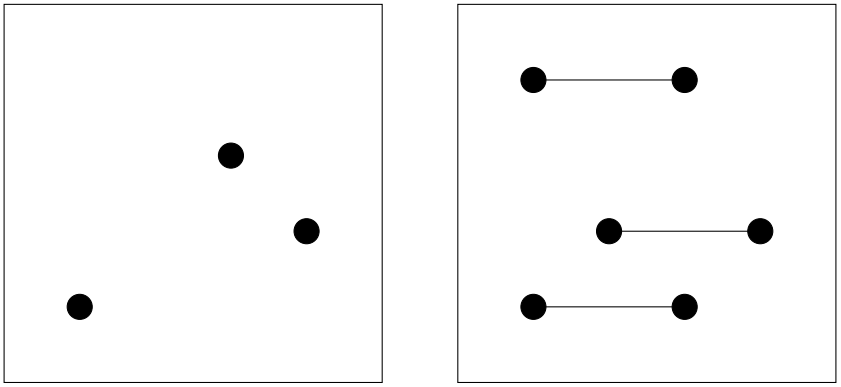

Figure 2.7: Regular graphs with $d = 0$ (left) and $d = 1$ (right)

Figure 2.7 shows that 0-regular graphs are just isolated nodes. 1-regular graphs must consist of pairs of nodes connected by one edge each (every node must be connected to one other node, but to none other beyond that).

Figure 2.8 demonstrates that 2-regular graphs may be isolated nodes with loops, pairs connected with double edges, or closed cycles like polygons.

For 3-regular graphs, there are more possibilities. Cycles of nodes may be connected or have additional spokes. Also, some three-dimensional structures may be constructed. The lower graph is actually a flattened version of a tetrahedron. Cubes and dodecahedra could be constructed in this way, too. The figure does not show all possibilities (for instance, pairs of nodes connected with triple edges are not depicted).

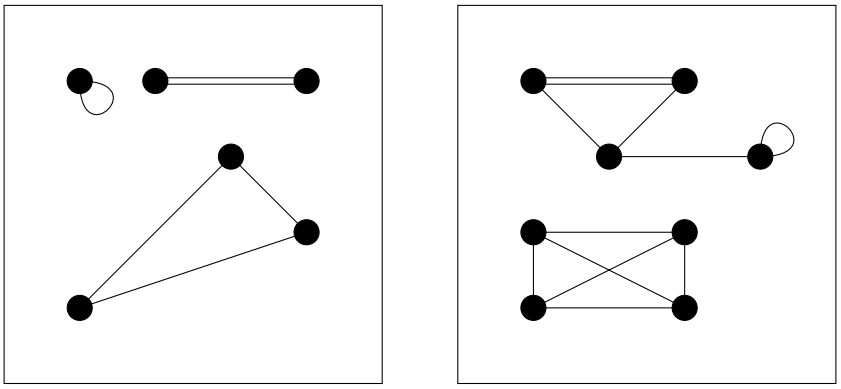

Figure 2.8: Regular graphs with $d = 2$ (left) and $d = 3$ (right)

An example for a 4-regular graph was presented above: The example graph $G_2$ (figure 2.6, p. 13).

One important feature of regular graphs is their fixed ratio of node and edge counts:

**Lemma 2.25** *An undirected $d$-regular graph $G = (V, e)$ with $n := |V|$ nodes has exactly*

$$\frac{nd}{2}$$

*edges.*

Proof: Each node is of degree $d$ and therefore has $d$ half-edges, according to definition 2.1 (p. 9). This makes for $n \cdot d$ half-edges. In order for $G$ to be a graph, all those must pair up to edges, which yields the stated number of edges. ■.

**Corollary 2.26** *If a graph $G = (V, e)$ is $d$-regular for odd $d$, its node count $n := |V|$ is even.*

Proof: This is a special case of corollary 2.3 (p. 11): In a $d$-regular graph with odd $d$, all the $n$ nodes have odd degree; therefore the node count has to be even. ■

### 2.2.2 Intermezzo: Regular Bipartite Graphs

We recall from definition 2.19 (p. 14) that bipartite graphs are 2-colorable. Often, such graphs are depicted with one part of the nodes (color 1) on the left-hand side, and the other (color 2) on the right-hand side. In that case, every edge must be between a left-hand and a right-hand node. We show the complete[3] bipartite graph $K_{5,3}$ (with five nodes of color 1 and three of color 2) as an example:

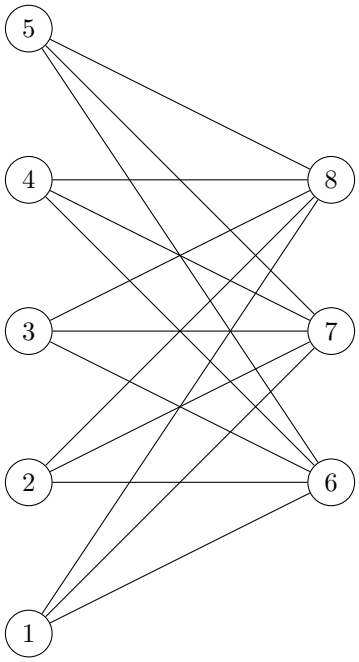


Figure 2.9: The bipartite graph $K_{5,3}$

We observe that every node on the left-hand side has degree 3, and every node on the right-hand side degree 5, respectively. This graph is *not* regular. But we may infer from this example the following

**Lemma 2.27** *A bipartite undirected graph $G = (V, e)$, $n := |V|$, can only be regular if $n$ is even, and if $G$ has $n/2$ nodes of each color. If it is $d$-regular and simple, then $d \leq n/2$.*

Proof: First, we observe that it is of course possible to construct bipartite graphs with an even node color distribution that are not regular (e.g. the complete bipartite graph $K_{d,d}$, with a single edge removed).

However, if $G$ is to be $d$-regular, each node must be incident with exactly $d$ half-edges that belong to edges connecting it with nodes of the other respective color. If there are $j$ nodes of

[3] "complete" to be understood in the sense that adding another edge would produce either a bipartite multi-graph, or destroy its bipartite property by connecting two nodes of the same color

color 1 and $k$ nodes of color 2, this implies $d \cdot j$ half-edges incident at color-1 nodes, and $d \cdot k$ half-edges incident at color-2 nodes. Since $G$ is bipartite, any edge must connect a color-1 node and a color-2 node – but this is only possible if the respective total numbers of half-edges are equal, i.e. $d \cdot j = d \cdot k$, which means $j = k = n/2$.

Also, for a simple bipartite $d$-regular graph, no more than $n/2$ edges can be incident at a single node because there are exactly $n/2$ nodes of each color and $G$ (being simple) has no multi-edges. Therefore $d \leq n/2$. ■

**Corollary 2.28** *The complete simple bipartite graphs $K_{d,d}$ (with $d$ nodes of each color) are $d$-regular.*

Proof: Each of the $d$ color-1 nodes is connected to each of the $d$ color-2 nodes by a single edge. ■

We now give a construction for regular simple bipartite graphs:

**Lemma 2.29** *Let $V := \{1, \cdots, n\}$ for an even $n \in \mathbb{N}$, and $0 < d \leq n/2$, then there is a $d$-regular simple bipartite graph $G = (V, e)$.*

Proof: Let the nodes $\{1, \cdots, n/2\}$ be of color 1, and $\{n/2+1, \cdots, n\}$ of color 2, respectively. Connect each color-1 node $j$ with each of the color-2 nodes $\{k_1(j), \cdots, k_d(j)\}$, where

$$k_r(j) := \left[((j-1) + (r-1)) \bmod \left(\frac{n}{2}\right)\right] + 1 + \frac{n}{2}$$

The subtraction from $j$ shifts the $j$ range to $\{0, \cdots, n/2-1\}$, which are the residues modulo $(n/2)$. Then, we add a shift of $\{0, \cdots, d-1\}$ depending on $r$. The modulo operation projects the result back into the $j$ range, so that the square bracket in the above formula amounts to a cyclic shift (an invertible operation). After that, we add an offset of 1 to re-adjust the label range back to $\{1, \cdots, n/2\}$, and add $(n/2)$ to take us to the color-2 range.

In other words, we perform a cyclic shift of the tuple of color-1 nodes (by a distance of $r$) and assign the shifted components to the tuple of color-2 nodes. This ensures that every color-1 node will be connected to $d$ different color-2 nodes in a uniform way. Since the above mapping is invertible, each of the color-2 nodes will also be connected to $d$ color-1 nodes, and the resulting graph is simple and $d$-regular. ■

As an example, we show the 3-regular construction for 12 nodes (observe the cyclic shift effect for $j = 5$ and $j = 6$):

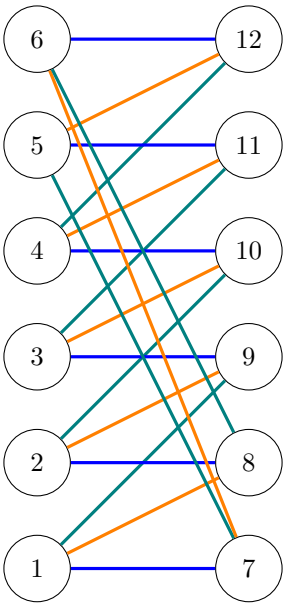


Figure 2.10: A 3-regular bipartite graph with 12 nodes

If we drew the graph on a cylindrical surface, with the color-1 nodes in the lower half and the color-2 nodes in the upper half, the picture would be translation-symmetric (nodes 6 and 12 would be next to nodes 5 and 11, respectively, as in the above figure, but also next to nodes 1 and 7).

**Corollary 2.30** *The construction in lemma 2.29 is d-edge-colorable.*

Proof: Assign edge color $r$ to each edge between $j$ and $k_r(j)$. Because of the unified construction, each color-1 node is then incident with edges colored from 1 to $d$. Since the construction's mapping $j \mapsto k_r(j)$ is a bijection and could be formulated equivalently in the other direction, each color-2 node is also incident with $d$ edges of pairwise-different edge colors:

- Edge color 1: Between $j$ and $j+n/2$, or, equivalently, between $k-n/2$ and $k$ (modulo range corrections)
- Edge color 2: Between $j$ and $j+1+n/2$, or, equivalently, between $k-1-n/2$ and $k$ (modulo range corrections)
- $\cdots$
- Edge color d: Between $j$ and $j+(d-1)+n/2$, or, equivalently, between $k-(d-1)-n/2$ and $k$ (modulo range corrections)

■

We show the different edge colors resulting from the above example:

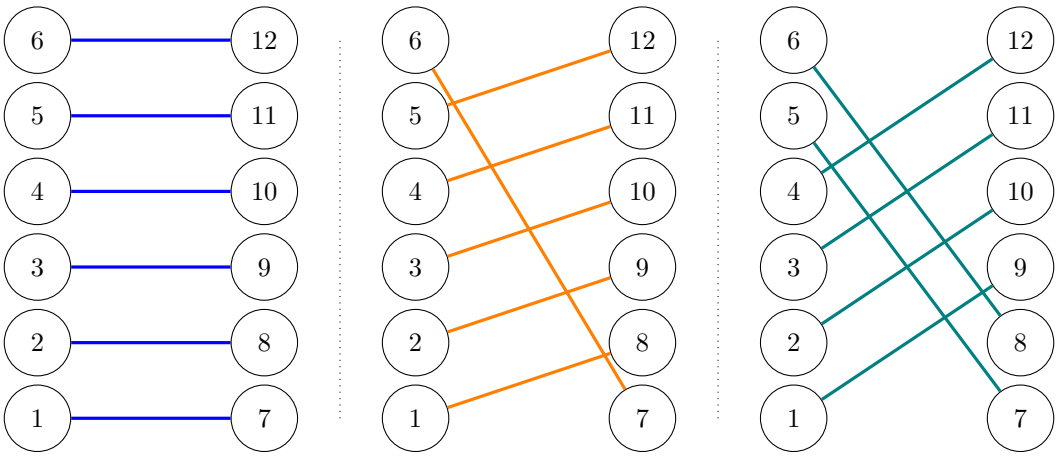


Figure 2.11: The previous graph. Edges with colors 1 (left), 2 (middle) and 3 (right)

We would like to point out that the above construction will in no way generate *all* $d$-regular bipartite graphs. In the main chapter 3, we will introduce in section 3.4, a more general way to build the complete set of $d$-regular $d$-edge-colorable bipartite graphs.

### 2.2.3 Spectral Properties of Regular Graphs

**Lemma 2.31** *The spectrum of an undirected d-regular graph $G = (V, e)$, $n := |V|$, always contains the eigenvalue d; and a vector comprised of n components with value 1, $\vec{v} := (1, \cdots, 1)^T$, is an eigenvector of the adjacency matrix $A := A(G)$ to that eigenvalue.*

Proof: Since the graph is $d$-regular, every node has degree $d$. As per definition 2.11 (p. 13), the components of any row of $A$ sum up to $d$. Consider the equation of the eigenvalue problem (definition C.1, p. 100) with that vector:

$$A \cdot \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix} = \lambda \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix}$$

If we pick the $j$-th component of the eigenvector, the equation is

$$\left[\lambda \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix}\right]_j = \lambda = \left[A \cdot \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix}\right]_j = \sum_k A_{jk} \left[\begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix}\right]_k = \sum_k A_{jk} = d \quad \blacksquare$$

**Corollary 2.32** *The spectrum of an undirected d-regular graph G with k connected components contains at least k times the eigenvalue d.*

Proof: Any connected component of $G$ can be viewed as a separate $d$-regular graph, whose spectrum features an eigenvalue $d$ as per lemma 2.31. Thus, and per claim 2.24 (p. 18), each of the components contributes an eigenvalue $d$ to the overall spectrum of $G$. ■

**Corollary 2.33** *If the spectrum of an undirected d-regular graph G contains the eigenvalue d just once, G is a connected graph.*

Proof: If $G$ had more than one connected components, its spectrum would feature the eigenvalue $d$ at least twice, as per corollary 2.32. ■

Next, we follow a proof in [BM11] (p. 44) to demonstrate that the spectrum of a connected $d$-regular graph (i.e. a graph with only one connected component) contains the eigenvalue $d$ exactly once (not more), and that all other eigenvalues are smaller than $d$. In the above corollary 2.32, the lower bound $k$ then becomes an upper bound as well, because a $d$-regular graph with $k$ connected components cannot have more than $k$ times the eigenvalue $d$.

**Lemma 2.34** *If an undirected d-regular graph $G = (V, e)$, $n := |V|$, with an adjacency matrix $A := A(G)$, is connected, then d is the largest eigenvalue of $A(G)$, and occurs only once.*

Proof: Let $\lambda$ be an eigenvalue of $A$ and $\vec{v}$ an eigenvector of $A$ for $\lambda$. We may assume that $\vec{v}$ scaled in such a way that its largest component $v_j$ (for some $j \in \{1, \cdots, n\}$) equals 1. This is possible due to corollary C.2 (p. 100).

We now consider the product of row $j$ of $A$ with $\vec{v}$, remembering that $v_j = 1$, that $v_k \leq v_j$ for all $k$, and that the sum of any row's components in $A$ is $d$, according to definitions 2.11 (p. 13) and 2.12 (p. 13):

$$\lambda = \lambda \cdot 1 = \lambda \cdot v_j = (A\vec{v})_j = \sum_k A_{jk} v_k \leq \sum_k A_{jk} v_j = \sum_k A_{jk} = d$$

Thus, any eigenvalue is less or equal to $d$.

We now recall that $A$ is symmetric, therefore all its eigenspaces are orthogonal according to corollary C.11 (p. 103), and have full dimension (i.e. an eigenspace features as many linearly independent eigenvectors as the algebraic multiplicity of its eigenvalue permits).

For our connected graph $G$, let $\vec{w}$ be an eigenvector of $A$ to the eigenvalue $d$. As above, we may scale that vector, so that its largest component $w_j$ is 1. But then:

$$d = d \cdot 1 = d \cdot w_j = (A\vec{w})_j = \sum_k A_{jk} w_k \leq \sum_k A_{jk} w_j = \sum_k A_{jk} = d$$

But since every $w_k$ in $\vec{w}$ is at most 1, and none of the components of $A$ is negative, and the sum of a row's components in $A$ is exactly $d$, we can only achieve the "equal" case in "$\leq$" if all the $w_k$ have maximum value, i.e. if they are all 1.

Thus, if $\vec{w}$ is eigenvector to $d$, it follows that it is a multiple of $(1, 1, \cdots, 1)^T$. Thus, the eigenspace of eigenvalue $d$ must be one-dimensional, and the eigenvalue $d$ can occur only once. ■

We combine the results from above:

**Corollary 2.35** *The adjacency matrix A of a connected undirected d-regular graph G always has an eigenvalue of d, which occurs exactly once and is the largest eigenvalue of the spectrum.*

If we denote the second-largest eigenvalue of a connected $d$-regular graph $G$ by $\lambda_2(G)$, we can show the following:

**Lemma 2.36** *For a connected undirected d-regular graph $G = (V, e)$, $n := |V|$, with second-largest eigenvalue $\lambda_2$ and adjacency matrix A:*

$$\lambda_2 = \max_{\vec{0} \neq \vec{x} \perp (1, \cdots, 1)^T} \left\{ \frac{\langle A \cdot \vec{x}, \vec{x} \rangle}{\langle \vec{x}, \vec{x} \rangle} \right\} = \max_{\vec{0} \neq \vec{x} \perp (1, \cdots, 1)^T} \left\{ \frac{\vec{x}^T A \vec{x}}{\vec{x}^T \vec{x}} \right\}$$

Proof: Since $A$ is a symmetric matrix, there is (as per theorem C.10 (p. 102)) an orthogonal basis of $\mathbb{R}^n$ consisting of $n$ eigenvectors. We may assume that all those vectors $\vec{v}_1, \cdots, \vec{v}_n$ are normalized, so $\langle \vec{v}_j, \vec{v}_k \rangle = \delta_{jk}$. Further, let $\vec{v}_1 := (1, \cdots, 1)^T/\sqrt{n}$ and let the spectrum be ordered per

$$\lambda_n \leq \cdots \leq \lambda_2 < \lambda_1 = d$$

Since the $\vec{v}$ vectors constitute an orthonormal basis, we may project any vector $\vec{x}$ onto those vectors to determine its components:

$$\vec{x} = \sum_j x_j \vec{v}_j$$

If we consider the two scalar products in the quotient, we observe that the quotient is independent of $\vec{x}$'s length, so we may also assume that

$$||\vec{x}|| = \sqrt{\sum_j x_j^2} = 1$$

We now calculate the quotient:

$$\frac{\langle A \cdot \vec{x}, \vec{x} \rangle}{\langle \vec{x}, \vec{x} \rangle} = \frac{\vec{x}^T A^T \vec{x}}{||\vec{x}||^2} = \vec{x}^T A \vec{x} = \sum_j x_j \vec{v}_j^T \cdot A \cdot \sum_k x_k \vec{v}_k = \sum_{j,k} x_j x_k \vec{v}_j^T \cdot A \cdot \vec{v}_k$$

$\vec{v}_k$ is an eigenvector of $A$ to the eigenvalue $\lambda_k$:

$$\cdots = \sum_{j,k} x_j x_k \lambda_k \vec{v}_j^T \cdot \vec{v}_k = \sum_{j,k} x_j x_k \lambda_k \langle \vec{v}_j, \vec{v}_k \rangle = \sum_{j,k} x_j x_k \lambda_k \delta_{jk} = \sum_j \lambda_j x_j^2$$

Now, if we only consider vectors $\vec{x} \perp \vec{v}_1$, then $x_1 = 0$, because the $\vec{v}$ basis is orthogonal. Also, all other eigenvalues $\lambda_j$ are bounded by $\lambda_2$:

$$\cdots = \sum_{j=2}^{n} \lambda_j x_j^2 \leq \lambda_2 \sum_{j=2}^{n} x_j^2 = \lambda_2$$

Thus, the quotient cannot exceed the value $\lambda_2$ – but it can reach it: If we consider the vector $\vec{x} := \vec{v}_2$, then $x_2 = 1$ and all other $x_j$ are zero, so that the quotient is exactly $\lambda_2$. Thus, the maximum of quotient values is indeed $\lambda_2$, the second-largest eigenvalue of $A$. ■

### 2.2.4 Edge Expanders (Definition)

Expander graphs are characterized by high connectivity: Cutting such graphs into two parts (in any way) requires the deletion of a number of edges – the more edges, the better the "expansion". In preparation of the construction algorithm [ASS08], we introduce a quantity that is variously known as "Edge Expansion", "Cheeger Constant" or "Isoperimetric Constant", depending on the source. It is called "Expansion Parameter" in [Sta17]. Because Alon et al. reserve the term "edge expansion" (see below) for a normalized version of this quantity, we will introduce it as "isoperimetric constant":

**Definition 2.37** *For an undirected graph $G = (V, e)$, $n := |V|$, the* Isoperimetric Constant $\mathcal{I}(G)$ *is defined per*

$$\mathcal{I}(G) := \min_{\substack{S \subset V \\ 0 < |S| \leq \frac{n}{2}}} \left\{ \frac{e(S, \bar{S})}{|S|} \right\},$$

*where $\bar{S} := V \setminus S$ is the complement of the subset $S$, and $e(S, \bar{S})$ is the number of edges between $S$ and $\bar{S}$, i.e. the number of all the edges connecting one node in $S$ with one node in $\bar{S}$.*

The isoperimetric constant can be calculated for any undirected graphs, not only for regular ones. We avoided mentioning it in the basics section because we will mainly use it in its relation to expander graphs. The next section will feature several example graphs, for which we have calculated the value of $\mathcal{I}$ with a simple Java program (cf. chapter F, p. 117, for details).

**Corollary 2.38** *The isoperimetric constant of a given graph $G = (V, e)$ is zero if and only if $G$ is not connected.*

Proof: If $G$ is not connected, it contains at least two disjoint connected components (cf. definition 2.15, p. 14, for details). Let one of those be the subset $S \subset V$. Then, $\bar{S} = V \setminus S$ contains all the other connected components of $G$, none of which can be reached from $S$ by traversing an edge: $e(S, \bar{S}) = 0$.

If $G$ is connected, then, per definition 2.13 (p. 13), all its nodes are indirectly connected to each other, and any subset of nodes will be connected to its complement by at least one edge. Since all the fractions in definition 2.37 therefore are positive (zero cannot be reached), the minimum of those fractions must be nonzero, too. ■

**Corollary 2.39** *For an undirected graph $G = (V, e)$, the isoperimetric constant $\mathcal{I}(G)$ satisfies*

$$\mathcal{I}(G) \leq \max_{j \in V} \{\deg(j)\}$$

*If $G$ is $d$-regular, then $\mathcal{I}(G) \leq d$.*

Proof: The fractions in definition 2.37 are largest if all the edges of a subset $S$ of nodes are connected to its complement. This edge count cannot be more than the sum of node degrees (cf. definition 2.11, p. 13), which is at most $|S|$ times the maximum node degree. ■

We will now restrict ourselves to regular graphs again, and introduce the notation used in [ASS08]:

**Definition 2.40** *An undirected $d$-regular Graph $G = (V, e)$, $n := |V|$, is called a* $\delta$-Edge-Expander*, denoted as an* $[n, d, \delta]$-Expander*, where*

$$\delta := \frac{\mathcal{I}(G)}{d}$$

**Corollary 2.41** *For an $[n, d, \delta]$-expander, the expansion $\delta$ satisfies $0 \leq \delta \leq 1$.*

Proof: This follows directly from the normalization in definition 2.40 and the corollaries 2.38 and 2.39. ■

## 2.3 Examples of Regular Graphs

We will first look at a popular cubic (i.e. 3-regular) graph devised by J. Petersen (cf. [Sta17], p. 21), which has comparatively good expander properties, and a contrasting example with poor expansion. After that, we present some complete, and some bipartite graphs. All graphs in this section will be regular, i.e. their nodes all have equal degree.

In this section, we will denote the adjacency matrices with $M$ instead of $A$, because the letters $A$ do $D$ will be needed several times for sub-matrices.

### 2.3.1 The Petersen Graph

This is a cubic graph ($d = 3$) with ten nodes. As per lemma 2.25 (p. 20), it has $3 \cdot 10/2 = 15$ edges; see figure 2.12:

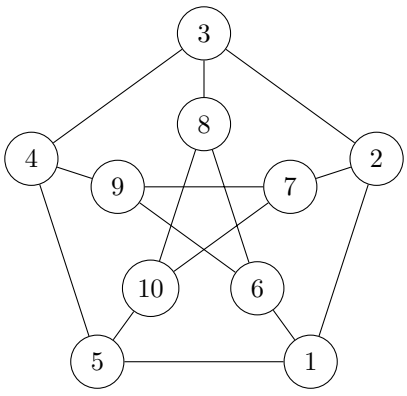


Figure 2.12: The Petersen graph

The corresponding adjacency matrix (written as a table) is:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 3 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 4 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 5 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 6 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 7 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 8 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 9 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |

Table 2.1: Adjacencies of the Petersen graph

We calculate the spectrum of $M$, i.e. we solve $0 \stackrel{!}{=} \det(M - \lambda \mathbb{1}_{10}) =: \det M_\lambda$. Instead of solving directly, using ten-dimensional Laplace expansion, we exploit the high degree of symmetry in $M$, which was already hinted at by the table lines:

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \qquad \text{with} \qquad B = C = \mathbb{1}_5$$

We employ a formula by J. Silvester [Sil00] (theorem 3): If $M$ is an $(2n) \times (2n)$ matrix with $n \times n$ blocks $A, B, C, D$ as written above, and if $CD = DC$ (i.e. $C, D$ commute, as per definition A.24, p. 76), then $\det M = \det(AD - BC)$. Since the unit matrix $C = \mathbb{1}_5$ certainly commutes with $D$, the formula applies. This still holds if we deduct $\lambda \mathbb{1}_{10}$: $M_\lambda$ has similar block structure with matrices $A_\lambda$ and $D_\lambda$, but $B, C$ unchanged. Thus, $\det M_\lambda = \det(A_\lambda D_\lambda - \mathbb{1}_5)$, because $BC = \mathbb{1}_5$.

We calculate the other matrix product:

$$\begin{aligned} A_\lambda D_\lambda &= \begin{pmatrix} -\lambda & 1 & 0 & 0 & 1 \\ 1 & -\lambda & 1 & 0 & 0 \\ 0 & 1 & -\lambda & 1 & 0 \\ 0 & 0 & 1 & -\lambda & 1 \\ 1 & 0 & 0 & 1 & -\lambda \end{pmatrix} \cdot \begin{pmatrix} -\lambda & 0 & 1 & 1 & 0 \\ 0 & -\lambda & 0 & 1 & 1 \\ 1 & 0 & -\lambda & 0 & 1 \\ 1 & 1 & 0 & -\lambda & 0 \\ 0 & 1 & 1 & 0 & -\lambda \end{pmatrix} \\ &= \begin{pmatrix} \lambda^2 & 1-\lambda & 1-\lambda & 1-\lambda & 1-\lambda \\ 1-\lambda & \lambda^2 & 1-\lambda & 1-\lambda & 1-\lambda \\ 1-\lambda & 1-\lambda & \lambda^2 & 1-\lambda & 1-\lambda \\ 1-\lambda & 1-\lambda & 1-\lambda & \lambda^2 & 1-\lambda \\ 1-\lambda & 1-\lambda & 1-\lambda & 1-\lambda & \lambda^2 \end{pmatrix} \end{aligned}$$

This means that we can use the formula from lemma 2.23 (p. 15). Subtracting $\mathbb{1}_5$ from the above, we obtain the $a, b$ pattern with $a = (\lambda^2 - 1)$ and $b = (1 - \lambda)$. Together with $n = 5$, this yields

$$0 = \det M_\lambda = (a - b)^4 \cdot [a + 4b]$$

We determine the expressions in brackets:

$$\begin{aligned} a - b &= \lambda^2 + \lambda - 2 = (\lambda - 1)(\lambda + 2) \\ a + 4b &= \lambda^2 - 4\lambda + 3 = (\lambda - 1)(\lambda - 3) \end{aligned}$$

And thus

$$0 = (\lambda - 1)^5 (\lambda + 2)^4 (\lambda - 3),$$

yielding a spectrum of

$$-2,\ -2,\ -2,\ -2,\ 1,\ 1,\ 1,\ 1,\ 1,\ 3$$

The algorithm presented in chapter F returns an isoperimetric constant (cf. definition 2.37, p. 24) of 1: the inner set of nodes $\{6, \cdots, 10\}$ is connected to the outer set $\{1, \cdots, 5\}$ by five edges, hence the quotient. Because the Petersen graph is cubic, this means that, according to definition 2.40 (p. 25), it is a $[10, 3, (1/3)]$-expander.

### 2.3.2 An Example with Poor Expansion

The following graph also has ten nodes and is cubic, but has decidedly poorer expansion. In keeping with the numbering of the examples in section 2.1, we will call this example $G_3$; see figure 2.13:

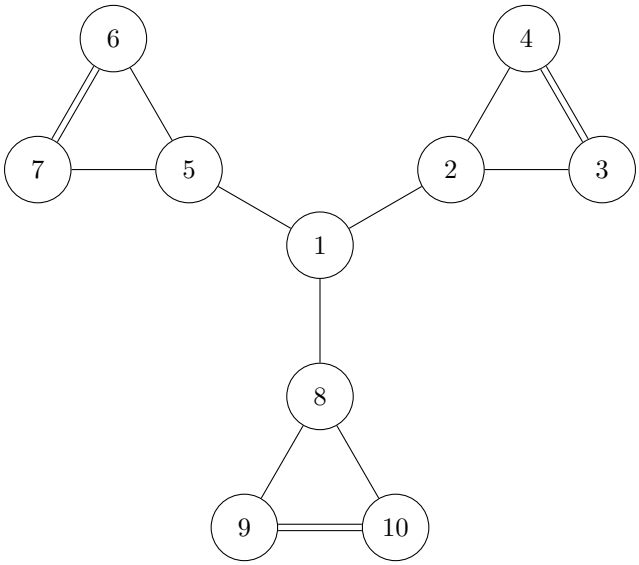


Figure 2.13: Example graph $G_3$

The adjacency table (again, suggestively spaced in anticipation of the spectrum calculation) is as follows:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 2 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 |
| 8 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 |

Table 2.2: Adjacencies of the Example graph $G_3$

This graph's adjacency matrix $M$ does not exhibit a block structure suitable for Sylvester's theorem, but the three-fold symmetry is clearly visible: If we deleted the edges incident at node 1, we would have three pairwise-disconnected graphs with identical structure; thus, $M$ is almost block-diagonal. We will use the shorthand $A$ for the $3 \times 3$ blocks on the diagonal, and $A_\lambda := A - \lambda \mathbb{1}_3$ like in the previous example. Also, we define $\vec{e} := (1, 0, 0)^T$; thus

$$M_\lambda = \begin{pmatrix} -\lambda & \vec{e}^T & \vec{e}^T & \vec{e}^T \\ \vec{e} & A_\lambda & 0 & 0 \\ \vec{e} & 0 & A_\lambda & 0 \\ \vec{e} & 0 & 0 & A_\lambda \end{pmatrix}$$

For the determinant calculation, we first add the third and fourth block-columns to the second one (each block addition consisting of three proper matrix column additions). After that, we add the second block row, scaled with $(-1)$ to the third and fourth:

$$0 = \det M_\lambda = \det \begin{pmatrix} -\lambda & 3\vec{e}^T & \vec{e}^T & \vec{e}^T \\ \vec{e} & A_\lambda & 0 & 0 \\ \vec{e} & A_\lambda & A_\lambda & 0 \\ \vec{e} & A_\lambda & 0 & A_\lambda \end{pmatrix} = \det \begin{pmatrix} -\lambda & 3\vec{e}^T & \vec{e}^T & \vec{e}^T \\ \vec{e} & A_\lambda & 0 & 0 \\ 0 & 0 & A_\lambda & 0 \\ 0 & 0 & 0 & A_\lambda \end{pmatrix}$$

This matrix has block-upper-triangular form, and we can use lemma B.48 (p. 98) to reduce its determinant to a threefold product:

$$\cdots = \det \begin{pmatrix} -\lambda & 3\vec{e}^T \\ \vec{e} & A_\lambda \end{pmatrix} \cdot (\det A_\lambda) \cdot (\det A_\lambda)$$

We expand the left-hand matrix along the first column to yield the expression in square brackets:

$$\cdots = (\det A_\lambda)^2 \cdot \left[ (-\lambda)(\det A_\lambda) - \det \begin{pmatrix} 3 & 0 & 0 \\ 1 & -\lambda & 2 \\ 1 & 2 & -\lambda \end{pmatrix} \right]$$

This leaves us with just two $3 \times 3$ determinants to calculate. Starting with the one expressed directly in the last equality, and expanding it along the first row, we obtain a contribution of

$$3(\lambda^2 - 4) = 3(\lambda + 2)(\lambda - 2)$$

For $A_\lambda$, we may employ Sarrus's rule (cf. the example in figure B.1, p. 94) after the proof of Leibniz's rule (theorem B.40, pp. 92ff). This yields:

$$\det A_\lambda = \det \begin{pmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 2 \\ 1 & 2 & -\lambda \end{pmatrix} = (-\lambda)^3 + 2 + 2 - (-\lambda) - (-\lambda) - (-4\lambda) = -\lambda^3 + 6\lambda + 4$$

In order to simplify the square bracket expression from above, it would be beneficial if $\det A_\lambda$ had a root of $\pm 2$; and indeed we find that $(-2)$ is a root of $\det A_\lambda$. Employing polynomial division (cf. lemma E.9, p. 108), we may split off the factor belonging to this root:

$$\det A_\lambda = -(\lambda + 2)(\lambda^2 - 2\lambda - 2)$$

Thus, our expression for $\det M_\lambda$ simplifies to

$$\cdots = (\lambda+2)^3(\lambda^2-2\lambda-2)^2 \left[\lambda(\lambda^2 - 2\lambda - 2) - 3(\lambda - 2)\right] = (\lambda+2)^3(\lambda^2-2\lambda-2)^2 \left[\lambda^3 - 2\lambda^2 - 5\lambda + 6\right]$$

For the square bracket term, we may use the fact that $G_3$ is 3-regular, and therefore the characteristic polynomial must have a $\lambda = d = 3$; this does not feature in the factors outside the square bracket, so the corresponding factor must divide the square bracket. Using polynomial division again, we obtain

$$\lambda^3 - 2\lambda^2 - 5\lambda + 6 = (\lambda - 3)(\lambda^2 + \lambda - 2) = (\lambda - 3)(\lambda - 1)(\lambda + 2)$$

We combine the previous results:

$$0 = (\lambda + 2)^4(\lambda^2 - 2\lambda - 2)^2(\lambda - 3)(\lambda - 1)$$

The resulting spectrum, then, is

$$-2,\ -2,\ -2,\ -2,\ 1 - \sqrt{3},\ 1 - \sqrt{3},\ 1,\ 1 + \sqrt{3},\ 1 + \sqrt{3},\ 3$$

The algorithm in chapter F returns an isoperimetric constant of $(1/3)$: If we cut through the (single) edge between nodes 1 and 2, we have separated a set of three nodes from the rest of $G_3$. According to definition 2.40 (p. 25), this makes $G_3$ a $[10, 3, (1/9)]$-expander, with only a third of the expansion of the Petersen graph from the previous subsection.

Generally speaking, having large node clusters with only a few bridging connections among each other will produce a poor expansion because it will only take a few edges to cut in order to separate the connected graph into two components.

### 2.3.3 Complete Graphs

Complete graphs $K_n$ ($n \in \mathbb{N}$, $n > 1$) are simple graphs where each node is connected to every of the respective other $(n-1)$ nodes, making $K_n$ $(n-1)$-regular. We show an example for $n = 9$ in figure 2.14:

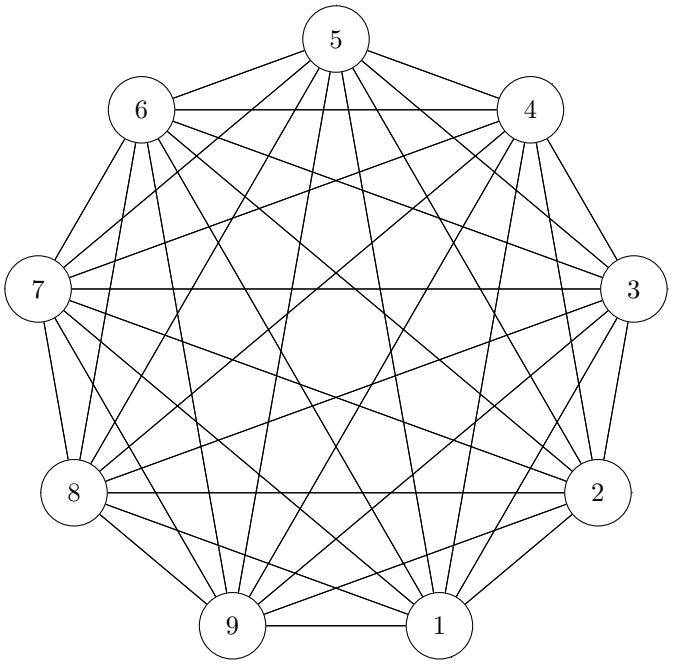


Figure 2.14: The complete graph $K_9$

The adjacency matrix of $K_n$ has zeros on its diagonal, and all off-diagonal components are 1. For the spectrum, we may therefore employ the formula in lemma 2.23 (p. 15), with $a = (-\lambda)$ and $b = 1$:

$$0 = \det M_\lambda = (-\lambda - 1)^{n-1} \cdot [-\lambda + (n-1)] = (-1)^{n-1} \cdot (\lambda + 1)^{n-1}[(n-1) - \lambda]$$

Thus, the spectrum consists of $(n-1)$ times the eigenvalue $(-1)$, and a single eigenvalue of $d = (n-1)$.

For the isoperimetric constant we refer to B. Mohar's article [Moh89], where he states that

$$\mathcal{I}(K_n) = \left\lceil \frac{n}{2} \right\rceil,$$

which we could verify for some small values for $n$ with our Java algorithm from chapter F.

Because of the high degree of symmetry, it is not hard to find a subset of nodes that reaches this minimum ratio of outer edge count to subset node count (cf. definition 2.37, p. 24):

- For even $n$, take $n/2$ nodes to form $S$. The complement $\bar{S}$ will contain $n/2$ nodes, too. Now, each of the nodes in $S$ will have an edge connecting it to any node in $\bar{S}$; this makes for $(n/2)^2$ edges. If we divide this by $|S|$, we obtain $n/2$.

  The expansion as per definition 2.40 (p. 25) is

  $$\delta = \frac{n}{2(n-1)} = \frac{1}{2 - \frac{2}{n}}$$

- For odd $n$, take $(n-1)/2$ nodes to form $S$. The complement $\bar{S}$ will contain $(n+1)/2$ nodes. By the same reasoning, this makes for $(n-1)(n+1)/4$ edges between $S$ and $\bar{S}$, which yields $(n+1)/2$ if divided by $|S|$.

  The expansion evaluates to

  $$\delta = \frac{n+1}{2(n-1)} = \frac{1}{2 - \frac{2}{n+1}}$$

On the face of it, this looks "better" than the Petersen graph for large enough $n$, and it cannot be denied that a sizable number of edges has to be cut in order to divide $K_n$ into two components – but the $K_n$ also have maximum degree for a simple graph, which does not fit the aim of sparse expanders.

In fact, Alon et al. argue[ASS08] that constant degree expanders (i.e. with fixed $d$) are preferable.

### 2.3.4 Complete Regular Bipartite Graphs

We already determined in lemma 2.27 (p. 20) that any bipartite graph that is also regular has an even number of nodes, with half the nodes of color 1, and the other half of color 2. We also stated in corollary 2.28 (p. 21) that the complete bipartite graphs $K_{d,d}$ are $d$-regular. As an example, we show the graph $K_{5,5}$ in figure 2.15:

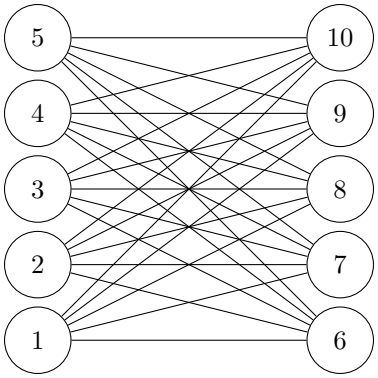


Figure 2.15: The complete regular bipartite graph $K_{5,5}$

For the spectrum, we observe that the adjacency matrix $M$ and the eigenvalue matrix $M_\lambda$ have block structure:

$$M_\lambda = \begin{pmatrix} -\lambda\mathbb{1}_d & 1 \\ 1 & -\lambda\mathbb{1}_d \end{pmatrix} =: \begin{pmatrix} A_\lambda & B \\ C & D_\lambda \end{pmatrix},$$

where $B$ and $C$ are $d \times d$-matrices consisting entirely of 1. The matrices $C$ and $D\lambda$ commute because

$$CD_\lambda = -\lambda C\mathbb{1}_d = -\lambda\mathbb{1}_d C = D_\lambda C$$

This means we may employ Silvester's formula (theorem 3 in [Sil00]) again (as we did with the Petersen graph), and obtain $\det M_\lambda = \det(A_\lambda D_\lambda - BC)$.

Now, $A_\lambda D_\lambda$ is just $\lambda^2\mathbb{1}_d$, and $BC$ is $d{\cdot}B$, $d$ times a matrix consisting only of 1. Thus, $A_\lambda D_\lambda - BC$ is a matrix with $a := (\lambda^2 - d)$ for its diagonal components; all the off-diagonal components are $b := (-d)$. We can calculate this determinant with our formula from lemma 2.23 (p. 15), with $a$ and $b$ as specified just now. This yields:

$$0 = \det M_\lambda = ((\lambda^2 - d) - (-d))^{d-1} \cdot [(\lambda^2 - d) + (d-1)(-d)] = (\lambda^2)^{d-1} \cdot [\lambda^2 - d^2]$$

Thus, the spectrum of $K_{d,d}$ consists of $(2d-2)$ zeros and the values $\pm d$.

(It is no coincidence that the spectrum is symmetric. Stanic [Sta17] (pp. 13f.) proves that any bipartite graph will have a symmetric spectrum of its adjacency matrix.

In the special case of bipartite regular graphs, we may always use Silvester's determinant formula and end up with a matrix with $\lambda^2$ on its diagonal elements, and has the form of an eigenvalue problem for eigenvalues $\mu := \lambda^2$. Solving the characteristic polynomial for $\mu$ (all of whose roots will be non-negative) will always yield symmetric eigenvalues $\lambda = \pm\mu$.)

As for the isoperimetric constant, Mohar [Moh89] states that

$$\mathcal{I}(K_{d,d}) = \begin{cases} \frac{d^2}{2d}, & \text{for even } d \\ \frac{d^2+1}{2d}, & \text{for odd } d \end{cases},$$

which makes for expansion values $\delta$ of exactly $1/2$ or $((1/2) + 1/(2d^2))$, respectively.

We verified the isoperimetric constants for some small values of $d$ with our Java algorithm F:

- For even $d$, take a subset consisting of half the nodes of each color, so $|S| = d$. Each of those nodes will have $(d/2)$ external edges leading to $\bar{S}$, making for $e(S, \bar{S}) = d^2/2$.

- For odd $d$, take $(d-1)/2$ nodes of one color (say, color 1) and $(d+1)/2$ of the other (here, color 2). Then, the color-1 nodes in $S$ will have external edges to the $(d-1)/2$ color-2 nodes belonging to $\bar{S}$, and the color-2 nodes in $S$, edges to the $(d+1)/2$ color-1 nodes. Thus, the edge count between $S$ and $\bar{S}$ is:

$$e(S,\bar{S}) = \left(\frac{d-1}{2}\right)^2 + \left(\frac{d+1}{2}\right)^2 = \frac{1}{4}[d^2 - 2d + 1 + d^2 + 2d + 1] = \frac{1}{2}(d^2+1)$$

  The subset size is $d$, like in the even $d$ case.

Again, as for the complete graphs, the expansion looks promising at first, but the node degrees are manifestly linear in the overall node count, and like the complete graphs from above, the complete regular bipartite graphs are not sparse, as can be apprehended from figure 2.15 already.

This concludes our general introduction to (regular) graphs, adjacency matrix spectra and edge expanders; we now proceed with the construction proofs from [ASS08] by Alon et al.

# Chapter 3

# Construction of Constant-Degree Edge Expanders

The article [ASS08] by N. Alon, O. Schwartz and A. Shapira describes how a class of large edge expanders with fixed degree (i.e. independent of the node count) may be constructed, and provides proofs for the most important concepts involved. Those include observations on the spectral gap of expander graphs, a special kind of product between graphs (particularly, its application in products of expanders), and a class of easily constructible expanders with non-constant degree that will later occur as factors in such products.

We will explore all those topics in preparation for the main construction section 3.6 (pp. 60ff.), and conclude with a small section containing specializations of the main construction.

## 3.1 About the Spectral Gap

We follow a proof from [ASS08] (theorem 1) to show an important result combining the *spectral gap*, i.e. the difference between the largest and second-largest eigenvalues, with its edge expansion $\delta$ (which is directly related to the isoperimetric constant, see definitions 2.37 (p. 24) and 2.40 (p. 25)).

**Theorem 3.1** *For any undirected $d$-regular graph $G$ with expansion $\delta$ and largest eigenvalues (of its adjacency matrix $A(G)$) $\lambda_1 := d$ and $\lambda_2$:*

$$\delta \geq \frac{1}{2}\left(1 - \frac{\lambda_2}{d}\right) = \frac{1}{2}\frac{d-\lambda_2}{d} = \frac{1}{2}\frac{\lambda_1 - \lambda_2}{d}$$

First, we observe that this is in accordance with corollary 2.38 (p. 24) about disconnected graphs. If $G$ is disconnected, it will have $\mathcal{I}(G) = 0$, and also, because of corollary 2.35 (p. 23), the eigenvalue $d$ occurs at least twice – therefore $\lambda_2 = \lambda_1 = d$, and the expression on the right-hand side of the statement is zero.

Secondly, we may readily multiply the inequality with $d > 0$, to yield an equivalent statement:

$$\mathcal{I}(G) \geq \frac{1}{2}(d - \lambda_2)$$

Before we commence the proof, we consider the spectral gaps and expansion values for our examples in section 2.3 (pp. 25ff.).

- For the Petersen graph, $\mathcal{I}(G) = 1$, $d = 3$ and $\lambda_2 = 1$; thus, $(d - \lambda_2)/2 = 2/2 = 1 \leq 1$. ✓
- For the poor expansion example $G_3$, $\mathcal{I}(G) = 1/3$, $d = 3$ and $\lambda_2 = 1 + \sqrt{3}$. The right-hand expression here evaluates to $(d - \lambda_2)/2 = (2 - \sqrt{3})/2 \approx 0.134 \leq 1/3$ ✓
- For the complete graphs $K_n$, $\mathcal{I}(G) = \lceil n/2 \rceil$, $d = (n-1)$ and $\lambda_2 = (-1)$. Thus,

$$\frac{d - \lambda_2}{2} = \frac{n}{2} \leq \left\lceil \frac{n}{2} \right\rceil \quad \checkmark$$

- For the complete bipartite graphs $K_{d,d}$ with $d > 1$, $\mathcal{I}(G) \geq d/2$ (depending on even or odd $d$), and $\lambda_2 = 0$. Thus, $(d - \lambda_2)/2 = d/2 \leq \mathcal{I}(G)$ ✓

### 3.1.1 Partition Vectors

We recall that, since $G = (V, e)$ is $d$-regular, the largest eigenvalue $\lambda_1$ is $d$, and $\vec{v}_1 := (1, \cdots, 1)^T$ is an associated eigenvector, as per lemma 2.31 (p. 22) and lemma 2.34 (p. 23).

In preparation of the proof, we examine the behavior of certain vectors for a given partition of the nodes into $S$ and $\bar{S} = V \setminus S$. This will allow us to reason about the inherent inequality stated in lemma 2.36 (p. 23).

**Definition 3.2** *For a given partition $S \uplus \bar{S} = V$ of a graph $G$ with node set $V$, define the characteristic vectors $\vec{x}_S$ and $\vec{x}_{\bar{S}}$ via*

$$(\vec{x}_S)_j := \begin{cases} 1, & j \in S \\ 0, & j \in \bar{S} \end{cases} \qquad (\vec{x}_{\bar{S}})_j := \begin{cases} 1, & j \in \bar{S} \\ 0, & j \in S \end{cases}$$

Since any node $j$ belongs to either $S$ or $\bar{S}$, $\vec{x}_S + \vec{x}_{\bar{S}} = \vec{v}_1 = (1, \cdots, 1)^T$. $\vec{x}_S$ contains exactly $|S|$ non-zero components; $\vec{x}_{\bar{S}}$, contains $|\bar{S}|$ such components.

We examine their scalar products among each other, and with the eigenvector $\vec{v}_1$:

$$\begin{aligned} \langle \vec{x}_S, \vec{x}_S \rangle &= \sum_j (\vec{x}_S)_j (\vec{x}_S)_j = |S| \\ \langle \vec{x}_{\bar{S}}, \vec{x}_{\bar{S}} \rangle &= \sum_j (\vec{x}_{\bar{S}})_j (\vec{x}_{\bar{S}})_j = |\bar{S}| \\ \langle \vec{x}_S, \vec{x}_{\bar{S}} \rangle &= \sum_j (\vec{x}_S)_j (\vec{x}_{\bar{S}})_j = 0 \\ \langle \vec{x}_S, \vec{v}_1 \rangle &= \sum_j (\vec{x}_S)_j (\vec{v}_1)_j = |S| \\ \langle \vec{x}_{\bar{S}}, \vec{v}_1 \rangle &= \sum_j (\vec{x}_{\bar{S}})_j (\vec{v}_1)_j = |\bar{S}| \end{aligned}$$

### 3.1.2 Additional Edge Counters

If we construct quadratic forms of the adjacency matrix $A$ with the characteristic vectors of the previous subsection, we obtain certain edge counts. For that, we define

**Definition 3.3** *For a given partition $S \uplus \bar{S} = V$ of an undirected graph $G = (V, e)$, define the following edge counters:*

- *Let $e(S)$ the sum of all edges among nodes in $S$.*
- *Let $e(\bar{S})$ the sum of all edges among nodes in $\bar{S}$.*
- *Let $e(S, \bar{S}) = e(\bar{S}, S)$ the sum of all edges between nodes in $S$ and nodes in $\bar{S}$ (like in definition 2.37, p. 24).*

**Corollary 3.4** *For a given partition $S \uplus \bar{S} = V$ of a $d$-regular graph $G = (V, e)$, $|V| = n$:*

$$e(S) + e(\bar{S}) + e(S, \bar{S}) = \frac{nd}{2} = \frac{d}{2}|V| = \frac{d}{2}\left(|S| + |\bar{S}|\right)$$

Proof: The left-hand expression sums up all the edges of $G$. Because $G$ is $d$-regular, this edge count is $nd/2$, according to lemma 2.25 (p. 20). ■

Now for the advertised quadratic forms:

$$\begin{aligned} \vec{x}_S^T A \vec{x}_S &= \sum_{j,k} (\vec{x}_S)_j A_{jk} (\vec{x}_S)_k = \sum_{\substack{j \in S \\ k \in S}} A_{jk} = 2e(S) \\ \vec{x}_{\bar{S}}^T A \vec{x}_{\bar{S}} &= \sum_{j,k} (\vec{x}_{\bar{S}})_j A_{jk} (\vec{x}_{\bar{S}})_k = \sum_{\substack{j \in \bar{S} \\ k \in \bar{S}}} A_{jk} = 2e(\bar{S}) \\ \vec{x}_S^T A \vec{x}_{\bar{S}} &= \sum_{j,k} (\vec{x}_S)_j A_{jk} (\vec{x}_{\bar{S}})_k = \sum_{\substack{j \in S \\ k \in \bar{S}}} A_{jk} = e(S, \bar{S}) \\ \vec{x}_{\bar{S}}^T A \vec{x}_S &= \sum_{j,k} (\vec{x}_{\bar{S}})_j A_{jk} (\vec{x}_S)_k = \sum_{\substack{j \in \bar{S} \\ k \in S}} A_{jk} = e(\bar{S}, S) = e(S, \bar{S}) \end{aligned}$$

The factors 2 in the first two above equations result from the fact that $A$ is symmetric. If $j$ and $k$ are taken from the same index sets, every edge will be counted twice. In the third equation, $j$ is from $S$ and $k$ from $\bar{S}$, so the symmetric contributions are not counted here. The same argument applies to the fourth equation.

### 3.1.3 Combining the Characteristic Vectors

We now examine a special linear combination of the characteristic vectors for a given partition of $V$:

$$\vec{x} := |\bar{S}|\vec{x}_S - |S|\vec{x}_{\bar{S}}$$

Taking its scalar product with the characteristic vectors yields (using the bilinear property of the real scalar product (cf. definition A.8, p. 69, and subsequent remarks)):

$$\begin{aligned}\langle\vec{x},\vec{x}_S\rangle &= |\bar{S}|\langle\vec{x}_S,\vec{x}_S\rangle - |S|\langle\vec{x}_{\bar{S}},\vec{x}_S\rangle = |\bar{S}||S| - |S|\cdot 0 = |\bar{S}||S| \\ \langle\vec{x},\vec{x}_{\bar{S}}\rangle &= |\bar{S}|\langle\vec{x}_S,\vec{x}_{\bar{S}}\rangle - |S|\langle\vec{x}_{\bar{S}},\vec{x}_{\bar{S}}\rangle = |\bar{S}|\cdot 0 - |S||\bar{S}| = -|S||\bar{S}|\end{aligned}$$

Therefore:

$$\langle\vec{x},\vec{v}_1\rangle = \langle\vec{x},(\vec{x}_S+\vec{x}_{\bar{S}})\rangle = \langle\vec{x},\vec{x}_S\rangle + \langle\vec{x},\vec{x}_{\bar{S}}\rangle = 0 \qquad \Leftrightarrow \qquad \vec{x}\perp\vec{v}_1$$

We also take the scalar products of $\vec{x}$ with itself, and its quadratic form with the adjacency matrix $A$, plugging in the results of the previous two subsections and using $n = |V| = |S| + |\bar{S}|$:

$$\begin{aligned}\langle\vec{x},\vec{x}\rangle &= \left\langle(|\bar{S}|\vec{x}_S - |S|\vec{x}_{\bar{S}}),(|\bar{S}|\vec{x}_S - |S|\vec{x}_{\bar{S}})\right\rangle = |\bar{S}|^2|S| - 0 - 0 + |S|^2|\bar{S}| = \left(|\bar{S}| + |S|\right)|S||\bar{S}| \\ &= n|S||\bar{S}| \\ \vec{x}^T A\vec{x} &= \left(|\bar{S}|\vec{x}_S^T - |S|\vec{x}_{\bar{S}}^T\right) A\left(|\bar{S}|\vec{x}_S - |S|\vec{x}_{\bar{S}}\right) \\ &= |\bar{S}|^2\vec{x}_S^T A\vec{x}_S - |\bar{S}||S|\vec{x}_S^T A\vec{x}_{\bar{S}} - |S||\bar{S}|\vec{x}_{\bar{S}}^T A\vec{x}_S + |S|^2\vec{x}_{\bar{S}}^T A\vec{x}_{\bar{S}} \\ &= 2|\bar{S}|^2 e(S) - 2|S||\bar{S}|e(S,\bar{S}) + 2|S|^2 e(\bar{S})\end{aligned}$$

We note that these two expressions occur in lemma 2.36 (p. 23), and that $\vec{x}$ is indeed perpendicular to $\vec{v}_1 = (1,\cdots,1)^T$. In order for the main argument to work, we have to rewrite $\vec{x}^T A\vec{x}$ in a way that hints at the expressions occurring in the definition 2.37 (p. 24) of the isoperimetric constant:

### 3.1.4 Rewriting the Edge Counters

We recall corollary 3.4 (p. 33). In fact, we can split this formula along the partition as well and obtain:

**Lemma 3.5** *For a given partition $S \uplus \bar{S} = V$ of a $d$-regular graph $G = (V, e)$, $|V| = |S| + |\bar{S}| = n$:*

$$\begin{aligned}e(S) + \frac{1}{2}e(S,\bar{S}) &= \frac{d|S|}{2} \\ e(\bar{S}) + \frac{1}{2}e(S,\bar{S}) &= \frac{d|\bar{S}|}{2}\end{aligned}$$

Proof: We first observe that the sum of those two equations yields exactly the contents of corollary 3.4.

Secondly, both sides of the equations may turn out to be half-integers, depending on the selected subset $S$, which does not necessarily contain an even number of nodes. This already suggests that we may proceed by multiplying the stated equations by 2:

$$2e(S) + e(S,\bar{S}) = d|S| \qquad \wedge \qquad 2e(\bar{S}) + e(S,\bar{S}) = d|\bar{S}|$$

In order to show this, we consider a new graph $\tilde{G}$ consisting of two copies of $G$ side by side, and subsets $S_1$ and $S_2$ that are both equal to the subset $S$ from $V$ within their respective copies of $G$, as illustrated in figure 3.1 (p. 35).

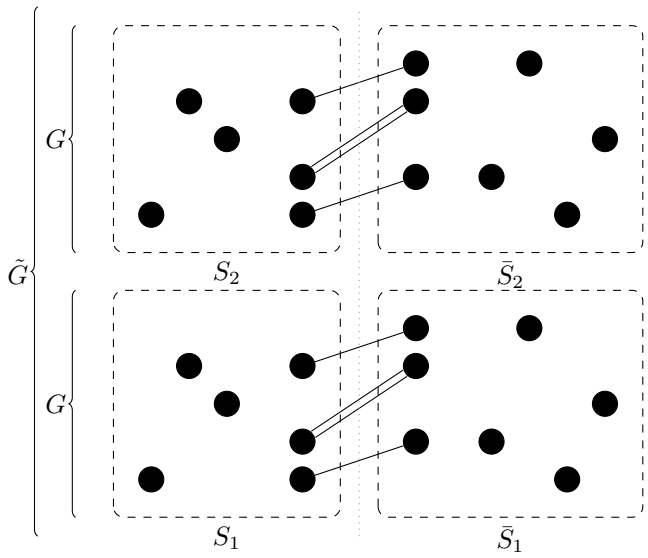


Figure 3.1: Doubled partitioned graph $\tilde{G}$ with partition-crossing edges

Both $S_1$ and $S_2$ each contain $e(S)$ internal edges, and each is connected via $e(S, \bar{S})$ edges to their respective complements $\bar{S}_1$ and $\bar{S}_2$.

In $S$, all purely internal nodes have $d$ edges to other nodes in $S$. The nodes with edges into $\bar{S}$ have fewer than $d$ edges connecting to nodes in $S$, by a total of $e(S, \bar{S})$.

Now, if we take the doubled graph $\tilde{G}$ and cut open all the partition-crossing edges, (in figure 3.1, that means cutting along the dotted vertical line), we end up with twice $e(S, \bar{S})$ half-edges. We connect every such half-edge in $S_1$ to its counterpart in $S_2$, and do the same for the half-edges in $\bar{S}_1$ and $\bar{S}_2$, as shown in figure 3.2:

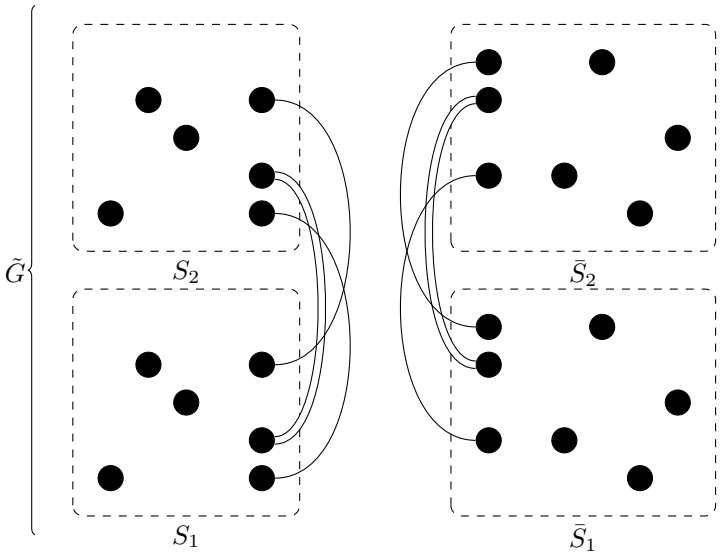


Figure 3.2: Doubled graph $\tilde{G}$, reconnected

We observe that the reconnected $\tilde{G}$ now consists of two unconnected sub-graphs, one made up of $S_1$ and $S_2$ (with $2|S|$ nodes), the other one of $\bar{S}_1$ and $\bar{S}_2$ (with $2|\bar{S}|$ nodes). Each of those sub-graphs is $d$-regular, with the full compliment of internal edges.

Because the sub-graphs are $d$-regular, their edge counts are $(2|S|)d/2 = d|S|$ and, respectively, $(2|\bar{S}|)d/2 = d|\bar{S}|$. But because of the doubling procedure from above, those numbers also equal $2e(S) + e(S, \bar{S})$ and $2e(\bar{S}) + e(S, \bar{S})$, respectively, by counting the internal edges of the respective subsets and adding the reconnected edges. This proves the stated equations. ■

### 3.1.5 Proof of Theorem 3.1

With these preparations, we are now able to prove the main assertion. We begin by restating lemma 2.36 from p. 23:

$$\lambda_2 = \max_{\vec{0}\neq\vec{x}\perp(1,\cdots,1)^T} \left\{ \frac{\vec{x}^T A\vec{x}}{\langle\vec{x},\vec{x}\rangle} \right\}$$

Now, for a non-empty subset $S$ of nodes, $\vec{x}$ is not the zero vector, and it is perpendicular to $\vec{v}_1$ (see above). We multiply the above equation and turn it into an inequality by omitting the maximum operation:

$$\lambda_2\langle\vec{x},\vec{x}\rangle \geq \vec{x}^T A\vec{x}$$

We rewrite the right-hand expression by plugging in our formulas from lemma 3.5 (p. 34) into what we had already determined:

$$\begin{aligned}
\vec{x}^T A\vec{x} &= 2|\bar{S}|^2 e(S) - 2|S||\bar{S}|e(S,\bar{S}) + 2|S|^2 e(\bar{S}) \\
&= |\bar{S}|^2\left(d|S| - e(S,\bar{S})\right) - 2|S||\bar{S}|e(S,\bar{S}) + |S|^2\left(d|\bar{S}| - e(S,\bar{S})\right) \\
&= d\left(|\bar{S}|^2|S| + |S|^2|\bar{S}|\right) - e(S,\bar{S})\left(|\bar{S}|^2 + 2|S||\bar{S}| + |S|^2\right) \\
&= d|S||\bar{S}|\left(|\bar{S}| + |S|\right) - e(S,\bar{S})\left(|\bar{S}| + |S|\right)^2 \\
&= n\left[d|S||\bar{S}| - n\cdot e(S,\bar{S})\right]
\end{aligned}$$

Now we collect the left-hand expression:

$$\lambda_2\langle\vec{x},\vec{x}\rangle = \lambda_2 \cdot n|S||\bar{S}|$$

We divide by $n$ and solve for $e(S,\bar{S})/|S|$:

$$\begin{aligned}
& \lambda_2 \cdot n|S||\bar{S}| \geq n\left[d|S||\bar{S}| - n\cdot e(S,\bar{S})\right] \\
\Leftrightarrow\quad & \lambda_2|S||\bar{S}| \geq d|S||\bar{S}| - n\cdot e(S,\bar{S}) \\
\Leftrightarrow\quad & n\cdot e(S,\bar{S}) \geq |S||\bar{S}|(d-\lambda_2) \\
\Leftrightarrow\quad & \frac{e(S,\bar{S})}{|S|} \geq (d-\lambda_2)\frac{|\bar{S}|}{n}
\end{aligned}$$

In the definition (2.37, p. 24) of the isoperimetric constant, only subsets $S$ with size up to $n/2$ are considered. Restricting $S$ in this way fixes $|\bar{S}| \geq (n/2)$, and therefore

$$\frac{e(S,\bar{S})}{|S|} \geq \frac{1}{2}(d-\lambda_2)$$

Now, this holds for *all* the partitions of $V$ into $S$ and $\bar{S}$ obeying the size restrictions, including any partition for which the left-hand expression becomes minimal – but this is just the isoperimetric constant $\mathcal{I}(G)$, which proves the statement of theorem 3.1. ■

## 3.2 Replacement Product

There are various possibilities to combine two graphs with $n_1$ and $n_2$ nodes (respectively) into a graph with $n_1 \cdot n_2$ nodes, e.g. the zig-zag product [Sta17] (pp. 189ff.) or the tensor product [BM11] (pp. 65f.).

We follow [ASS08] in a definition of a replacement product $G \circ H$, where every node of $G$ is replaced (hence the name) by a copy of $H$, with additional edges according to the previous edges in $G$. This product is not commutative, because it requires that the node count of $H$ equal the degree of $G$. It will feature prominently in the construction of constant-degree expanders.

### 3.2.1 Description

**Definition 3.6** *For a $D$-regular and $D$-edge-colorable graph $G$ with $n$ nodes, and a $d$-regular graph $H$ with $D$ nodes, the* replacement product *$G \circ H$ obtained by the following procedure:*

1. *For any node $j$ of $G$, $1 \leq j \leq n$, let $H_j$ be a copy of $H$ with nodes $1 \leq k_j \leq D$, and with all the edges of $H$ reproduced in each copy.*
2. *For all edges of $G$ with color $c$, $1 \leq c \leq D$, between nodes $r$ and $s$, add a $d$-fold multi-edge between the nodes $k_r := c$ and $k_s := c$ in the copies $H_r$ and $H_s$.*
3. *$G \circ H$ consists of all the nodes in the $n$ copies of $H$, with the connectivity as described in the previous two steps.*

**Corollary 3.7** *For $G, H$ as in definition 3.6, the replacement product $G \circ H$ is a $(2d)$-regular graph with $n \cdot D$ nodes.*

Proof: Since $G$ had $n$ nodes, and $H$, $D$ nodes, respectively, the first construction step yields a graph with $n \cdot D$ nodes. It is at this point $d$-regular, because the various copies $H_j$ are not connected with each other.

Since $G$ is $D$-edge-colorable, let us assume $G$ has been edge-colored in some way. Because $G$ is also $D$-regular, each of $G$'s nodes is incident with exactly one edge of each color $c$, $1 \leq c \leq D$. We recall corollary 2.17 (p. 14) to ascertain that $G$ has no loops. In that case, all the edges of any node in $G$ must connect it to some other node of $G$ (multi-edges are not forbidden, though).

The graph $H$ consists of exactly $D$ nodes. If we number them $1, \cdots, D$ in the usual way, we actually obtain a (trivial) node-coloring of $H$, provided that $H$ is loop-free (cf. corollary 2.20, p. 14) – although this has no bearing on this proof. We may, however, observe that each node of $H$ corresponds to exactly one of the $D$ edge colors of $G$.

Together with the above observation on $G$, this means that the second construction step will add one $d$-fold multi-edge to every node of every copy of $H$, and that every such multi-edge is between two nodes with equal internal number (corresponding to the color $c = k_r = k_s$) in two different copies $H_r \neq H_s$ of $H$.

Since every node in $G \circ H$ belongs to exactly one copy of $H$, every such node is incident with the $d$ edges inside its copy of $H$, and with the $d$ edges of the added $d$-fold multi-edge from step 2, which makes $G \circ H$ $(2d)$-regular.

If $G$ had a multi-edge between its nodes $r$ and $s$, this implies multiple edges with different colors (because $G$ is $D$-edge-colorable); in that case, there are several multi-edges added between the copies $H_r$ and $H_s$, but each between different node pairs of those copies. ■

**Corollary 3.8** *For $G, H$ as in definition 3.6, if $H$ is $C$-edge-colorable, then $G \circ H$ is $(C+d)$-edge-colorable. Also, $C \geq d$.*

Proof: Take the colors $1, \cdots, C$ to color the edges of all the copies in a uniform way during the first construction step. All the $d$-fold multi-edges between nodes of different copies $H_j$ that are added in the second construction step may be colored $C+1, \cdots C+d$.

Because $H$ is $d$-regular, any proper edge coloring of $H$ must use at least $d$ different colors, which means $C \geq d$ (cf. corollary 2.18, p. 14). ■

We now show some examples to illustrate the construction of replacement products.

### 3.2.2 Examples

**A Minimal Example: $C_4 \circ K_2$**

$C_4$ is the cycle-graph with $n = 4$ nodes; it is $(D = 2)$-regular. Since $n$ is an even number, $C_4$ is 2-edge-colorable with alternating colors. Thus, $C_4$ qualifies as the left-hand graph $G$ of a replacement product.

The complete graph with two nodes $K_2$ consists of only $D = 2$ nodes connected by a single edge; it is $d = 1$-regular, and qualifies as right-hand graph $H$ to the replacement product with $C_4$.

We show the two original graphs in figure 3.3 (p. 38) and recall that the node numbers of $H$ correspond to the color numbers of $G$, whereas the node numbers of $G$ represent the various copies of $H$ obtained in the product's first construction step.

Because $K_2$ is a loop-free graph, it can be (properly) node-colored with two colors (otherwise, the node colors shown in figure 3.3 would only show the intended correspondence to the edge colors of $C_4$). We use the colors blue and orange for $c = 1$ and $c = 2$, respectively.

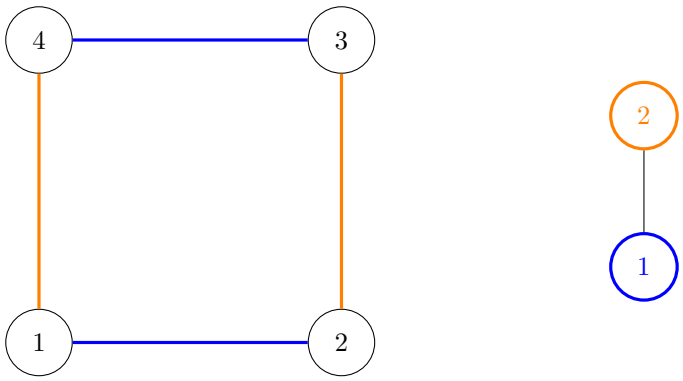


Figure 3.3: Graphs $G := C_4$ (left) and $H := K_2$ (right)

We now perform the first construction step (see figure 3.4): Each of the four nodes of $G$ is replaced by a copy of $H$.

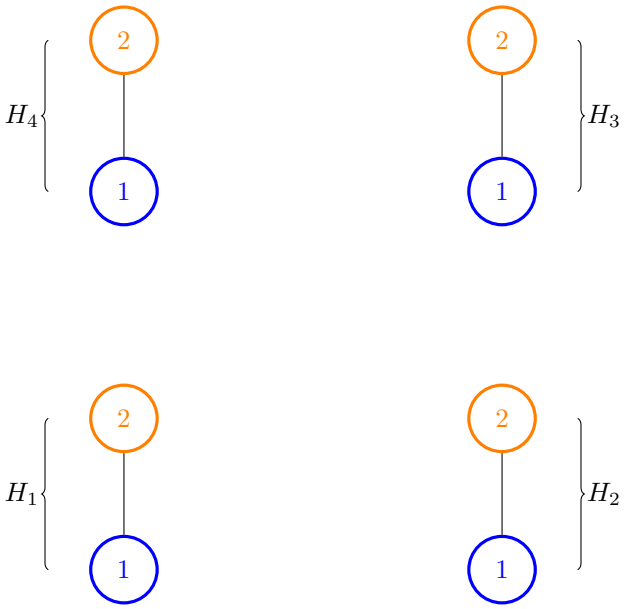


Figure 3.4: $G \circ H$, step 1

Because $d = 1$, the second construction step involves only single edges added to each of the nodes. Node 1 of $H_1$ will be connected to node 1 of $H_2$ because there is a blue ($c = 1$) edge in $G$ between its nodes 1 and 2. Node 2 of $H_1$ will be connected to node 2 of $H_4$ because there is an orange ($c = 2$) edge in $G$ between its nodes 1 and 4. We show the completed second step in figure 3.5 (p. 39), the added multi-edges (here: 1-fold) drawn thicker.

We observe that the resulting graph $G \circ H$ is indeed 2-regular ($2 = 2d$). Because $H$ is $d$-edge-colorable, the product graph would be ($2d = 2$)-edge-colorable (in figure 3.5, we can interpret the thin edges as one color, and the thick edges as another color).

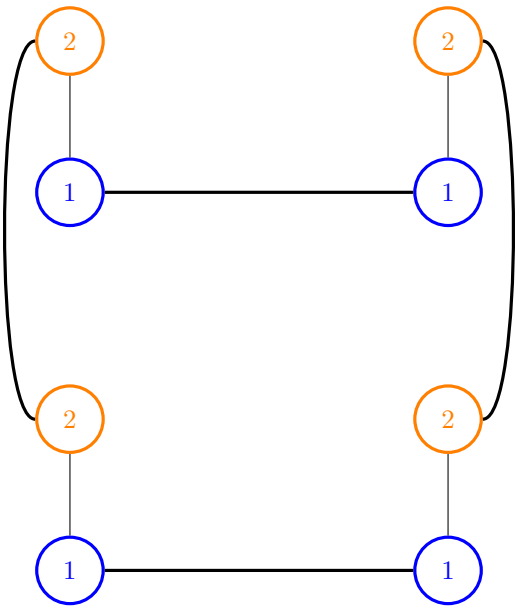


Figure 3.5: $G \circ H$, step 2

**A Slightly Bigger Example:** $K_{4,4} \circ C_4$

The complete bipartite graph $G := K_{4,4}$ is $(D = 4)$-regular and $(D = 4)$-edge-colorable. The cycle graph $H := C_4$ is $(d = 2)$-regular and contains $D = 4$ nodes (see the previous example) (it is also 2-edge-colorable and simple, i.e. loop-free). For the additional colors, we use teal $(c = 3)$ and purple $(c = 4)$. In anticipation of the second construction step, we draw the cycle graph in a way that its internal edges will not overlap with the multi-edges (which will be twofold in this case).

Because the resulting product graph will contain eight copies of $H$, we omit the node numbers in $H$ and instead draw colored dots; see figure 3.6:

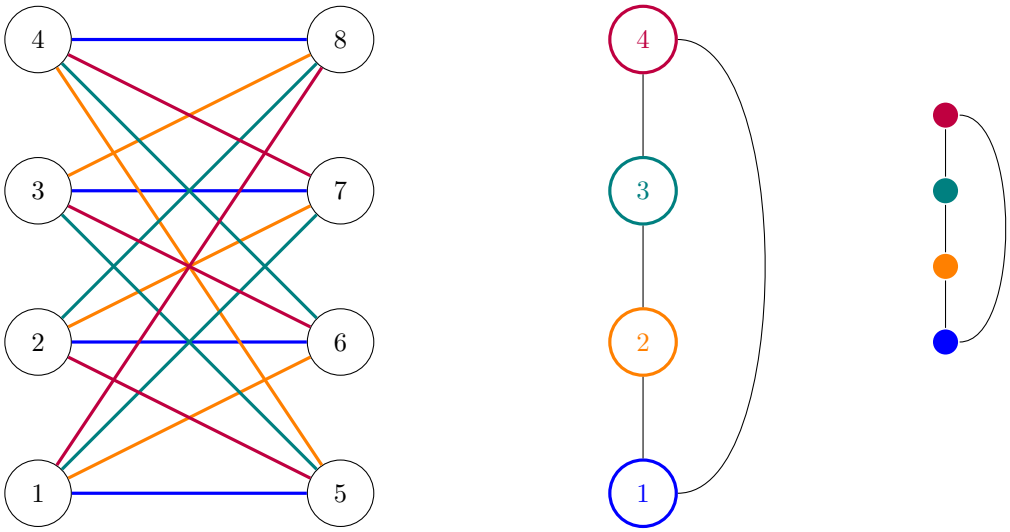


Figure 3.6: Graphs $G := K_{4,4}$ (left) and $H := C_4$ (middle); shorthand for $H$ (right)

We show the product $G \circ H$ in figure 3.7 (p. 40). This is a 4-regular graph with 32 nodes, and it could be 4-edge-colored if so desired.

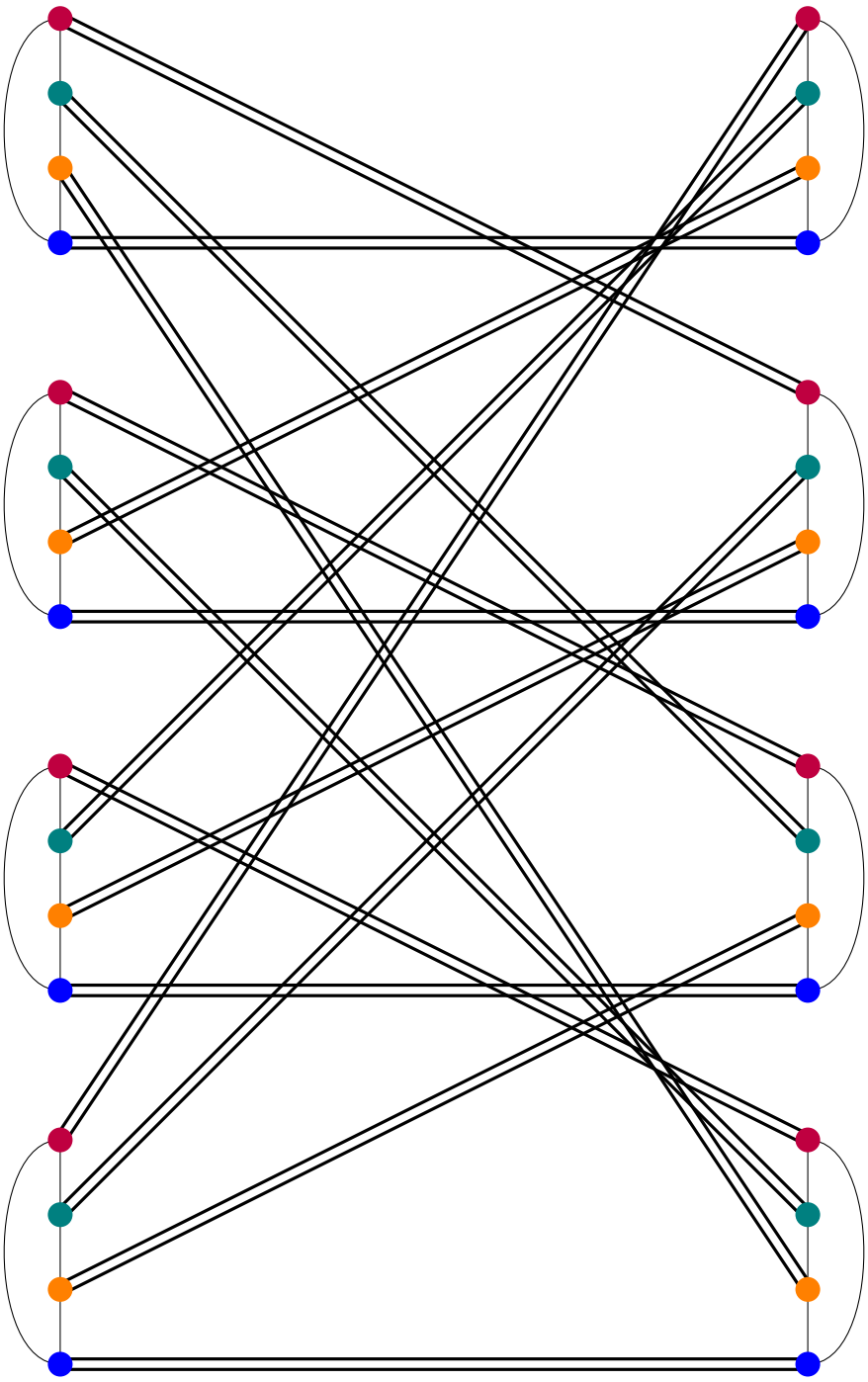

Figure 3.7: $G \circ H$

### An Example with Multi-Edges in $G$

Our last example features existing multi-edges in the left-hand graph $G$ – we take a $(D = 3)$-regular bipartite graph with $(D = 3)$-edge-coloring and $n = (5 + 5)$ nodes. For $H$, we choose the cycle graph $C_3$, which has $D = 3$ nodes and is $(d = 2)$-regular (but not 2-edge-colorable, because it has odd length). See figure 3.8 (p. 41).

Note that $G$ contains a triple edge between nodes 2 and 8 – a separate connected component. There is also a double edge between nodes 3 and 10.

The product $G \circ H$ is shown in figure 3.9, p. 41. We observe that the product graph also consists of two connected components, one of which is made up of the copies $H_2$ and $H_8$, with three parallel double edges added in construction step 2.

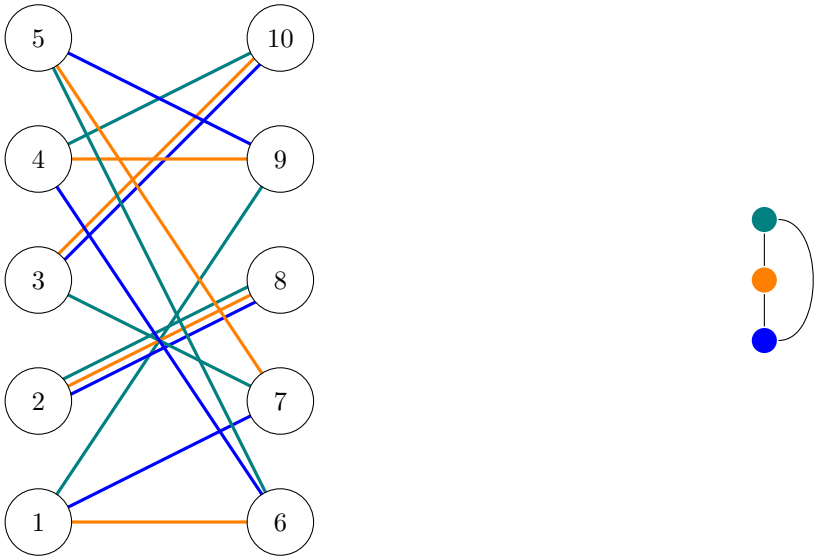


Figure 3.8: A 3-regular bipartite graph $G$ (left) and $H := C_3$ (right, shorthand)

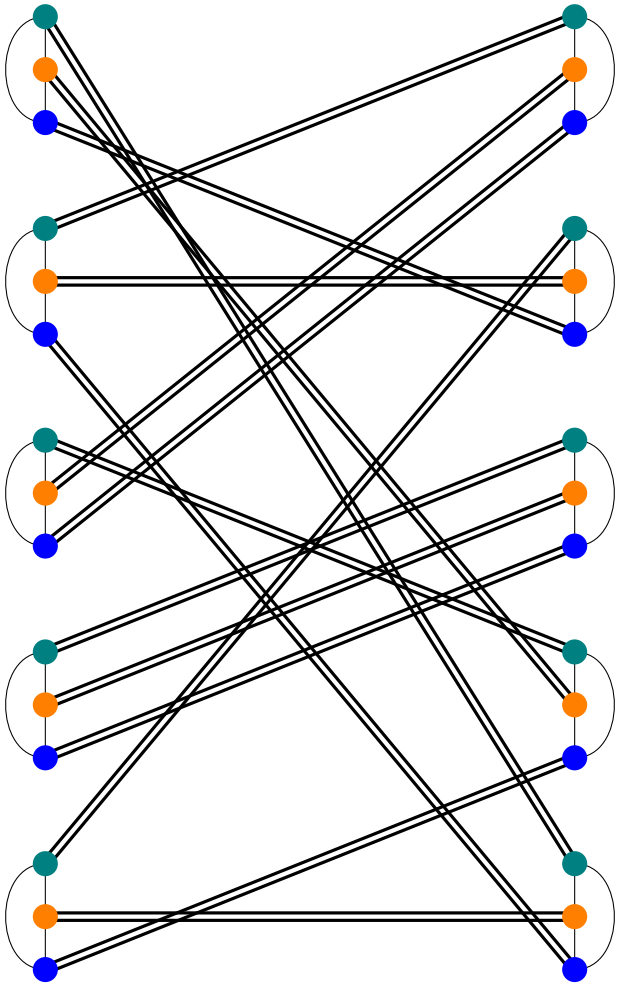

Figure 3.9: $G \circ H$

## 3.3 Replacement Product of Two Expanders

We recall from corollary 2.38 (p. 24) that all expanders with $\delta = 0$ consist of more than one connected component, which in turn means that all connected undirected graphs have $\delta > 0$.

We have seen in an example (figure 3.9, p. 41) where, for an disconnected $G$, the replacement product $G \circ H$ is also disconnected.

Before we further examine the expansion parameter of replacement products, we formalize the connectedness of a replacement product in the following

**Lemma 3.9** *Given two graphs $G, H$ satisfying the conditions of definition 3.6 (p. 37), the replacement product $G \circ H$ is connected if and only if both $G$ and $H$ are connected.*

Proof:

- If $G$ or $H$ are disconnected, the product $G \circ H$ is disconnected as well:
  - If $G = (V_G, e_G)$ is disconnected, the nodes of $G$ may be separated into two disjoint nonempty sets $V_{G,1} \uplus V_{G,2} = V_G$ with no edges between those sets. There might be more than two connected components, but it suffices to separate just one from the rest of the graph.

    We pick nodes $r \in V_{G,1}$ and $s \in V_{G,2}$. In $G$, $r$ and $s$ are not indirectly connected according to definition 2.13 (p. 13). But this also means that the second step of definition 3.6 stipulates that no multi-edge will be added between the copies $H_r$, $H_s$.

    This holds for any $r$ and $s$ we choose – which means that we may group the corresponding copies $H_r$ and $H_s$ into two sets $H_{(1)}$ and $H_{(2)}$, and that there are no multi-edges added in the second step of definition 3.6 between any two copies of $H$, where one is in $H_{(1)}$ and the other in $H_{(2)}$. Because the first step of definition 3.6 only creates edges within the confines of the various copies of $H$, there will be no edges between $H_{(1)}$ and $H_{(2)}$ in $G \circ H$; thus, the product is disconnected.
  - If $H = (V_H, e_H)$ is disconnected, we may separate its nodes into two disjoint nonempty sets $V_{H,1} \uplus V_{H,2} = V_H$ with no edges between those sets. This also holds for all the copies $H_j$ of $H$ that are required to construct $G \circ H$. The numbering of nodes is the same in all copies.

    But step 2 of definition 3.6 will only add multi-edges between nodes with the same internal number in two different copies $H_r$, $H_s$, according to the respective edge color in $G$. Thus, any such two nodes in copies of $H$ will correspond to a single node in the original $H$, which belongs to either $V_{H,1}$ or $V_{H,2}$.

    If we group all the nodes of $G \circ H$ corresponding to $V_{H,1}$ and $V_{H,2}$, into two sets $V_{(1)}$ and $V_{(2)}$, respectively, then there are no edges between those node sets in $G \circ H$. Step 1 of definition 3.6 replicates the connectivities in $H$, and step 2 will add multi-edges only within $V_{(1)}$ and $V_{(2)}$, not across – this makes $G \circ H$ disconnected.
- If $G$ and $H$ are both connected, we observe that...
  - ... all the nodes in $H$ are indirectly connected, i.e. there are paths of edges between them. This is faithfully reproduced within each of the copies $H_j$ in $G \circ H$.
  - ... all the nodes in $G$ are indirectly connected, too. Any path between two arbitrary nodes $r, s$ in $G$ can be reproduced as a path between the copies $H_r$ and $H_s$ via the multi-edges added in step 2 of definition 3.6:

    If we can reach $s$ from $r$ in $G$ via a single node $j$ by traveling along two edges, then we may pick any node in $H_r$. Because $H_r$ is connected, we can reach the node with a multi-edge towards $H_j$. Because $H_j$ is connected, we can reach the node in $H_j$ that has a multi-edge towards $H_s$. Having traveled there, we can reach any node in $H_s$.

    The same argument holds if there are more than one intermediate nodes on the path from $r$ to $s$. Since $r$ and $s$ were arbitrary, the whole of $G \circ H$ is therefore connected.

■

Thus, $G \circ H$ can only be connected if both $G$ and $H$ are. Alon et al. give the following lower bound for the expansion of a replacement product of two expanders:

**Theorem 3.10** *Let $G$ be an $[n, D, \delta_G]$-expander and $H$ a $[D, d, \delta_H]$-expander. Then, $F := G \circ H$ is an $[nD, 2d, \delta]$-expander, with $\delta \geq (\delta_G^2 \cdot \delta_H)/80$.*

Proof: Corollary 3.7 (p. 37) confirms that the resulting product $F$ has $nD$ nodes and is $(2d)$-regular. What remains is the examination of $\delta$, which we provide according to the proof in [ASS08] (theorem 3).

First, we observe that the stated inequality allows for $\delta = 0$ if any of $\delta_G, \delta_H$ are zero; this is in agreement with the preceding lemma 3.9. For the following, let $G$ and $H$ both be connected; thus $\delta > 0$.

Secondly, we recall that (per definition 3.6, p. 37), in a replacement product $F = G \circ H$, the connectivity of $H$ is reproduced in each $H_j$, while the edges of $G$ are used to define connections between the various copies $H_j$.

The overall goal is to show that for any selection $S$ of at most half the nodes of $F = (V, e)$, the edge count between $S$ and $\bar{S} := V \setminus S$ satisfies the condition

$$e(S, \bar{S}) \geq \delta \cdot (2d) \cdot |S| \tag{*1}$$

with $\delta$ as stated above (using definitions 2.37 and 2.40 (pp. 24 and 25, respectively)). Because there will be several references to various inequalities, we use labeled equations in this section so as to enhance clarity.

Both $G$ and $H$ are connected graphs with positive expansion parameters. The proof by Alon et al. employs a number of ways to organize the nodes of $F$ into sets in order to reason about the expansion properties of $G$ or $H$, depending on context. We provide a schematic illustration of those sets in figure 3.10 (p. 44) and introduce the quantities in full before we commence the actual proof. In order to limit symbolic cluttering, we will use the labels of graphs and their node sets synonymously.

Let $S$ be a selection of nodes in $F$, and $\bar{S} := F \setminus S$ its complement, with $|S| \leq |F|/2 = nD/2$:

- For each node $j$ in $G$, and corresponding copy $H_j$ of $H$, let $S_j := S \cap H_j$; $\bar{S}_j := H_j \setminus S_j$. The $H_\bullet$ are visualized as rectangles in figure 3.10.

- The index set $I'$ contains the indices $j$ of those $H_j$ for which the node count $|S_j|$ is at most $(1 - \delta_G/4)D$. We recall from definition 3.6 (p. 37) that every one of the $n$ copies $H_\bullet$ consists of $D$ nodes.

  The index set $I'' := \{1, \cdots, n\} \setminus I'$ contains all the other indices, belonging to copies $H_k$ where $|S_k| > (1 - \delta_G/4)D$.

- Given the index sets (which depend on the concrete selection $S$), we define six further sets of nodes (see an example partition schematic in figure 3.10):

$$S' := \bigcup_{j \in I'} S_j \qquad \bar{S}' := \bigcup_{j \in I'} \bar{S}_j$$

$$S'' := \bigcup_{j \in I''} S_j \qquad \bar{S}'' := \bigcup_{j \in I''} \bar{S}_j$$

$$H' := \bigcup_{j \in I'} H_j \qquad H'' := \bigcup_{j \in I''} H_j$$

- Because the $H_\bullet$, and, consequently, the $S_\bullet$ and $\bar{S}_\bullet$ are mutually disjoint, this implies that $F = H' \uplus H''$; $H' = S' \uplus \bar{S}'$; $H'' = S'' \uplus \bar{S}''$; $S = S' \uplus S''$ and $\bar{S} = \bar{S}' \uplus \bar{S}''$.

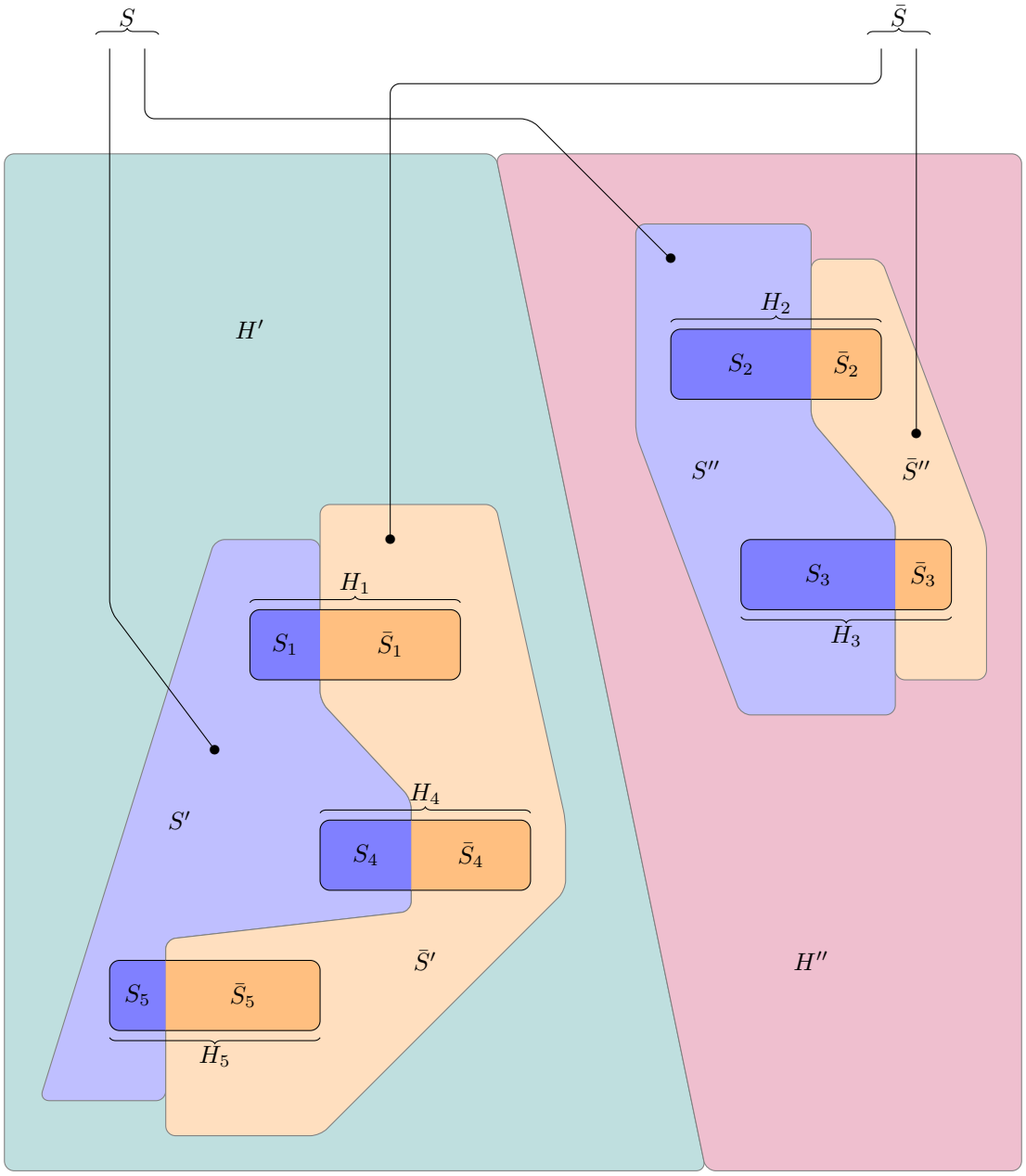


Figure 3.10: Schematic drawing of a product graph $F$, featuring the subsets used in the proof of theorem 3.10, with example index sets $I' = \{1, 4, 5\}$ and $I'' = \{2, 3\}$ for a given selection $S$. Nodes (within the $H_\bullet$) and edges (within the $H_\bullet$ and between different $H_\bullet$) omitted.

Broadly speaking, the index set $I''$ describes copies of $H$ where the selected nodes take up a large proportion of the nodes (remember that $\delta_G$ is between 0 and 1, so the $S_\bullet$ with those indices take up at least three-quarters of the nodes in the respective $H_\bullet$).

If the selection $S$ is centered on few, but highly populated, copies of $H$, the edges between $S$ and $\bar{S}$ are mainly between the $H\bullet$, and the expansion behavior is most influenced by $G$ and $\delta_G$. The other border case would be for a selection distributed over many, but sparsely populated, copies of $H$ – then, many of the edges between $S$ and $\bar{S}$ would be within the respective $H_\bullet$, and the expansion behavior would be most influenced by $H$ and $\delta_H$.

Beginning with the proof, we first consider the index set $I'$ and its associated node sets. For any $j \in I'$, the selected node set $S_j$ has size of up to $(1 - \delta_G/4)D$, which in turn means that the corresponding complement $\bar{S}_j$ has size $(D - |S_j|)$, which is at least $\delta_G D/4$. Because $|S_j| \leq D$, we may continue this inequality:

$$|\bar{S}_j| \geq \frac{\delta_G}{4} D \geq \frac{\delta_G}{4} |S_j|$$

We now have to distinguish two cases:

- If $|\bar{S}_j| \leq |S_j|$, then the node set $\bar{S}_j$ contains at most half the nodes of the copy $H_j$ of $H$.

Because $H$ is a $[D, d, \delta_H]$-expander, this also holds for the copy $H_j$, disregarding edges added in step 2 of the product construction (cf. definition 3.6, p. 37). Thus, we may conclude that the edge count between $\bar{S}_j$ and its complement in $H_j$, $\overline{\bar{S}_j} = S_j$, satisfies

$$e(\bar{S}_j, S_j) \geq \delta_H \cdot d \cdot |\bar{S}_j| \geq \delta_H \cdot d \cdot \frac{\delta_G}{4}|S_j|$$

- If $|S_j| < |\bar{S}_j|$, then the node set $S_j$ itself contains at most half the nodes of $H_j$, and we may plug in the expansion property of $H$ directly and obtain $e(S_j, \bar{S}_j) \geq \delta_H \cdot d \cdot |S_j|$. We may continue this inequality by introducing the factor $\delta_G/4$, which is at most $1/4$ but not negative. After that, the inequality agrees with the first case.

We collect both cases into a single result:

$$\forall j \in I' : e(S_j, \bar{S}_j) \geq \frac{1}{4}\delta_G\delta_H \cdot d \cdot |S_j| \tag{*2}$$

We sum over all $I'$ and introduce a factor of $2/2$ to reflect that $F$ is $(2d)$-regular:

$$e(S', \bar{S}') = \sum_{j \in I'} e(S_j, \bar{S}_j) \geq \sum_{j \in I'} \frac{1}{4}\delta_G\delta_H \cdot d \cdot |S_j| = \frac{1}{4}\delta_G\delta_H \cdot d \cdot |S'| = \frac{1}{8}\delta_G\delta_H \cdot 2d \cdot |S'| \tag{*3}$$

Now, if the size of $S'$, namely the node count of $S$ in copies of $H$ that are not "densely populated" (in the sense of the above definition), were at least $\delta_G \cdot |S|/10$, we could argue, using (*3), that

$$\boxed{e(S, \bar{S}) \geq e(S', \bar{S}) \geq e(S', \bar{S}') \geq \frac{1}{8}\delta_G\delta_H \cdot 2d \cdot \frac{1}{10}\delta_G|S| = \frac{1}{80}\delta_G^2\delta_H \cdot 2d \cdot |S|}, \tag{*4}$$

which would prove the statement $\delta \geq \delta_G^2\delta_H/80$, formulated in (*1).

The other case is rather more involved. For the following, therefore, $|S'| < \delta_G \cdot |S|/10$, and thus

$$|S''| = |S| - |S'| > \left(1 - \frac{\delta_G}{10}\right)|S| \geq \frac{9}{10}|S| \tag{*5}$$

(The last inequality in (*5) reflects the fact that $\delta_G \in [0, 1]$.)

Here, most of the selected nodes are in "densely populated" copies of $H$, and we will exploit the expansion property of $G$ (because many of the edges connecting $S$ and $\bar{S}$ will be between different $H_\bullet$), and thus consider the edges of $G$, or, respectively, the multi-edges between copies $H_\bullet$ in $F$. The additional effort stems from the fact that the nodes of $G$ have been replaced by whole sets of nodes (i.e. the copies of $H$) in $F$, and we will have to build our arguments from the node scale to the $H_\bullet$ scale.

Starting with observations on the index sets per se, which constitute a partition of the nodes in $G$, we will expand on the relationships between $I''$ and various set sizes, before we can finally combine all intermediary results.

First, we observe the following restrictions for the $S_j$ of the index set $I''$:

$$\forall j \in I'' : D \geq |S_j| > \left(1 - \frac{\delta_G}{4}\right) D \tag{*6}$$

We sum over all $I''$:

$$|I''| \cdot D \geq |S''| > |I''| \cdot \left(1 - \frac{\delta_G}{4}\right) D \tag{*7}$$

And we reorganize (*7) into two inequalities relating to $|I''|$:

$$|I''| \geq \frac{|S''|}{D} \qquad\qquad |I''| < \frac{|S''|}{\left(1 - \frac{\delta_G}{4}\right) D} \tag{*8}$$

Using $|S''| \leq |S| \leq nD/2$, we obtain, using the right-hand inequality of (*8), and $\delta_G \in [0, 1]$:

$$|I''| < \frac{nD/2}{\left(1 - \frac{\delta_G}{4}\right) D} = \frac{2n}{4 - \delta_G} \leq \frac{2n}{3} \tag{*9}$$

Thus, at most two-thirds of the copies of $H$ can be "densely populated" by $S$ in $F$.

From this, and $|I'| + |I''| = n$, we may directly conclude that

$$|I'| = n - |I''| \geq \frac{n}{3} \geq \frac{1}{2}|I''| \tag{*10}$$

We use the expansion property of $G$, which is an $[n, D, \delta_G]$-expander, and distinguish two possible cases for $|I'|$:

- If $|I'| \geq \frac{n}{2}$, then $|I''| \leq \frac{n}{2}$, and thus, there are at least $\delta_G \cdot D \cdot |I''|$ edges between $I''$ and $I' = \overline{I''}$ in $G$ (recalling that the sets $I'$ and $I''$ partition the node set of $G$).
- If $|I'| \leq \frac{n}{2}$, then there are at least $\delta_G \cdot D \cdot |I'|$ edges between the two node sets. We may express this in terms of $|I''|$ by using (*10).

Collecting both cases, we conclude that, for the graph $G$:

$$e_G(I', I'') \geq \frac{1}{2}\delta_G \cdot D \cdot |I''| \tag{*11}$$

Now, while $G$ describes connections between the single nodes in $I'$ and $I''$, the product graph $F$ features a $d$-fold multi-edge between different $H_\bullet$ for every edge in $G$.

We recall the definitions on p. 43, and express (*11) in the terms of the product graph $F$:

$$e(H', H'') \geq \frac{1}{2}\delta_G \cdot dD \cdot |I''| \tag{*12}$$

We will now work out various relationships between the selection sets $S'$ and $S''$, their complements, and $H'$, $H''$. The goal is to obtain a suitable restriction of $e(S, \bar{S})$ to convenient subsets about which it is easier to reason, like in the first two inequalities in (*4). In the end, we will be comparing the edge count between $S''$ and $\bar{S}'$ in order to arrive at a final estimation.

First, we observe that, because of the size restrictions, each $\bar{S}_j$ in $\bar{S}''$ (i.e. for $j \in I''$) will have a size of at most $\delta_G \cdot D/4$. Summing over $I''$ yields

$$|\bar{S}''| \leq \frac{1}{4}\delta_G \cdot D \cdot |I''| \tag{*13}$$

The nodes in $\bar{S}''$ are each fitted with a $d$-fold multi-edge in the second construction step of the replacement product, and therefore can support at most $d \cdot |\bar{S}''|$ edges between different $H_\bullet$. Since $\bar{S}'' \subset H''$, and $H' \cap H'' = \emptyset$, only at most $d \cdot |\bar{S}''|$ edges can connect $\bar{S}''$ with nodes $H'$, which yields

$$e(\bar{S}'', H') \leq \frac{1}{4}\delta_G \cdot dD \cdot |I''| \tag{*14}$$

We recall from the definitions above that $H'' = S'' \uplus \bar{S}''$. Thus, if the edge count between $H'$ and $\bar{S}''$ is limited by (*14), but the edge count between $H'$ and $H''$ has a lower bound from (*12), that means that at least the difference of those two estimates must contribute to the edge count between $H'$ and $H'' \setminus \bar{S}'' = S''$:

$$e(S'', H') \geq \frac{1}{4}\delta_G \cdot dD \cdot |I''| \tag{*15}$$

Up to now, we have proceeded from the expander property of $G$, via $e(H', H'')$ to $e(S'', H')$. Our aim now is to reduce $H'$ to $\bar{S}'$, because we want to obtain a formula expressing the edge count between some subset of the selection $S$ and some subset of its complement $\bar{S}$.

We recall that the selected nodes in $H'$ are fewer than $\delta_G \cdot |S|/10$, or our efforts would have stopped at the inequality in (*4). Because the formula in (*15) uses $D \cdot |I''|$, we first adapt the $S'$ size estimation accordingly. The authors of [ASS08] also replace the factor of (1/10) by (1/6), which means:

$$|S'| < \frac{1}{10}\delta_G \cdot |S| = \frac{1}{6}\delta_G \cdot D \cdot |I''| \cdot \left(\frac{6}{10} \cdot \frac{|S|}{D \cdot |I''|}\right) \tag{*16}$$

Now, since $D \cdot |I''|$ is the complete node count in $H''$, this term is a conservative upper bound for $|S''|$ (cf. the left-hand inequality in (*8)), which only counts the selected nodes in $H''$. We already determined in (*5) that $|S''|$ is at least nine-tenths of $|S|$. With this, we can estimate the bracketed term in (*16), canceling out $|S|$:

$$\frac{6}{10} \cdot \frac{|S|}{D \cdot |I''|} \leq \frac{6}{10} \cdot \frac{|S|}{\frac{9}{10}|S|} = \frac{2}{3} < 1$$

Thus, we may continue the inequality in (*16) by leaving the bracket term out completely:

$$|S'| < \frac{1}{6}\delta_G \cdot D \cdot |I''| \tag{*17}$$

Like before, we observe that the selected nodes in $S'$ are each connected with $d$-fold multi-edges to nodes in some $H_\bullet$. The size limit from (*17) therefore also limits the edge count between $S'$ and $H''$, which holds all the $H_\bullet$ necessarily disjoint with $S'$. Since $S'' \subset H''$, this also means:

$$e(S', S'') < \frac{1}{6}\delta_G \cdot dD \cdot |I''| \tag{*18}$$

This lower bound, together with $H' = S' \uplus \bar{S}'$, and with the estimate from (*15), means that that the edge count between $\bar{S}'$ and $S''$ must be at least the difference of the two bounds:

$$e(S'', \bar{S}') \geq \left(\frac{1}{4} - \frac{1}{6}\right)\delta_G \cdot dD \cdot |I''| = \frac{1}{12}\delta_G \cdot dD \cdot |I''| = \frac{1}{24}\delta_G \cdot 2d \cdot D|I''| \tag{*19}$$

We recall from above that $D|I''|$ is at least nine-tenth of $|S|$, which implies it is certainly larger than three-tenths of $|S|$, which allows us to finally state that

$$\boxed{e(S, \bar{S}) \geq e(S'', \bar{S}) \geq e(S'', \bar{S}') \geq \frac{1}{24}\delta_G \cdot 2d \cdot \frac{3}{10}|S| = \frac{1}{80}\delta_G \cdot 2d \cdot |S| \geq \frac{1}{80}\delta_G^2\delta_H \cdot 2d \cdot |S|}, \tag{*20}$$

where we introduced a factor of $\delta_G\delta_H \in [0, 1]$ in the last inequality in order to reproduce the statement of the theorem. This completes the proof for both cases of the set size $|S'|$. ■

## 3.4 Existence of $d$-Regular, $d$-Edge-Colorable Expanders

We introduced a simple way to construct $d$-regular and $d$-edge-colored bipartite graphs in subsection 2.2.2 (pp. 20ff., in lemma 2.29 and corollary 2.30), which yields a connected graph (with positive expansion $\delta$) for $d \geq 2$.

The authors of [ASS08] state (theorem 2) that there is some $\delta > 0$ such that there is a $d$-edge-colorable $[n, d, \delta]$-expander for any even $n$ and any $d \geq 3$. In lieu of a proof, they recommend considering random $d$-regular bipartite graphs, which we will do shortly (we will also expand on arbitrary random $d$-regular $d$-edge-colorable graphs).

However, if that $\delta$ were a numeric constant independent of $d$ and $n$, its value or some kind of bound would presumably have been offered at this point. Alon et al. mention a conference contribution by M. Pinsker[1] that proves the existence of constant-degree expanders.

We tend to interpret theorem 2 of [ASS08] differently, namely that for any $n$ and $d$ as specified above, there is a connected $d$-regular $d$-edge-colorable undirected graph. The connectedness then implies that its expansion parameter $\delta$ is positive. This we will, if not conclusively prove, at least strongly motivate in the following two subsections.

In addition, we would like to refer to work by B. Bollobás [Bol88] who demonstrates, translated to this work's vernacular, that for any $\varepsilon \in (0, \frac{1}{2})$, there is a degree $d$ such that almost all $d$-regular graphs will have expansion $\delta = \frac{1}{2} - \varepsilon$.

### 3.4.1 Random $d$-Regular $d$-Edge-Colored Bipartite Graphs

It can be shown [Fri21] that all $d$-regular bipartite (loop-free) graphs are $d$-edge-colorable. We will now demonstrate how to construct all possible $d$-regular and $d$-edge-colored bipartite graphs for given $n \in 2\mathbb{N}$ and $d \in \mathbb{N}$. The randomness alluded to in the subsection title then amounts to choosing one particular realization of such a graph; more on which shortly.

First, we stress that requiring $d$-regularity and the possibility for a $d$-edge-coloring simplifies construction because it implies that the resulting graph has no loops (multi-edges are, however, possible), and that every node will be incident with $d$ edges of all the colors $1 \leq c \leq d$.

For this subsection, we assume that the node subsets of the bipartite graph are $\{1, \cdots, \frac{n}{2}\}$ for node color 1, and $\{(\frac{n}{2} + 1), \cdots, n\}$ for node color 2. If this were not the case, there would be additional freedom for the random choice, but this would not lead to new graphs because of isomorphism (one could always find a permutation that re-labels the nodes according to the above assumption).

[1] On the Complexity of a Concentrator. $7^{th}$ Annual Teletraffic Conference (1973), pp. 1–4.

We recall the concept of half-edges from definition 2.1 (p. 9) right at the beginning of this work. In order to construct a $d$-regular $d$-edge-colorable graphs, we start by fitting every proto-node with $d$ half-edges, one each of the colors $\{1, \cdots, d\}$.

For the later graph to be $d$-edge-colorable, each half-edge of color $c$ of a node with node color $x$ needs to be connected to exactly one half-edge of color $c$ from a different node (loops are impossible to construct in this scenario). Because the graph is to be bipartite, this limits the choice of candidates to nodes of the respective other node color $(3 - x)$ that still have free (unconnected) half-edges of color $c$.

We propose a systematic approach, constructing all the $\frac{n}{2}$ edges of a single color, then moving on to the next color, etc. For each color, we must pair off the color-1 nodes with the color-2 nodes; in other words, we determine a bipartite *matching*. Because the node count is even, this will be a perfect matching, in that every node will be connected to exactly one other node. Equivalently, we construct an intermediate 1-regular graph (such graphs are always bipartite and 1-edge-colorable because they consist of node pairs joined by single edges; cf. figure 2.7, p. 19).

Because the two node sets are of equal size, it is always possible to create $\frac{n}{2}$ pairs such that every color-1 node is connected to exactly one color-2 node, and vice versa. We now recall the numbering we initially stipulated, and find that it suffices to choose a total bijective map of the numbers $\{1, \cdots, \frac{n}{2}\}$ on themselves, and then add $\frac{n}{2}$ to the mapping result. This will, for each node of color 1, determine exactly one of the color-2 nodes.

Such total bijections of finite number sets are called *permutations*, and are introduced in section B.1 (pp. 80ff.) of the appendix. In particular, we recall that the permutation for our problem belongs to $S_{n/2}$ (cf. definition B.2, p. 80), and that there are $(\frac{n}{2})!$ different permutations in the group $S_{n/2}$ (per lemma B.1, p. 80).

This can be repeated independently for all the $d$ edge colors. In the end, all the edges constructed in this way can be superimposed to yield a $d$-regular bipartite graph.

Given the initial subsets of nodes with the labeling as indicated above, this makes for

$$\left(\left(\frac{n}{2}\right)!\right)^d$$

different possible edge-colored graphs, which we may construct by drawing one partition of $S_{n/2}$ for each edge color.

If we were not interested in the actual edge-*coloring*, only the $d$-edge-*colorability*, we would have to subtract from that number all redundant graphs that share the same connectivity. Suffice it to remark that every concrete edge-coloring implies that a such a graph is also edge-colorable.

If the permutations are drawn at random, it is unlikely (the larger $d$, the more unlikely) that the exact same permutation is drawn multiple times. If we drew $d$ times the same permutation, we would end up with $\frac{n}{2}$ node pairs connected by $d$-fold multi-edges. This is suppressed by a factor of $((\frac{n}{2})!)^{1-d}$. There are other possibilities for disconnected resulting graphs, but for $d \geq 3$, many of the graphs constructed in this manner will be connected, and thus have positive expansion $\delta$.

We recall the construction presented in subsection 2.2.2 (pp. 20ff.): In that case, the permutations for color $c$ were just the powers $(1; \cdots; \frac{n}{2})^{c-1}$, producing cyclic shifts of $(1, \cdots, \frac{n}{2})$ by $(c-1)$ positions.

By way of an example, we reconstruct the cubic (3-regular) bipartite graph in figure 3.8 (p. 41). The matchings for the three colors are shown in figure 3.11 (p. 49). We list the three permutations (they can be reverse-engineered from figure 3.11 by subtracting 5 from the right-hand node labels):

| $j$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $\sigma_1(j)$ | 2 | 3 | 5 | 1 | 4 |
| $\sigma_1(j)+5$ | 7 | 8 | 10 | 6 | 9 |
| $\sigma_2(j)$ | 1 | 3 | 5 | 4 | 2 |
| $\sigma_2(j)+5$ | 6 | 8 | 10 | 9 | 7 |
| $\sigma_3(j)$ | 4 | 3 | 2 | 5 | 1 |
| $\sigma_3(j)+5$ | 9 | 8 | 7 | 10 | 6 |

Table 3.1: The bipartite cubic graph from figure 3.8. Permutations of the three colorings.

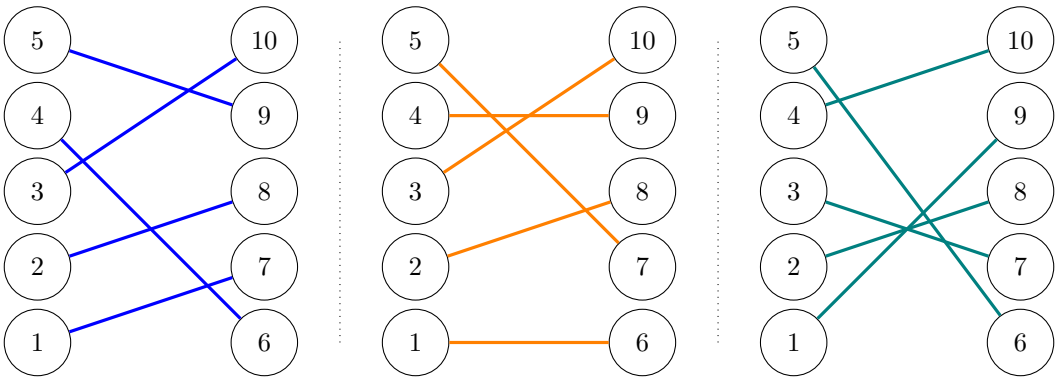


Figure 3.11: The bipartite cubic graph from figure 3.8. Edges with colors 1 (left), 2 (middle) and 3 (right).

### 3.4.2 Random $d$-Regular $d$-Edge-Colored Graphs

Our approach to general $d$-regular and $d$-edge-colored graphs is just a generalization of the previous technique; it has the same drawback in that it is easy to create all the possible colorings, but that there will be many redundant versions of the same $d$-edge-*colorable* graphs. However, this aesthetic caveat is not relevant to the expander construction in section 3.6, because that only will require the first connected $d$-regular $d$-edge-colorable graph for a given node count, and thus, the first $d$-edge-*colored* one will be sufficient for the task.

In the preceding subsection, we had stipulated that there is a given bipartition with the nodes $\{1, \cdots, \frac{n}{2}\}$ of color 1 and the others up to $n$ of color 2. If we were given just the $n$ nodes, we could construct different graphs, but only as regards labeling – because of isomorphism. The situation is different if the graph need not be bipartite.

We may again start with $n \in 2\mathbb{N}$ nodes, each fitted with $d \in \mathbb{N}$ colored half-edges. But in this case, there is no previous selection of the nodes into two classes. In order to create a $d$-regular and $d$-edge-colored graph, we still have to connect pairs of half-edges of the same color – which is always possible because $n$ is even. We also may treat the $d$ edge colors independently. For every color $c$, we have to construct a perfect matching of the $n$ nodes, i.e. a 1-regular graph consisting of node pairs, each connected by one $c$-colored edge.

However, here we may partition the nodes independently for each color. One way to do this is to draw a subset of size $\frac{n}{2}$ and treat its members as the color-1 nodes of a temporary bipartite graph; the subset's complement (of equal size) then holds its color-2 nodes. There are

$$\binom{n}{n/2} = \frac{n!}{(\frac{n}{2})!\,(\frac{n}{2})!}$$

different possible such subsets.

After this, we have the same situation as in the bipartite case of the preceding subsection. We may give all the nodes some intermediary labels such that the color-1 nodes are temporarily called $\{1, \cdots, \frac{n}{2}\}$, and the color-2 nodes, $\{\frac{n}{2}, \cdots, n\}$, respectively. We then can draw one of the permutations from $S_{n/2}$, and calculate the new edges of color $c$ by taking the permutation function's values and adding $\frac{n}{2}$ to each. After that, we can define the actual new edges of the graph, using the assignment table for the intermediary node labels (we will demonstrate this shortly with an example).

This, however, leads to an over-estimation of the number of $d$-edge-colored graphs which results from the arbitrary partitioning into intermediary color-1 and color-2 nodes.. If we draw all possible subsets with $\frac{n}{2}$ nodes, we double-define every edge, because for $j \neq k$, if we connect a color-1 node $j$ with a color-2 node $k$, this yields the same edge as if we connected a color-1 node $k$ to a color-2 node $j$. This holds for any of the $\frac{n}{2}$ pairs of nodes generated for each color, and we may correct for this by dividing the number of possible edge colorings by $2^{n/2}$.

Overall, this yields the following number of 1-edge-colorings:

$$\begin{aligned}\binom{n}{n/2} \cdot \left(\frac{n}{2}\right)! \cdot \frac{1}{2^{n/2}} &= \frac{n!}{2^{n/2}\,(\frac{n}{2})!} = \frac{n!}{2(\frac{n}{2}) \cdot 2(\frac{n}{2}-1) \cdot \cdots \cdot 2(1)} = \frac{n!}{(n)(n-2) \cdot \cdots \cdot (2)} \\ &= (n-1)(n-3) \cdot \cdots \cdot (1)\end{aligned}$$

(This is, by the way, equal to the number of permutations in $S_n$ that are composed of mutually disjoint (i.e. canonical) cycles of length 2 (i.e. transpositions), where we encounter the same situation that $\tau_{jk} = \tau_{kj}$. This is because the graph of canonical cycles is a perfect matching for those permutations, too. More on canonical cycles may be found in appendix section B.1, pp. 80ff.)

Thus, the overall number of $d$-regular, $d$-edge-colored graphs with $n \in 2\mathbb{N}$ nodes is

$$\left( \binom{n}{n/2} \cdot \left(\frac{n}{2}\right)! \cdot \frac{1}{2^{n/2}} \right)^d$$

As an example, we construct such a graph with $n = 10$ and $d = 3$. For convenience, we re-use the permutations listed in table 3.1 (p. 48), which leaves us with the choice of three partitions of $\{1, \cdots, 10\}$ into subset pairs of equal size, which we choose as

$$\begin{array}{lcl} S_1 := \{2,5,6,8,9\} & \Rightarrow & \bar{S}_1 := \{1,4,5,7,10\} \\ S_2 := \{1,2,3,8,10\} & \Rightarrow & \bar{S}_2 := \{4,5,6,7,9\} \\ S_3 := \{1,3,6,7,10\} & \Rightarrow & \bar{S}_3 := \{2,4,5,8,9\} \end{array}$$

For the intermediary labeling, we assign the labels in the order listed above (meaning, e.g., that we label the tuple $(2,5,6,8,9)$ as $(1,2,3,4,5)$, etc.), and obtain the following matchings (denoting nodes of temporary color-1 for edge color $c$ by $r_c$, and their counterparts of node color 2 by $s_c$; $j$ is used for the respective intermediary color-1 node label):

| $j$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $r_1$ | 2 | 5 | 6 | 8 | 9 |
| $s_1$ | 3 | 4 | 10 | 1 | 7 |
| $r_2$ | 1 | 2 | 3 | 8 | 10 |
| $s_2$ | 4 | 6 | 9 | 7 | 5 |
| $r_3$ | 1 | 3 | 6 | 7 | 10 |
| $s_3$ | 8 | 5 | 4 | 9 | 2 |

Table 3.2: A cubic graph with 3-edge-coloring. Matchings for the three colorings.

The graph edges are not from $j$ to $(\sigma_c(j) + \frac{n}{2})$ as in the bipartite case, but from $r_c(j) to s_c(j)$. We show the resulting graph in figure 3.12:

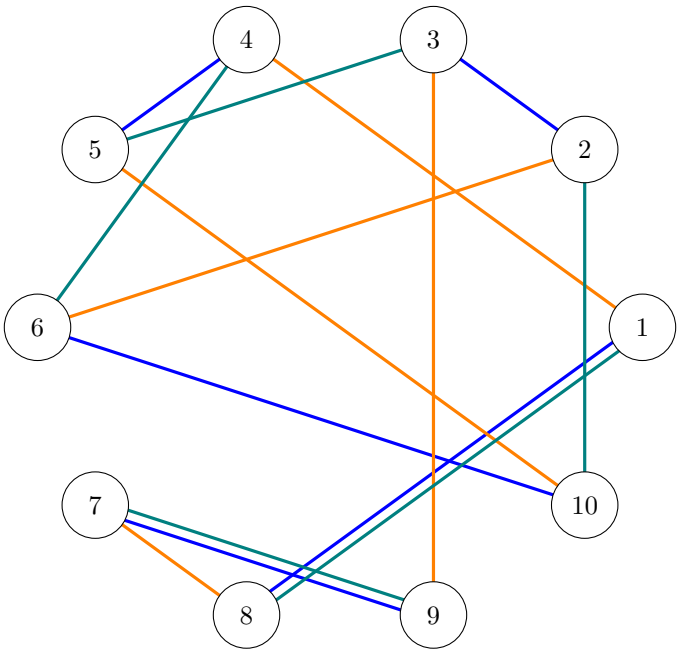


Figure 3.12: A cubic graph with 3-edge-coloring.

We observe that this graph is connected, but *not* bipartite because it contains a cycle of odd length (cf. theorem 1.4, p. 8, in [Nic18]). In particular, there is a cycle of length 3 containing the nodes 2, 6 and 10. If we colored node 2 with node color 1, then both nodes 6 and 10 would need to be node color 2, but they are directly connected by an edge; thus, the graph cannot be 2-node-colorable / bipartite (cf. definition 2.19, p. 14).

## 3.5 A Special Class of Expanders

At the start of section 2 in [ASS08], the authors introduce a special kind of parametrized graph, called $LD(q, r)$, which is a specialized version of a more general Cayley graph treated by N. Alon and Y. Roichman in [AR94]. The name evidently derives from the fact that their graphs have logarithmic diameter[2] (depending on their node count). Since this has no bearing on our problem, we will adapt the notation and denote such graphs with $G$.

Using the spectral gap theorem 3.1 (p. 32), Alon et al. show that, under certain conditions, those graphs are expanders with $\delta \geq \frac{1}{4}$. For this, we need a calculation of the largest two eigenvalues of such a graph's adjacency matrix.

The $G(q, r)$ graphs will later be used to construct constant-degree expanders in the following section 3.6, which will involve two replacement products (cf. sections 3.2f., pp. 36ff.). We recall that definition 3.6 (p. 37) stipulates that the left-hand graph is $D$-regular and $D$-edge-colorable.

However, the $G(q, r)$ graphs (as introduced shortly) will, in fact, not be edge-colorable at all (according to corollary 2.17, p. 14). because they contain loops. We will point out this (and another) difficulty when we have introduced the graphs, but then continue to follow the proof of a theorem on their second-largest eigenvalues (theorem 5 of [ASS08]).

After this, we propose an adaptations of not only the construction of $G(q, r)$ but also of the replacement product, which, in this special context, will fix the problem of edge-colorability and graph loops.

### 3.5.1 Constructing the Graphs $G(q, r)$ as Defined in [ASS08]

#### Preliminaries

For the following, we will rely on the contents of chapter E (pp. 107ff.), especially the section E.3 on Galois fields.

We recall the polynomial construction of a Galois field $GF(p^k)$ for $p$ prime and $k > 1$, using an irreducible modulus polynomial $m(x) \in \mathbb{Z}_p[x]$ of degree $k$ (according to lemma E.22, p. 114), which is necessary to facilitate multiplication. The arithmetics of $GF(p^k)$ are summarized in definition E.24 (p. 114).

We also recall from chapter E.3 that the elements of $GF(p^k)$ can be viewed as polynomials, which in turn may be represented by $p$-ary strings of coefficients, or coefficient tuples. A $p$-ary string of digits 0 to $(p-1)$ can be viewed as an encoding of a natural number; in fact, all the numbers 0 to $(p^k - 1)$ can be encoded using such strings – with the caveat that multiplication is taken modulo $m(x)$, and will yield different results than unbounded non-modulo multiplication if the factor polynomials have sufficiently large degrees. In those cases it is inevitable to consult the multiplication table, which is dependent on the choice of $m(x)$.

#### Relevant Algebraic Structures

For the construction in [ASS08], we let $t \in \mathbb{N}$, $q := 2^t$, and $\mathbb{F} := GF(q)$. Since 2 is prime, such Galois fields do exist, and the polynomial coefficients are from $\mathbb{Z}_2 = \{0, 1\}$. For binary strings (as representations of elements (polynomials) in $\mathbb{F}$), addition amounts to a bitwise `xor`. Also, there is no difference between addition and subtraction since, in $\mathbb{Z}_2$, $1 + 1 = 1 - 1 = 0$; this will be elementary for the following argument.

Since $\mathbb{F}$ is a field, we may construct a vector space (cf. definition A.7, p. 68) over $\mathbb{F}$, with the elements of $\mathbb{F}$ as scalars, and as components of vectors. In our case, we consider the Cartesian product $\mathbb{F}^{r+1}$, with $r \in \mathbb{N}$. Addition and scalar multiplication can be carried out component-wise like in the definitions A.9 and A.10 (p. 70) for Euclidean spaces.[3] Other than for the (non-finite) Euclidean space $\mathbb{R}^n$, addition and scalar multiplication behave according to the arithmetics of $\mathbb{F}$. This implies that the vector space $\mathbb{F}^{r+1}$ is finite itself: There are only $q$ different elements in $\mathbb{F}$. Any vector from $\mathbb{F}^{r+1}$ must therefore be one of the $q^{r+1}$ different vectors we can build. The arithmetics in $\mathbb{F}$ ensure that neither addition nor scalar multiplication (which can only involve components and scalar factors from $\mathbb{F}$), can lead outside the finite structure of $\mathbb{F}^{r+1}$.

[2] The diameter of an undirected graph is [Nic18] (p. 12) the maximum of distances between any two of its nodes; the distance being the length of a shortest path between two such nodes.

[3] $\mathbb{F}^{r+1}$ is, however not an Euclidean space itself – among other things, because we do not define any inner product (cf. definition A.8, p. 69).

We summarize the cascade of structures used in the following:

- $\mathbb{Z}_2 = \{0, 1\}$, the residue system modulo 2.
- $\mathbb{F} = GF(q) = GF(2^t) \cong \mathbb{Z}_2[x]_{m(x)}$, a finite field consisting of the $q$ polynomials of lesser degree than $t$, with coefficients from $\mathbb{Z}_2$, using an irreducible modulus polynomial $m(x)$ of degree $t$ for multiplication.
- $\mathbb{F}^{r+1}$, a finite vector space over $\mathbb{F}$ consisting of the $q^{r+1}$ different vectors with $(r + 1)$ components that can be constructed using only elements of $\mathbb{F}$.

**Small Example**

The smallest $t$ for which the polynomial structure of $GF(2^t)$ becomes relevant is $t = 2$, which means $q = 4$ and $\mathbb{F} = \{0, 1, 10, 11\}$ in binary string notation. There is only one irreducible polynomial of degree 2, namely $m(x) = x^2 + x + 1$, because $(x^2 + x)$ contains the factor $x$, and $(x^2 + 1)$ contains the factor $(x + 1)$: $(x + 1) \cdot (x + 1) = x^2 + x + x + 1 = x^2 + 1$. $m(x)$, on the other hand, does not have any nontrivial factors.

We represent $m(x)$ by the binary string 111 and derive the multiplication table (taking the modulus (by applying `xor` with $m(x)$) where product strings have length greater than 2, and remembering to calculate products with $\mathbb{Z}_2$ modulo rules, e.g. $11 \cdot 11 = 110 + 11 = 101$):

| $\cdot$ | 1 | 10 | 11 |
|---|---|---|---|
| 1 | 1 | 10 | 11 |
| 10 | 10 | 11 | 1 |
| 11 | 11 | 1 | 10 |

| $\cdot$ | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 1 | 2 | 3 |
| 2 | 2 | 3 | 1 |
| 3 | 3 | 1 | 2 |

Table 3.3: Multiplication in $\mathbb{F} = GF(2^2)$ with $m(x) = 111$. Binary strings (left) and corresponding numbers (right).

The structure $\mathbb{F}^2$ needed for the graph edges (see the definition below) contains the sixteen entries $(0, 0), \cdots, (11, 11)$.

In order to distinguish this from the vector space $\mathbb{F}^{r+1}$, we choose $r = 2$; thus, the finite vector space for the graph nodes (again, cf. the next definition) is $\mathbb{F}^3$, and consists of all the vectors with three components, each taken from $\mathbb{F}$; this makes for 64 elements:

$$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ 10 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ 11 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \cdots, \quad \begin{pmatrix} 11 \\ 11 \\ 11 \end{pmatrix}$$

We demonstrate the vector space operations. For addition, we recall that adding polynomials (binary strings) is performed component-wise, and the polynomial components are just the digits of the binary strings; therefore addition is bitwise, and modulo 2 (as per the rules of $\mathbb{Z}_2$). We add leading zeros to highlight the bitwise operations

$$\begin{pmatrix} 01 \\ 11 \\ 10 \end{pmatrix} + \begin{pmatrix} 01 \\ 01 \\ 11 \end{pmatrix} = \begin{pmatrix} 00 \\ 10 \\ 01 \end{pmatrix}$$

For scalar multiplication, we have to consult the multiplication table of $\mathbb{F}$:

$$11 \cdot \begin{pmatrix} 01 \\ 11 \\ 10 \end{pmatrix} = \begin{pmatrix} 11 \\ 10 \\ 01 \end{pmatrix}$$

This may serve as a demonstration that the operations in $\mathbb{F}^{r+1}$ always yield vectors belonging to $\mathbb{F}^{r+1}$, as it should be for any vector space, and that, due to the various modulo operations, the vector space is finite.

**Construction**

**Definition 3.11** *For $r, t \in \mathbb{N}$ and $q := 2^t$, the graph $G(q, r)$ is constructed in the following way:*

- *The nodes are identified by the $q^{r+1}$ vectors in $\mathbb{F}^{r+1}$, where $\mathbb{F} := GF(q)$.*

- *Two nodes $\vec{a}, \vec{b}$ are connected by an edge if there is a pair $(x, y) \in \mathbb{F}^2$ such that*

$$\vec{b} = \vec{a} + y \cdot \vec{p}_x, \qquad \textit{where} \qquad \vec{p}_x := \begin{pmatrix} 1 \\ x \\ x^2 \\ \dots \\ x^r \end{pmatrix}$$

  *The edge between those nodes has color $(x, y)$.*

(The powers of $x$ can be calculated by repeated multiplications in $\mathbb{F}$.)

The authors state that such a graph is $q^2$-edge-colorable because there are $q^2$ different pairs $(x, y)$, and because of

**Corollary 3.12** *For $(x, y) \in \mathbb{F}^2$, if $\vec{b} = \vec{a} + y \cdot \vec{p}_x$, then $\vec{a} = \vec{b} + y \cdot \vec{p}_x$.*

Proof: The vector equation in definition 3.11 stipulates that $\vec{a} = \vec{b} - y \cdot \vec{p}_x$, but since subtraction is carried out component-wise, the rules of $\mathbb{F}$ apply. In the polynomial representation of the Galois field $\mathbb{F}$, addition and subtraction are calculated coefficient-wise modulo $\mathbb{Z}_2$ – and, as we pointed out above, there is no difference between addition and subtraction in $\mathbb{Z}_2$. ■

The same pair $(x, y)$ connects $\vec{a}$ to $\vec{b}$ and $\vec{b}$ to $\vec{a}$ (not to any other node), and may therefore serve as edge color.

Also, because there are $q^2$ such different pairs, any node is incident with $q^2$ edges, making $G(q, r)$ a $q^2$-regular graph.

#### The Dilemma of Edge-Colorability

If we consider the edge function of $G(q, r)$ from definition 3.11, we observe that there are some cases where $\vec{b} = \vec{a}$, or, equivalently, $\vec{b} = \vec{a} + \vec{0}$. Since $\vec{p}_x$ cannot be the zero vector for any $x$ because of its first component, this is happens precisely when $y = 0$ – which holds for $q$ different pairs $(x, 0) \in \mathbb{F}^2$.

Thus, $G(q, r)$ as defined above has $q$ loops (because different pairs $(x, y)$ imply different edges) at each of its nodes, and $(q^2 - q)$ edges leading to other nodes, respectively. This is a problem in two ways:

- In our preliminaries chapter 2, we had expressly stated that edge-colorable graphs cannot have loops (edges connecting a node to itself) in corollary 2.17 (p. 14).
- Even if we permitted loops to carry colors, our notion of node degree would stipulate that $q$ loops with different colors contribute $2q$, making the graph $(q^2 + q)$-regular. This would be useless for the expander construction in the next section 3.6, because this involves replacement products (cf. definition 3.6, p. 37), and the $G(q, r)$ appearing on their respective left-hand sides would have to feature matching numbers for degree and edge-colorability.

The second point might be mitigated if we counted loops once, not twice, when calculating a node's degree. While this is possible, it would imply that a node's degree would no longer match the corresponding row (or column) sum of the graph's adjacency matrix (cf. definition 2.11, p. 13) – a property we already exploited above, particularly in the proof of lemma 2.34 (p. 23) when showing that a $d$-regular connected graph will only have one largest eigenvalue $d$ in its spectrum. Also, all the other eigenvalue calculations would require us to treat diagonal and off-diagonal elements of the adjacency matrix differently.

In addition, we would have to re-define the notion of a $d$-regular multi-graph. If loops counted once, then a single node with a loop would constitute a 1-regular graph, in contradiction to corollary 2.26 (p. 20) stipulating that odd-degree regular graphs must have an even node count.

We will propose a solution for both these problems in the subsection after next, but will continue for now according to [ASS08] because the question of edge coloring has no bearing on the next subsection.

### 3.5.2 Establishing the Expander Property of $G(q, r)$

The authors of [ASS08] prove a statement regarding the edge expansion of $G(q, r)$, using the spectral gap theorem 3.1 (p. 32). For this, we need the two largest eigenvalues of the adjacency matrix. The largest one ($\lambda_1$) should equal the graph's degree ($G(q, r)$ is regular). For $G(q, r)$ to be an expander, the second-largest eigenvalue needs to be smaller, i.e. $\lambda_2 < \lambda_1$.

### About the Eigenvalue Problem

The field $\mathbb{F}^{r+1}$ has $n := q^{r+1}$ elements that can readily be labeled to correspond to the integer numbers 1 to $n$, which may serve as node indices for the eigenvalue calculations.

Like in the preliminaries chapter 2, the adjacency matrix $A$ contains the connectivity data of the nodes; thus, $A \in \mathbb{N}_0^{n \times n} \subset \mathbb{R}^{n \times n}$, and the eigenvectors of $A$ will be from $\mathbb{R}^n$.

Because of the underlying structure, every component of an eigenvector refers to a node index, which (other than in chapter 2) corresponds to a vector in $\mathbb{F}^{r+1}$.

Now, the determinant of $A_\lambda$ would be quite unwieldy to calculate for large enough $n$ (we recall that $n = (2^t)^{r+1} = 2^{t(r+1)}$) – but we do not need the full spectrum of $A$. Alon et al. present a way to solve the eigenvalue problem for the two largest cases using the original equation as in definition C.1 (p. 100):

$$A \cdot \vec{v} = \lambda \vec{v}$$

We recall theorem C.10 (p. 102) which states that there is an orthogonal basis of $\mathbb{R}^n$ made up of eigenvectors of $A$, because $A$ is symmetric. This implies that if we find $n$ mutually orthogonal vectors that each satisfy the eigenvalue equation, all eigenvectors are accounted for.

### Eigenvectors for $A$

We recall that the node indices for $G(q, r)$ may be uniquely mapped to the vectors of $\mathbb{F}^{r+1}$ and vice versa; thus we may interchange an index $a$ and its corresponding vector $\vec{a}$ as desired.

**Definition 3.13** *For some arbitrary but fixed surjective linear map $L : \mathbb{F} \mapsto \{0, 1\}$, and for $a \in \{1, \cdots, n\}$, $n := q^{r+1}$, $q, r$ as in definition 3.11, a set of $n$ vectors $\{\vec{v}_a\}_a$ is defined over their respective components via*

$$(\vec{v}_a)_b = (\vec{v}_{\vec{a}})_{\vec{b}} := (-1)^{L(\sum_j a_j b_j)},$$

*where the sum is taken over all $j \in \{0, 1, \cdots, r\}$.*

Note that the sum yields an element of $\mathbb{F}$, not unlike a Euclidean scalar product.

The authors give an example for one such linear map, namely the map that retrieves the least significant bit of a binary string in $\mathbb{F}$. It is surjective because any field $\mathbb{F} = GF(2^t)$ will contain elements where the zeroth polynomial coefficient is 0 or 1 – in fact, exactly half of the elements end with 0, the others with 1. This is true even for $t = 1$, the special case where no polynomials are needed (cf. subsection E.3.4, p. 116).

For the following, two important properties of the $(\vec{v}_a)$ are shown:

**Corollary 3.14** *Let $\vec{a}$ be one of the vectors from definition 3.13. Then, for $b, c \in \{1, \cdots, n\}$:*

$$(\vec{v}_a)_{(b+c)} = (\vec{v}_a)_b \cdot (\vec{v}_a)_c$$

Proof: We plug in the behavior specified in definition 3.13, and exploit the linearity of $L$, i.e. $L(\alpha x + \beta y) = \alpha L(x) + \beta L(y)$:

$$\begin{aligned}
(\vec{v}_a)_{(b+c)} &= (-1)^{L(\sum_j a_j (b_j + c_j))} = (-1)^{L((\sum_j a_j b_j) + (\sum_j a_j c_j))} \\
&= (-1)^{L(\sum_j a_j b_j) + L(\sum_j a_j c_j)} = (-1)^{L(\sum_j a_j b_j)} \cdot (-1)^{L(\sum_j a_j c_j)} \\
&= (\vec{v}_a)_b \cdot (\vec{v}_a)_c \quad \blacksquare
\end{aligned}$$

**Lemma 3.15** *The vectors from definition 3.13 are mutually orthogonal.*

Proof: First, we recall that the vectors are defined in $\mathbb{R}^n$, so we may take any two such vectors and calculate their Euclidean scalar product. Since none of the vectors can be zero, they are orthogonal if (and only if) that product is zero. We therefore calculate

$$\langle \vec{v}_a, \vec{v}_b \rangle = \sum_{c=1}^{n} (\vec{v}_a)_c \cdot (\vec{v}_b)_c = \sum_{\vec{c} \in \mathbb{F}^{r+1}} (\vec{v}_{\vec{a}})_{\vec{c}} \cdot (\vec{v}_{\vec{b}})_{\vec{c}}$$

We plug in the definition and use the linearity of $L$ to calculate the sum terms:

$$\begin{aligned}
(\vec{v}_{\vec{a}})_{\vec{c}} \cdot (\vec{v}_{\vec{b}})_{\vec{c}} &= (-1)^{L(\sum_j a_j c_j)} \cdot (-1)^{L(\sum_j b_j c_j)} = (-1)^{(L(\sum_j a_j c_j)) + (L(\sum_j b_j c_j))} \\
&= (-1)^{L(\sum_j (a_j + b_j) c_j)}
\end{aligned}$$

If $a = b$, then $a_j = b_j$ for all $j$, and because addition is carried out bitwise in the polynomial coefficients: $a_j + b_j = 0$. But then $L(0) = 0$ because of the linearity of $L$, and all the sum terms in the scalar product are $(-1)^0 = 1$, leading to $\langle \vec{v}_a, \vec{v}_b \rangle = n > 0$.

If, however, $a \neq b$, then there is a fixed vector $\vec{d} := \vec{a} + \vec{b}$, such that

$$(\vec{v}_{\vec{a}})_{\vec{c}} \cdot (\vec{v}_{\vec{b}})_{\vec{c}} = (-1)^{L(\sum_j d_j c_j)} = (\vec{v}_{\vec{d}})_{\vec{c}}$$

Thus:

$$\langle \vec{v}_a, \vec{v}_b \rangle = \sum_{\vec{c} \in \mathbb{F}^{r+1}} (-1)^{L(\sum_j d_j c_j)}$$

Because the sum is taken over all $\vec{c}$ in $\mathbb{F}^{r+1}$, the exponent varies uniformly over $\{0, 1\}$ (i.e. both values occur equally often), which in turn means the sum terms vary uniformly over $\{-1, 1\}$, and that the total sum is zero. Therefore, $\langle \vec{v}_a, \vec{v}_b \rangle = 0$. ■

**Lemma 3.16** *The vectors from definition 3.13 (p. 54) are eigenvectors of the adjacency matrix $A(G(q, r))$.*

Proof: The vectors are mutually orthogonal as per lemma 3.15. If each of them satisfies the eigenvalue equation of $A := A(G(q, r))$, then the $\{\vec{v}_a\}_a$ constitute a complete set of eigenvectors of $A$.

For any $a \in \{1, \cdots, n\}$, we calculate the eigenvalue equation, and examine its $b$-th component:

$$(A \cdot \vec{v}_a)_b = \sum_{c=1}^{n} A_{bc} (\vec{v}_a)_c$$

Now, the matrix element $A_{bc}$ can only be non-zero if there is an edge between the nodes $\vec{b}$ and $\vec{c}$. In order to get the sum over all possible such edges, we gather from definition 3.11 (p. 52) that we have to consider all pairs $(x, y) \in \mathbb{F}^2$ for which $\vec{c} = \vec{b} + y \cdot \vec{p}_x$. We express this with a Kronecker symbol (cf. definition A.15, p. 72):

$$\cdots = \sum_{c=1}^{n} \left( \sum_{(x,y) \in \mathbb{F}^2} \delta_{\vec{c}, (\vec{b} + y \cdot \vec{p}_x)} \right) (\vec{v}_a)_c = \sum_{\vec{c} \in \mathbb{F}^{r+1}} \left( \sum_{(x,y) \in \mathbb{F}^2} \delta_{\vec{c}, (\vec{b} + y \cdot \vec{p}_x)} \right) (\vec{v}_{\vec{a}})_{\vec{c}}$$

We switch the sums and evaluate the Kronecker symbol, which fixes $\vec{c}$ to $\vec{b} + y \cdot \vec{p}_x$:

$$\cdots = \sum_{(x,y) \in \mathbb{F}^2} \sum_{\vec{c} \in \mathbb{F}^{r+1}} \delta_{\vec{c}, (\vec{b} + y \cdot \vec{p}_x)} (\vec{v}_{\vec{a}})_{\vec{c}} = \sum_{(x,y) \in \mathbb{F}^2} (\vec{v}_{\vec{a}})_{(\vec{b} + y \cdot \vec{p}_x)}$$

Using corollary 3.14 (p. 54), we obtain:

$$\cdots = \sum_{(x,y) \in \mathbb{F}^2} (\vec{v}_{\vec{a}})_{\vec{b}} \cdot (\vec{v}_{\vec{a}})_{(y \cdot \vec{p}_x)} = \left( \sum_{(x,y) \in \mathbb{F}^2} (\vec{v}_{\vec{a}})_{(y \cdot \vec{p}_x)} \right) \cdot (\vec{v}_{\vec{a}})_{\vec{b}} = \left( \sum_{(x,y) \in \mathbb{F}^2} (\vec{v}_{\vec{a}})_{(y \cdot \vec{p}_x)} \right) \cdot (\vec{v}_a)_b$$

Collecting all the components $b$ yields the eigenvalue equation:

$$A \cdot \vec{v}_a = \left( \sum_{(x,y) \in \mathbb{F}^2} (\vec{v}_{\vec{a}})_{(y \cdot \vec{p}_x)} \right) \cdot (\vec{v}_a) =: \lambda_{\vec{a}} \cdot \vec{v}_a \quad \blacksquare$$

**Eigenvalue Gap and Conclusion**

**Lemma 3.17** *For $G(q, r)$ as in definition 3.11 (p. 52) and $r < q$, the second-largest eigenvalue of $A(G(q, r))$ satisfies $\lambda_2 \leq rq$.*

Proof (theorem 5 in [ASS08]): We continue from above and examine the expression for $\lambda_{\vec{a}}$, with the shorthand

$$p_{\vec{a}}(x) := \sum_{j=0}^{r} a_j x^j = \sum_{j=0}^{r} a_j (\vec{p}_x)_j$$

Then:

$$\lambda_{\vec{a}} = \sum_{(x,y) \in \mathbb{F}^2} (\vec{v}_{\vec{a}})_{(y \cdot \vec{p}_x)} = \sum_{(x,y) \in \mathbb{F}^2} (-1)^{L(\sum_j a_j \cdot y \cdot (\vec{p}_x)_j)} = \sum_{(x,y) \in \mathbb{F}^2} (-1)^{L(y \cdot p_{\vec{a}}(x))}$$

The sum over $x$ is split into two parts – one where $p_a(x) = 0$, the other for non-zero $p_a(x)$:

$$\cdots = \left( \sum_{\substack{x \in \mathbb{F} \\ p_{\vec{a}}(x)=0}} \left[ \sum_{y \in \mathbb{F}} (-1)^{L(y \cdot p_{\vec{a}}(x))} \right] \right) + \left( \sum_{\substack{x \in \mathbb{F} \\ p_{\vec{a}}(x) \neq 0}} \left[ \sum_{y \in \mathbb{F}} (-1)^{L(y \cdot p_{\vec{a}}(x))} \right] \right)$$

In the left-hand contribution, the inner sum is $q$ times $(-1)^0$ because $p_{\vec{a}}(x)$ is zero, therefore it simplifies to $q$ times the roots of the polynomial function $p_{\vec{a}}(x)$ in $\mathbb{F}$.

For the right-hand contribution, we recall that $a$ (and its corresponding vector $\vec{a}$) is arbitrary but fixed; so we may use the same argument as in the proof of lemma 3.15 (p. 54) – because the sum in square brackets is over the whole field $\mathbb{F}$ and $p_{\vec{a}}(x)$ is fixed, $L$ varies uniformly over $\{0, 1\}$, and the square bracket evaluates to zero. Thus:

$$\lambda_{\vec{a}} = q \cdot \sum_{\substack{x \in \mathbb{F} \\ p_{\vec{a}}(x)=0}} 1$$

Now, for $\vec{a} = (0, \cdots, 0)^T \in \mathbb{F}^{r+1}$, all terms of the sum $p_{\vec{a}}(x) = \cdots$ are zero, so $p_{\vec{a}}(x)$ vanishes for every of the $q$ different $x \in \mathbb{F}$; this makes for $\lambda_1 := \lambda_{\vec{0}} = q^2$. If $\vec{a} \neq \vec{0}$, the polynomial given by $p_{\vec{a}}(x)$ is not the zero polynomial, but it may be reducible, i.e. it may have roots in $\mathbb{F}$. Since it is a polynomial of degree $r$, there cannot be more than $r$ such roots because $p_{\vec{a}}(x)$ can only have up to $r$ factors of degree 1. Therefore the sum can have at most $r$ terms, and $\lambda_{\vec{a}} \leq rq$; this holds for all $\vec{a} \neq \vec{0}$, and, consequently, for all corresponding eigenvalues, including $\lambda_2$.

According to our premise, $r < q$, so $\lambda_2 \leq rq < q^2 = \lambda_1$. ■

**Corollary 3.18** *For $G(q, r)$ as in definition 3.11 (p. 52) and $r \leq \frac{q}{2}$, $G(q, r)$ is a $[q^{r+1}, q^2, \frac{1}{4}]$-expander.*

Proof: Since $r \leq \frac{q}{2} < q$, we may employ lemma 3.17 (p. 55) and calculate an eigenvalue gap of

$$\lambda_1 - \lambda_2 \geq q^2 - \frac{q^2}{2} = \frac{q^2}{2}$$

Theorem 3.1 (p. 32) about the expander spectral gap then yields a lower bound for the graph's edge expansion (plugging in a degree of $q^2$):

$$\delta \geq \frac{\lambda_1 - \lambda_2}{2q^2} \geq \frac{1}{4} \quad ■$$

(Corollary 2.38 (p. 24) confirms that such a graph is connected.)

### 3.5.3 Solving the Edge-Colorability Dilemma

We now return to the two problems outlined above, namely the existence of loops in the $G(q, r)$ graphs and the conflicting methods of counting and/or coloring those.

First, we observe that the problem would not occur if the edge function of $G(q, r)$ (cf. definition 3.11, p. 52) were adapted to prevent any loops. We recall that $\vec{a}$ is connected to $\vec{b}$ if $\vec{b} = \vec{a} + y\vec{p}_x$ for some $x, y \in \mathbb{F}$. In order for the eigenvalue problem to be calculated as in [ASS08], we note two important characteristics:

- $\vec{b}$ should have an offset $\vec{a}$ so that corollary 3.14 (p. 54) can be used to separate the eigenvector and pull it out of the sum so that the eigenvalue equation holds.
- $(\vec{b} - \vec{a})$ should be linear in $y$ for the simplification of the sums, where it reads "$L$ varies uniformly". If we introduce additional contributions, this does no longer hold.

A second idea might be to restrict the range of $(x, y)$ to $\mathbb{F} \times (\mathbb{F} \setminus \{0\})$, which prevents any loops in $G(q, r)$ because $y \neq 0$. We could keep the eigenvalue calculation (with a small adaptation; see the following subsection), but we would end up with a graph of degree $q^2 - q$ (and the same number for its edge-colorability).

While this is $O(q^2)$, and even the expander property of corollary 3.18 could be maintained (in fact, due to the smaller degree, $\delta$ would even be larger than $\frac{1}{4}$), the following section requires degree

and edge-colorability to be exactly $q^2$ because the $G(q,r)$ graphs are involved in two replacement products (cf. definition 3.6, p. 37), where the degree of the left-hand factor's graph must match the node count of the right-hand factor's.

Thus, the restriction of $(x,y)$ alone cannot be sufficient – but it points us in a fruitful direction. We recall that in $G(q,r)$, every node has equally many loops. We may use this to define an alternative graph $\tilde{G}(q,r)$, constructed with the $(x,y)$ restriction mentioned just now, and with $\frac{q}{2}$ loops added to every node afterwards. Using our notion of a loop's contribution to the node degree, this adds $q$ to every node's degree, and, consequently, to the entire graph's degree, making $\tilde{G}(q,r)$ $q^2$-regular.

This can always be done because $q = 2^t$ is even for any $t \in \mathbb{N}$, so $\frac{q}{2}$ is integer. We will demonstrate the adapted eigenvalue calculation in the following subsection; only a small change is necessary.

Up to now, we do, however, remain with the problem of edge-colorability, which is (according to our definitions) not possible if a graph contains loops. We propose to circumvent this problem by making an ad-hoc exception only for the graphs $\tilde{G}(q,r)$, and only under the condition that those graphs occur as left-hand factors in a graph replacement product, which will have to be adapted slightly, too, in order to reflect the properties of $\tilde{G}(q,r)$. We will show, that such a product yields a loop-free graph that is not only of the desired degree but may readily be used as factor in another replacement product if so desired.

### Adapted Graphs $\tilde{G}(q,r)$

**Definition 3.19** *For a given $q = 2^t$, $t \in \mathbb{N}$, $\mathbb{F} = GF(q)$ and $r \in \mathbb{N}$, the graph $\tilde{G}(q,r)$ can be constructed as follows (remembering definition 3.11, p. 52):*

1. *Create $q^{r+1}$ nodes identified by the vectors $\vec{a} \in \mathbb{F}^{r+1}$.*
2. *For any node $\vec{a} \in \mathbb{F}^{r+1}$, create edges between nodes $\vec{a}$ and $\vec{b}$ for any $(x,y) \in \mathbb{F} \times (\mathbb{F} \setminus \{0\})$ satisfying*

$$\vec{b} = \vec{a} + y \cdot \vec{p}_x, \qquad \text{where} \qquad \vec{p}_x := \begin{pmatrix} 1 \\ x \\ x^2 \\ \cdots \\ x^r \end{pmatrix}$$

   *Color such edges with $(x,y)$*

3. *For any node $\vec{a} \in \mathbb{F}^{r+1}$, create $\frac{q}{2}$ bi-colored loops, carrying colors in consecutive pairs, i.e. $(0,0)$ and $(1,0)$ for loop 1, $(10,0)$ and $(11,0)$ for loop 2, etc., up to $(\mathrm{bin}(q-2),0)$ and $(\mathrm{bin}(q-1),0)$ for loop $\frac{q}{2}$.*

**Corollary 3.20** *The graphs $\tilde{G}(q,r)$ from definition 3.19 are $q^2$-regular and $q^2$-edge-colorable (using an ad-hoc exception allowing for bi-colored loops).*

Proof: Step 2 adds $q^2 - q$ edges and uses up as many edge colors $(x,y)$, which is all those with $y \neq 0$. Step 3 adds $\frac{q}{2}$ loops, i.e. $q$ incident half-edges to each node, making each node's degree $q^2$, which yields the stated graph degree.

Those $q$ half-edges are colored with the remaining $q$ colors $(x,0)$, $x \in \mathbb{F}$. The ad-hoc exception mentioned in the statement consists in allowing half-edges of different colors to be connected to form a *bi-colored loop*. This is, of course, not a properly edge-colored graph, but, as we mentioned before, the exception is only granted on the proviso that such a graph appears as left-hand factor in an adapted replacement product (see below).

This procedure uses exactly the $(q^2-q)+q = q^2$ edge colors in $\mathbb{F}^2$ for every node, which makes the graph $q^2$-edge-colorable. ■

### Adapted Replacement Product

**Definition 3.21** *For a $q^2$-regular and $q^2$-edge-colorable graph $\tilde{G}(q,r)$ as per definition 3.19 with $q^{r+1}$ nodes, and a $d$-regular graph $H$ with $q^2$ nodes, the* replacement product *$\tilde{G}(q,r) \circ H$ obtained by the following procedure:*

1. *For any node $\vec{a} \in \mathbb{F}^{r+1}$ of $\tilde{G}(q,r)$, let $H_{\vec{a}}$ be a copy of $H$ with nodes $k_{\vec{a}} \in \mathbb{F}^2$, and with all the edges of $H$ reproduced in each copy.*
2. *For all edges of $\tilde{G}(q,r)$ with color $(x,y) \in \mathbb{F} \times (\mathbb{F} \setminus \{0\})$, between nodes $\vec{a}$ and $\vec{b} \neq \vec{a}$, add a $d$-fold multi-edge between the nodes $k_{\vec{a}} := (x,y)$ and $k_{\vec{b}} := (x,y)$ in the copies $H_{\vec{a}}$ and $H_{\vec{b}}$.*
3. *For all nodes $\vec{a} \in \mathbb{F}^{r+1}$ and all their respective bi-colored loops in $\tilde{G}(q,r)$ with color pairs $(c_1,c_2) \in (\mathbb{F} \times \{0\})^2$, add a $d$-fold multi-edge between the nodes $k_{\vec{a}} = c_1$ and $k_{\vec{a}} = c_2$ in the copy $H_{\vec{a}}$.*
4. *$\tilde{G}(q,r) \circ H$ consists of all the nodes in the $q^{r+1}$ copies of $H$, with the connectivity as described in the previous three steps.*

**Corollary 3.22** *A graph $\tilde{G}(q,r) \circ H$ as in definition 3.21 is $(2d)$-regular. It is loop-free if $H$ is.*

Proof: Steps 1 and 2 are covered by corollary 3.7 (p. 37); thus, all the nodes $(x,y)$ with $y \neq 0$ have degree $(2d)$, for every copy $H_{\vec{a}}$.

Step 3 adds $d$-fold multi-edges between nodes of different internal indices $c_1 \neq c_2$ within copies $H_{\vec{a}}$. Those nodes were not affected by step 2, and they are different by virtue of the construction of $\tilde{G}(q,r)$. There are $\frac{q}{2}$ pairs of such nodes, and they each are incident with an additional $d$-fold multi-edge after step 3.

Thus, every node of $\tilde{G}(q,r)$ has degree $2d$ after step 3.

If $H$ is loop-free, step 1 creates a loop-free graph. All the loops of $\tilde{G}(q,r)$ are bi-colored and are used to add $d$-fold multi-edges in copies of $H$ with different node indices (determined by the relevant loop's two colors). As per the construction of $\tilde{G}(q,r)$, the two colors of such loops are never equal; thus, neither are the nodes connected in step 3. ■

### Example for the Adapted Replacement Product

Because the $\tilde{G}(q,r)$ are large graphs, we choose to present a simpler example to illustrate the adapted replacement product from above, namely a 5-regular bipartite graph $G$ with some bi-colored loops, and the cycle graph $C_5$. We would like to point out that this is not a permitted exception to the notion of edge-coloring, and is only to serve as a demonstration of the principles in definition 3.21. The graphs $G$ and $H$ are shown in figure 3.13, and the resulting product in figure 3.14 (both: p. 59).

We observe that the product is 4-regular and loop-free, as expected from the above. When creating an edge-coloring for the bipartite graph $G$, it is important that the multiplicities of the colors in loops match on for each node class: Creating a loop on a color-1 node takes away two edge colors, say ($a$ and $b$), which will not feature in edges to a color-2 node. Therefore, some color-2 nodes need to have incident half-edges colored $a$ and $b$ belonging some loops (not necessarily the same one), too, so that edges to color-1 nodes with those colors are missing there as well. This informs why we enforced a rigid coloring scheme for the loops in the $\tilde{G}(q,r)$ definition above.

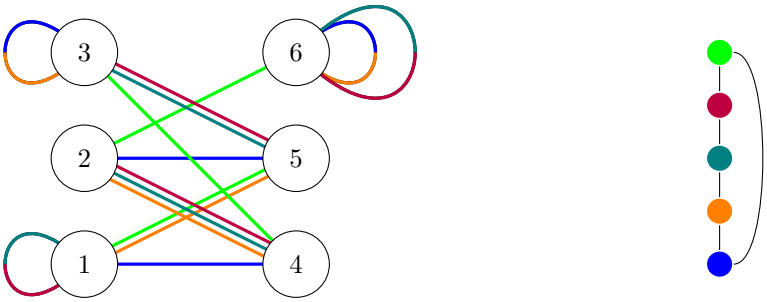


Figure 3.13: A 5-regular bipartite graph with bi-colored loops $G$ (left) and $H := C_5$ (right, shorthand)

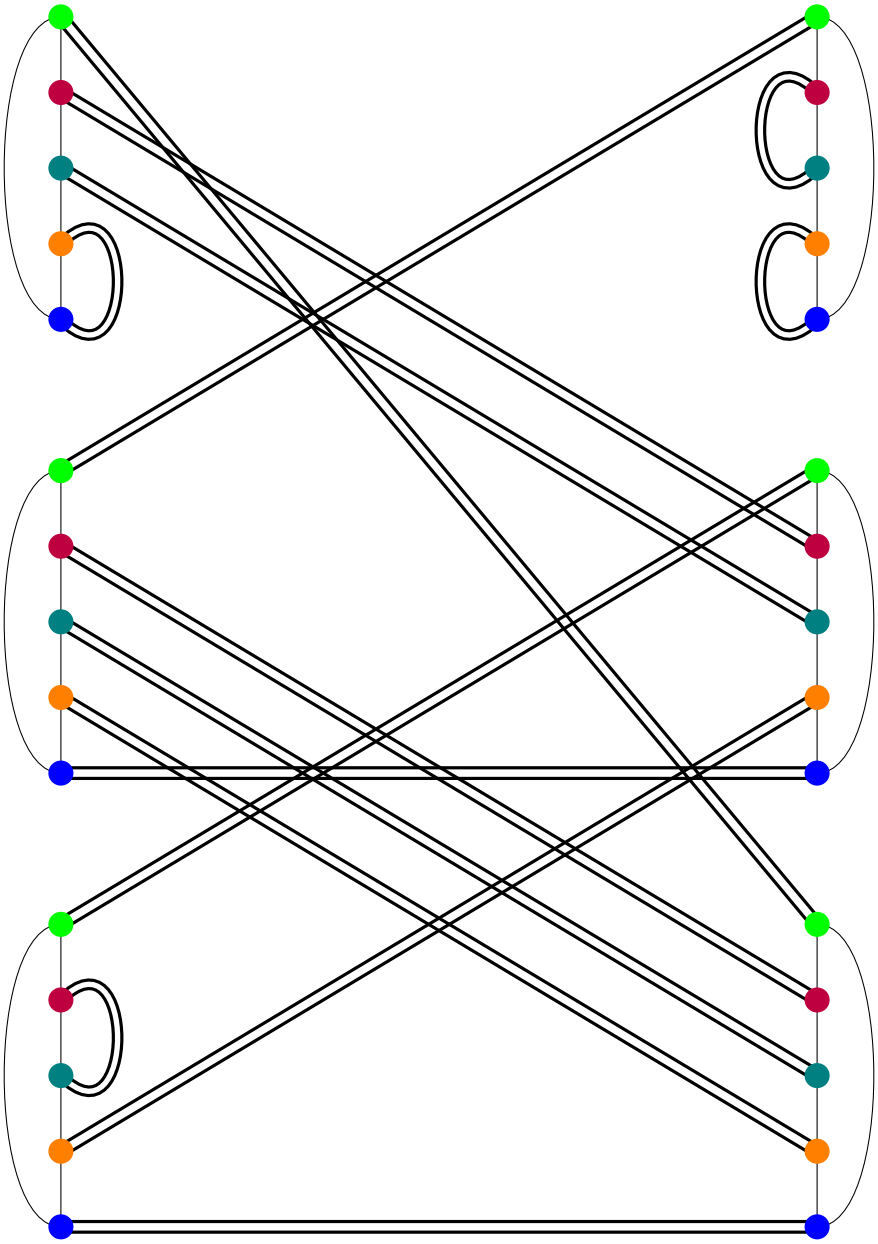

Figure 3.14: $G \circ H$

**Adapted Eigenvalue Problem**

We show that the eigenvalue problem solved in lemmas 3.16 (p. 55) and 3.17 (p. 55) needs only a small adaption to reflect the new graphs $\tilde{G}(q,r)$.

**Lemma 3.23** *The eigenvalues and eigenvectors of $A(\tilde{G}(q,r))$ are identical to those calculated for $G(q,r)$ in lemmas 3.16 and 3.17.*

Proof: We split the sum occurring in the left-hand side of the eigenvalue equation into two parts:

$$(A \cdot \vec{v}_a)_b = \sum_{c=1}^{n} A_{bc}(\vec{v}_a)_c = \left( \sum_{c \in \{1,\cdots,n\} \setminus \{b\}} A_{bc}(\vec{v}_a)_c \right) + A_{bb}(\vec{v}_a)_b$$

$A_{bb}$ contains the number of half-edges belonging to loops, which is $q = 2 \cdot \frac{q}{2}$. The vector component outside the bracket is already of the desired form for the right-hand side of the eigenvalue equation.

We now deal with the sum in brackets. Here, the vectors $\vec{b}$ and $\vec{c}$ are different, so the matrix elements $A_{bc}$ are the sum of non-loop edges incident at $\vec{b}$. As above, those sums can be determined by considering all pairs $(x,y) \in \mathbb{F}^2$ with a Kronecker delta. There is no need to restrict $y$ at this point because the contribution $c = b$ is not part of the outer sum. Evaluating the sum over $c$ with the Kronecker delta, then, yields the $y$ restriction:

$$\left( \cdots \right) = \sum_{\substack{x \in \mathbb{F} \\ y \in \mathbb{F} \setminus \{0\}}} (\vec{v}_{\vec{a}})_{(\vec{b} + y \cdot \vec{p}_x)} = \left( \sum_{\substack{x \in \mathbb{F} \\ y \in \mathbb{F} \setminus \{0\}}} (\vec{v}_{\vec{a}})_{(y \cdot \vec{p}_x)} \right) (\vec{v}_a)_b$$

Combining these results yields:

$$A \cdot \vec{v}_a = \left( q + \sum_{\substack{x \in \mathbb{F} \\ y \in \mathbb{F} \setminus \{0\}}} (\vec{v}_{\vec{a}})_{(y \cdot \vec{p}_x)} \right) \cdot \vec{v}_a =: \lambda_a \cdot \vec{v}_a$$

Using the same polynomial abbreviation $p_{\vec{a}}(x)$ as in lemma 3.17, we have:

$$\lambda_a = q + \sum_{\substack{x \in \mathbb{F} \\ y \in \mathbb{F} \setminus \{0\}}} (-1)^{L(y \cdot p_{\vec{a}}(x))}$$

In order to exploit the uniformity of the values for $L$, it is desirable to sum $y$ over the whole of $\mathbb{F}$, including 0, such that all binary patterns occur in positive and inverted form; otherwise the symmetry argument would not work. We therefore subtract and add the contributions $y = 0$:

$$\lambda_a = q + \left( \sum_{(x,y) \in \mathbb{F}^2} (-1)^{L(y \cdot p_{\vec{a}}(x))} \right) - \sum_{x \in \mathbb{F}} (-1)^{L(0 \cdot p_{\vec{a}}(x))}$$

Now, the sum outside the bracket is just $q$ times $(-1)^0$, which cancels the $q$ before the bracket, and leaves us with exactly the same expression as in the proof of lemma 3.17 – which in turn yields the same observations on the eigenvalues (including the edge expansion as stated in corollary 3.18, p. 56). ■

Therefore, we may use the graphs $\tilde{G}(q,r)$ instead of $G(q,r)$ for the following deliberations.

## 3.6 Constructing a Constant-Degree Expander in Polynomial Time

We now combine all the previous results to show that it is possible to construct expander graphs in polynomial time (i.e. regarding their final node count).

The first step (as outlined in [ASS08] (theorem 4)) is to search for a basic expander like described in section 3.4 (pp. 47ff.). This may seem counterintuitive at first glance, because the strategies from that section already yield expanders with constant degree (which is just the number of edge colors) – that is, until we recall that the algorithm is randomized, so there is no control over the

resulting graph's expansion parameter. If we aim for a certain threshold for $\delta$ in our expander, we are not only forced to repeat the process until a satisfactory graph has been found, but, more crucially, we have to calculate the isoperimetric constant for every graph. Since we cannot presume any symmetries, we have to check all possible partitions for their count of connecting edges – this takes exponential time.[4] This quickly becomes infeasible for large enough node count.

On the other hand, the graphs $\tilde{G}(q, r)$ from the preceding section can certainly be made very large, and even satisfy $\delta \geq \frac{1}{4}$ – but their degree is dependent on the node count and therefore not constant.

Alon et al. propose to combine the two scenarios by using the replacement product twice – we will show that one such product would not be sufficient for the desired polynomial construction time, and motivate their choice of parameters.

### Finding an Initial Expander $H$

Since the replacement product will have twice the degree of its right-hand factor, it is not necessary to start with a high degree, particularly when anticipating that two such products will be necessary. The minimum degree of any "serious" expander is three, because a (connected) degree-2 expander would just be a big cycle graph, with poor $\delta$.

Because $H$ is going to be a right-hand factor of a replacement product, its node count is tied to the future left-hand factor's degree (and edge-colorability). The $\tilde{G}(q, r)$ have degree $q^2$; thus, the authors of [ASS08] recommend searching for a 3-edge-colorable $[q^2, 3, \delta_H]$-expander $H$ with a suitable value of $\delta_H$. We recall from above that $q = 2^t$ for some integer $t$. Since $q^2$ is always even, the remarks of section 3.4 apply, and we may find such a graph in exponential time, regarding its size:

**Lemma 3.24** *For $q = 2^t$, $t \in \mathbb{N}$, and some value $\delta$, a 3-edge-colorable $[q^2, 3, \delta_H]$-expander $H$ can be found in time $q^{O(q^2)}$.*

Proof: We provide a rough estimation. As we pointed out above, an isoperimetric check will cost $2^{q^2}$ steps. Because we cannot expect to hit upon a suitable expander on the first attempt, this number will be multiplied with some fraction of the number of possible 3-edge-colored 3-regular graphs, which is of order $((q^2)!)^3$.

We use Stirling's formula ([K$^+$88]) to estimate $(q^2)!$ is of order $q \cdot (q^2)^{q^2} = q^{1+2q^2}$.

Combining this, we obtain a construction time of order $2^{q^2} \cdot q^{3(1+2q^2)} = q^{O(q^2)}$. ■

### Applying the First Replacement Product

**Lemma 3.25** *Given a 3-edge-colorable $[q^2, 3, \delta_H]$-expander $H$ for some $q = 2^t$ as in lemma 3.24, one replacement product $G_1 := \tilde{G}(q, r_1) \circ H$, with an expander $\tilde{G}(q, r_1)$ for some $r_1$, will not suffice to ensure a construction time polynomial in the node count of $G_1$.*

Proof: We choose an integer $r_1 \leq \frac{q}{2}$ in order to ensure that $\tilde{G}(q, r_1)$ is a $[q^{r_1+1}, q^2, \frac{1}{4}]$-expander, recalling corollary 3.18 (p. 56).

Theorem 3.10 (p. 42) then stipulates that the product graph $G_1$ is a $[q^{r_1+3}, 6, \delta_1]$-expander. Its node count is of order $q^{r_1}$, and since $r_1 \leq q$, this yields $q^{O(q)}$.

But $\text{poly}(q^{O(q)})$ is less than $q^{O(q^2)}$, which could only be reached by an additional exponentiation with $O(q)$, not a polynomial of $q^{O(q)}$. ■

### Parameter Considerations, and the Second Replacement Product

Because of lemma 3.25, we will endeavor to take a second replacement product, with a graph $G_2 := \tilde{G}(q_2, r_2) \circ G_1$. Since $\tilde{G}(q_2, r_2)$ is to be a $[q_2^{r_2+1}, q_2^2, \frac{1}{4}]$-expander, the replacement product requires that $q_2^2 = q^{r_1+3}$, which can be used to fix a value for $r_1$.

Before that, we formulate our aim: If the node count of $\tilde{G}(q_2, r_2)$ is of order $q^{\text{poly}(q^2)}$, bounded in both directions, then the lower bound ensures that $q^{O(q^2)}$ is at most polynomial in that node count, and cannot be worse than polynomial in the node count of the final graph $G_2$ (which we rename later). If we can meet this aim with a choice of parameters, just two replacement products, applied to $H$, are sufficient – which in turn ensures that the initial expansion parameter of $H$, while greatly reduced, does not become infinitesimal.

[4] If we just were interested in connectedness, i.e. $\delta > 0$; this could be checked in linear time using depth-first or breadth-first search.

Now, $\tilde{G}(q_2, r_2)$ has $q_2^{r_2+1}$ nodes, and if is to be an expander as stated above, $r_2 \leq \frac{q_2}{2}$. We recall from above that

$$q_2^2 = q^{r_1+3} \quad \Rightarrow \quad q_2 = q^{\frac{r_1+3}{2}},$$

which allows for any odd integer $r_1$. The first solutions are:

- $r_1 = 1 \quad \Rightarrow \quad q_2 = q^2$
- $r_1 = 3 \quad \Rightarrow \quad q_2 = q^3$
- $r_1 = 5 \quad \Rightarrow \quad q_2 = q^4$

If we fix $r_2$ to be of order $q_2$ (in both directions), then the node count of $\tilde{G}(q_2, r_2)$ is bounded in both directions by

$$\left(q^{(r_1+3)/2}\right)^{q^{(r_1+3)/2}} = q^{\left(\frac{r_1+3}{2} \cdot q^{(r_1+3)/2}\right)}$$

Now, since $q^2$ is only linear in $q^2$, and $q^3$ is not an integer power of $q^2$, the first safe and aesthetically pleasant choice would be $q_2 := q^4$. Thus, we fix

$$r_1 := 5; \qquad q_2 := q^4; \qquad p \cdot q^4 \leq r_2 \leq \frac{q^4}{2} \text{ with } p < \frac{1}{2} \text{ allowing for some integer choices for } r_2$$

This in turn means that $\tilde{G}(q, r_1) = \tilde{G}(q, 5)$ is of type $[q^6, q^2, \frac{1}{4}]$, and that $\tilde{G}(q_2, r_2) = \tilde{G}(q^4, r_2)$ is a $[q^{4r_2+4}, q^8, \frac{1}{4}]$-expander.

We observe that corollary 3.18 (p. 56) states that $\tilde{G}(q, 5)$ has expansion of at least $\frac{1}{4}$ if $5 \leq \frac{q}{2}$, so $q = 2^t$ with $t \geq 4$.

We may relax this to $5 < q$, i.e. $t \geq 3$ due to lemma 3.17 (p. 55). In the latter case, $\tilde{G}(q, 5)$ still has positive edge expansion, but no longer with a constant lower bound. If we only demand that $r_1 < q$, the second-largest eigenvalue will be $\lambda_2 \leq q^2 - q$, which leads to an eigenvalue gap of, not at least $\frac{q^2}{2}$, but of at least $q$, and consequently an expansion of at least $\frac{1}{2q^2}$. In the border case $q = 8$, this yields a lower bound of $\frac{1}{128}$ for the expansion parameter.

We also observe from this example the following

**Corollary 3.26** *The replacement product from definition 3.6 (p. 37) is not associative; nor is the adapted product from definition 3.21 (p. 58).*

Proof: By counterexample. In the above remarks, we construct

$$G_2 := \tilde{G}(q^4, r_2) \circ G_1 = \tilde{G}(q^4, r_2) \circ \left(\tilde{G}(q, 5) \circ H\right)$$

However, the product $(\tilde{G}(q^4, r_2) \circ \tilde{G}(q, 5))$ is not defined because the left-hand graph is of degree $q^8$, but the right-hand one's node count is only $q^6 \neq q^8$. Thus, we may not change the brackets, i.e. the order of product calculations. ■

We close our observations by stating that $G_1$, as per the above, will be a $[q^8, 6, \delta_1]$-expander, and $G_2$, a $[q^{4r+12}, 12, \delta_2]$-expander.

**Final Statement**

All the previous results lead to theorem 4 of [ASS08] (modified to include our restriction of $t$):

**Theorem 3.27** *There is a $0 < \delta < 1$ such that for any integer $t > 2$, $q := 2^t$, and any integer $r$ satisfying*

$$\frac{q^4}{100} \leq r \leq \frac{q^4}{2},$$

*there is an expander $E$ of type $[q^{4r+12}, 12, \delta]$ that can be constructed in polynomial time (regarding its node count). $E$ can be designed to be 12-edge-colorable.*

Proof: The statement combines the contents of this section. $E$ is just a re-labeled $G_2$:

$$E := \tilde{G}(q^4, r) \circ \left(\tilde{G}(q, 5) \circ H\right),$$

for some 3-regular and 3-edge-colorable expander $H$ of type $[q^2, 3, \delta_H]$ as described above. Because all factors of the replacement products involved in the construction are edge-colorable with exactly as many colors as the graph's respective node counts, by virtue of corollary 3.8 (p. 37), the same applies to the final graph $E$ from theorem 3.27: it is 12-regular and 12-edge-colorable.

Alon et al. also point out that the construction of the $\tilde{G}$ graphs requires time because representations of $GF(q) = GF(2^t)$ and $GF(q^4) = GF(2^{4t})$ need to be found. This requires a search for irreducible modulus polynomials of degree $t$ and $4t$, respectively. However, the number of possible polynomials of such degrees is bounded by the respective field sizes; even the naive brute-force method from section E.3 is benign in that respect. Building the multiplication tables is also bound by the field sizes, such that all this effort is negligible compared to building the vector spaces over the respective fields (which itself is linear in the node respective counts, and therefore subsumed in polynomial order of $E$'s node count). ■

## 3.7 Specializations

### 3.7.1 Narrowing the Node Count

Theorem 3.27 (p. 62) established that 12-regular, 12-edge-colorable expanders of type $[q^{4r+12}, 12, \delta]$ exist for integer $t > 2$ and $q := 2^t$, and can be constructed in polynomial time regarding their node count. We now present a method by Alon et al. to construct 12-regular and 12-edge-colorable expanders with node count $\Theta(n)$[5] for any given $n$ that is sufficiently large, using the construction from the theorem.

First, we define sets of integers that can be exact node counts for the graphs constructed in theorem 3.27, for $t \in \mathbb{N}$:

$$N_t := \left\{ q^{4r+12} \,\middle|\, q = 2^t \quad \wedge \quad r \in \mathbb{N} \cap \left[ \frac{q^4}{100}, \frac{q^4}{2} \right] \right\}$$

If we recall the restriction $t > 2^6$, this means that the available node degrees are

$$\bigcup_{t>2} N_t$$

The smallest of these is

$$\min N_3 = 8^{4\frac{8^4}{100}+12} \approx 6.3 \cdot 10^{158}$$

(For $t \geq 2$, we would have obtained $\min N_2 \approx 2.5 \cdot 10^{13}$)

We now show that the intervals covered by the $N_t$ sets have no gaps for $t > 1$ (which is satisfied anyway since we demand $t > 2$); thus, all the possible node counts in question lie in some $N_t$. For this, we demonstrate that $\max N_t > \min N_{t+1}$, remembering that $2q = 2^{t+1}$

$$\begin{aligned} \max N_t &= q^{4\frac{q^4}{2}+12} = (2^t)^{2q^4+12} = 2^{t(2q^4+12)} \\ \min N_{t+1} &= (2q)^{4\frac{(2q)^4}{100}+12} = (2^{t+1})^{\frac{64}{100}q^4+12} = 2^{(t+1)\cdot(\frac{64}{100}q^4+12)} \end{aligned}$$

Because the exponential function is strictly monotonous, and increasing for positive base, it is sufficient to compare the exponents:

$$\begin{aligned} t(2q^4+12) - (t+1)\cdot\left(\frac{64}{100}q^4+12\right) &= q^4\left(2t - \frac{64}{100}t - \frac{64}{100}\right) + (12t - 12t - 12) \\ &= q^4\left(\frac{136}{100}t - \frac{64}{100}\right) - 12 \end{aligned}$$

For $t = 1$ this expression yields $16 \cdot \frac{72}{100} - 12 = \frac{1152}{100} - 12 < 0$, but for $t = 2$ the factor $q^4 = 2^{4t}$ already dominates. The bracketed term becomes negligible because it is larger than 1 but only linear in $t$, and the expression is larger than $(256 - 12) > 0$.

Thus, for $t \geq 2$ (and certainly for $t > 2$), $\max N_t > \min N_{t+1}$.

(We observe that the overlap of the $N_t$ sets depends on the lower bound of $r$, and may retroactively motivate the choice of $\frac{q^4}{100}$, which appeared somewhat arbitrary in theorem 3.27 (p. 62).)

[5] This means that the actual node count is bounded in both directions by numbers linear in $n$

[6] The authors of [ASS08] do not restrict $t$ in the theorem and therefore continue here with $t \geq 2$

Consequently, for $n \geq \min N_3$, there is some $t > 2$ for which $n \in [\min N_t, \max N_t]$. For this $t$, we obtain a $q := 2^t$ and a range for possible $r$ values. We choose $r_0$ to be the one satisfying

$$q^{4r_0+12} \leq n \leq q^{4(r_0+1)+12} = q^4 \cdot q^{4r_0+12}$$

Dividing the right-hand inequality by $q^4$ yields

$$\frac{n}{q^4} \leq q^{4r_0+12} \leq n \qquad (*)$$

We define the node count for an expander $E_0$ according to theorem 3.27 with those values for $q$ and $r_0$ as

$$n_0 := q^{4r_0+12}$$

The right-hand inequality in (*) yields $n_0 = O(n)$. If $q^4$ were constant (most importantly: independent of $n$), then $n_0$ would also be $\Omega(n)$, and thus $\Theta(n)$.

As $q$ certainly depends on $n$, we must fix the lower bound in another way. Alon et al. distinguish two cases:

- Either $n_0 \geq \frac{n}{32}$; then we do have a lower bound linear in $n$. An expander $E_0$ constructed according to theorem 3.27 with $n_0$ nodes is acceptable because $n_0 = \Theta(n)$.
- Or $n_0 < \frac{n}{32}$, meaning that we will have to exploit the left-hand inequality in (*), which reads $n \leq q^4 \cdot n_0$, in some way in order to fix a lower bound for the expander node count.

  We construct $E_0$ anyway, but will modify it in a way that the resulting expander indeed has node count $\Theta(n)$.

  First, we define a ratio

  $$\varrho := \left\lfloor \frac{1}{16} \frac{n}{n_0} \right\rfloor < q^4,$$

  where we have used (*) for the upper bound.

  We now find a 6-edge-colorable, 6-regular expander $H$ of type $[12\varrho, 6, \delta_H]$. Since $\varrho < q^4$, lemma 3.24 (p. 61) ensures this can be done brute-force in time $q^{O(q^4)}$ (the different degree only accounts for a constant factor in the exponent).

  Since the expander $E_0$ satisfies $r_0 > \frac{q^4}{100}$, its node count $n_0$, which is larger than $q^{4r_0}$, is also larger than $q^{q^4/25}$; thus, finding $H$ can be achieved in time $\text{poly}(n_0)$.

  Now, from $E_0$ we construct an expander $G$ of type $[n_0, 12\varrho, \delta_G]$ by replacing each of $E_0$'s edges with $\varrho$-fold multi-edges (which, of course affects the expansion parameter). Because $E_0$ was 12-edge-colorable, $G$ is $12\varrho$-edge-colorable.

  We construct the final expander $E$ as a replacement product: $E := G \circ H$, which is of type $[12\varrho \cdot n_0, 12, \delta]$.

  Defining $m := 12\varrho \cdot n_0$, we now show that $m = \Theta(n)$; if so, then $E$ satisfies our requirements.

  In fact, according to the definition of $\varrho$, we obtain

  $$m = 12\varrho \cdot n_0 = \left\lfloor \frac{3}{4} n \right\rfloor,$$

  which readily yields

  $$\frac{n}{2} \leq m \leq n,$$

  so $m$ is indeed in $\Theta(n)$.

### 3.7.2 Expanders with Smaller Degree

The expanders described in theorem 3.27 (p. 62) are 12-regular and 12-edge-colorable.

In order to obtain an expander with smaller degree, the authors of [ASS08] recommend calculating a replacement product $E \circ G$ with an expander $G$ of type $[12, d, \delta_G]$, which is well-defined. Since $E$ is loop-free, the original replacement product from definition 3.6 (p. 37) is sufficient here.

The result will be a $2d$-regular graph, according to corollary 3.7 (p. 37), and will be $2d$-edge-colorable as per corollary 3.8 (p. 37) if $G$ is $d$-edge-colorable.

If an odd degree $(2d-1)$ is desired, we can use the $d$-regular graph $G$, but only add $(d-1)$ instead of $d$ multi-edges in step 2 of the replacement product. For $d \geq 2$, this will still yield a connected graph because at least one edge is added for each node pair in step 2; thus, the connections between the various copies $G_j$ (replacing the nodes of $E$) will be established.

# Chapter 4

# Conclusion

We have introduced the necessary requirements to understand the nature of regular expander graph. In our main chapter, we have presented and motivated most of the work in [ASS08], which gives a concise way to construct expander graphs of various node counts but with fixed degree.

Several concepts that the authors declare "easy to see" have required quite a few intermediary steps to make them accessible to the interested reader not versed in advanced combinatorics and (spectral) graph theory.

We do, however, appreciate that Alon et al. have offered proofs for all their assertions (perhaps with the exception of theorem 2 [ASS08], for which we have provided a motivation of our own), and that in a matter of this complexity, not everything can be expressed in the confines of an eight-page article.

We have adapted the special expander graphs $LD(q, r)$ and the replacement product presented in [ASS08] to conform to the notions of loops, node degree and edge-colorability as introduced in our preparatory chapter 2, with a limited exception to corollary 2.17 (p. 14) by way of bi-colored loops. The spectral and expansion properties are not affected by those changes.

At some points, we have changed notation in the hope for greater clarity and consistency with chapter 2, and we have provided several illustrations in order to demonstrate or motivate certain concepts, like e.g. the replacement product.

An extensive appendix provides mathematical prerequisites that are (or, at least, could be) taught in undergraduate maths education for the sciences, but may not feature in every interested reader's later work. While we tried to be thorough, we had to leave some gaps (for instance, a true understanding of Galois fields would require substantial preparations in algebra). The appendix has also served as a central mathematical reference for most of the calculations in the previous chapters.

# Appendix A

# Mathematical Basics

## A.1 About the Mathematical Background

We begin this appendix with a chapter on basic algebraic structures which are helpful for a consistent formulation of the actual mathematical concepts we want to explain. Some initial rigidity will therefore unavoidable, yet we will try to avoid formal detail where it does not greatly improve clarity.

For example, while we want to mention the special linear group $SL$, we will omit the notion of normal subgroups. And although we need the concept of a finite field for some aspects in the above thesis, we will not elaborate on Galois theory, the uniqueness of Galois fields or the notions of algebraic closure. Also, we will concentrate on real vector spaces instead of complex ones where possible. Some notational hand-waviness related to residue classes and their representatives will be explained and motivated in the relevant section on residue rings.

After the algebraic vocabulary is established, we will summarize some important results from linear algebra.

The aim of this appendix is to make the thesis accessible to anyone with undergraduate-level mathematical education. Some space (not only in this chapter) is devoted to the treatment of matrices and linear algebra, which may be redundant for those familiar with higher mathematics.

All the statements collected here can be verified using textbooks of (linear) algebra. If a statement is denoted "claim", this suggests a more involved proof. Proofs of "theorems" are only adapted from literature, and we refer to the original. Some of the "smaller" statements (lemmas) are also adapted from literature proofs.

This chapter uses definitions from [Hof14, KM21, K$^+$88] and attempts to achieve a good compromise on brevity and completeness.

## A.2 Algebraic Structures, with Simple Examples

*Sets* of objects are among the most fundamental structures in mathematics because they allow us to group similar objects together. The behavior of such elements can be described with *maps*, i.e. instructions that define which object(s) an object (or several objects) is (are) connected to.

An *algebraic structure* combines a set of objects with one or more maps that describe either *internal operations* (connecting elements within the set) or *external operations* (connecting elements from the set and elements from other sets).

A *magma* is a simple example of an algebraic structure, combining a set of objects with an internal connection that is closed – any two elements of the set map to some element of the same set. In the following, we develop a cascade of restrictions to obtain algebraic structures with higher specification that will allow for more detailed reasoning.

### A.2.1 Groups

**Definition A.1** *The pair $(G, \circ)$, with a set $G$ of objects and a binary operation $\circ : G \times G \to G$ is called a* semigroup *if it satisfies:*

*G1:* $\forall a, b \in G : a \circ b \in G$ *(closure of $G$ under $\circ$)*

*G2:* $\forall a, b, c \in G : a \circ (b \circ c) = (a \circ b) \circ c$ *(associativity of $\circ$)*

Example: the positive integers $\mathbb{N}$ under the addition $+$.

**Definition A.2** *A semigroup $(G, \circ)$ is called a* monoid *if $G$ contains an identity element $e$, such that*

*G3:* $\forall a \in G : e \circ a = a \circ e = a$

Examples:

- the non-negative integers $\mathbb{N}_0$ under the addition $+$, with the identity element 0
- the integers $\mathbb{N}$ (or $\mathbb{N}_0$) under multiplication, with the identity element 1
- square matrices from $\mathbb{R}^{n \times n}$ under matrix multiplication (cf. the following section on linear algebra for details on matrices), with the identity element $\mathbb{1}_n = \mathrm{diag}(1, \ldots, 1)$ (the unit matrix with $n$ entries of 1 on its diagonal)

Note that the identity is unique, because if there were another element $e'$ satisfying G3, then

$$e = e \circ e' = e',$$

where we use G3 for $e'$ in the first equality, and G3 for $e$ in the second.

**Definition A.3** *A monoid $(G, \circ)$ is called a* group*, if it satisfies the additional constraint*

*G4:* $\forall a \in G : \exists a^{-1} \in G : a \circ a^{-1} = a^{-1} \circ a = e$

*The group is called* commutative *(or* Abelian*) if all elements $a, b$ of $G$ satisfy*

$$a \circ b = b \circ a$$

Examples:

- the integers $\mathbb{Z}$ under the addition. The inverse element of $a$ is $-a$.
- the rational, real and complex numbers ($\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$) under addition. The inverse element of $a$ is $-a$.
- the rational, real and complex numbers ($\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$), each without 0, under multiplication. The inverse element of $a$ is $(1/a)$.
- invertible square matrices from $\mathbb{R}^{n \times n}$ under matrix multiplication. Unlike the prior examples, this group is not Abelian, because matrix multiplication is not generally commutative. This group is also called the *general linear group* $GL(\mathbb{R}, n)$ and will be briefly mentioned again in the first algebra chapter, after the notion of inverse matrices has been introduced.

Note that for any element $a$ of a group, its inverse $a^{-1}$ is unique, because if there were another element $b$ satisfying G4, then

$$b = b \circ e = b \circ (a \circ a^{-1}) = (b \circ a) \circ a^{-1} = e \circ a^{-1} = a^{-1},$$

where we use the following sequences of the group axioms for the equalities: G3, G4 for $a^{-1}$, G2, G4 for $b$, G3.

**Definition A.4** *A group $(\tilde{G}, \tilde{\circ})$ is called a* subgroup *of $(G, \circ)$, if $\tilde{G} \subset G$ and if $\tilde{\circ}$ is the restriction of $\circ$ to $\tilde{G}$.*

Examples: Because $\mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$, all the subset relations imply subgroup relations under addition. For all of the above sets but $\mathbb{Z}$, taken without $\{0\}$, there are also subgroup relations under multiplication.

Note that the identity $e$ of $G$ is also the identity of $\tilde{G}$, because any $a$ in $\tilde{G}$ is also in $G$, and therefore can be (according to G3 for $G$) combined with $e$ from $G$ to yield $a$ itself – which satisfies G3 for $\tilde{G}$ as well. Since the identity is unique in $\tilde{G}$, too, it cannot be another element than $e$ from $G$, or $G$ would contain two different identities.

### A.2.2 Rings and Fields

While groups only consider a single operation linking elements of a set, there are plenty of cases where two operations are defined; often these are addition "+" and multiplication "$\cdot$". Rings and fields combine a set with two such operations, and we will use the above symbols for the operations.

**Definition A.5** *A triple $(R, +, \cdot)$, where $R$ is a set of objects, and the operations $+$ and $\cdot$ connect via $+, \cdot : R \times R \to R$, is called a* ring *if it satisfies the following conditions:*

*R1:* $(R, +)$ *is an Abelian group, with identity element 0*

*R2:* $(R, \cdot)$ *is a semigroup*

*R3:* $\cdot$ *distributes with respect to* $+$*, i.e.* $\forall a, b, c \in R$ :

$$a \cdot (b + c) = (a \cdot b) + (a \cdot c)$$
$$(a + b) \cdot c = (a \cdot c) + (b \cdot c)$$

*If $(R, \cdot)$ is a monoid, i.e. if $R$ contains the identity of the multiplication, then $(R, +, \cdot)$ is called a* ring with unity.
*If the multiplication is commutative, $(R, +, \cdot)$ is called a* commutative ring.

Examples:

- $(\mathbb{Z}, +, \cdot)$ is a commutative ring with unity, the identities are 0 (for $+$) and 1 (for $\cdot$). The same holds for $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$, respectively.
- $(\mathbb{R}^{n \times n}, +, \cdot)$ with the matrix multiplication for "$\cdot$" is a ring with unity, but not a commutative ring.

Note that for commutative rings, the two distributive laws in R3 amount to the same.

If we constrain the notion of a commutative ring with unity even further, we obtain the definition of a field:

**Definition A.6** *A triple $(F, +, \cdot)$ is called a* field *if it is a commutative ring with unity, and every element $a \in F \setminus \{0\}$ has a non-zero multiplicative inverse in $F$. We denote the identity element of multiplication with 1 ($\neq 0$).*

Examples: The number sets $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ with the usual addition and multiplication (the integers $\mathbb{Z}$ do not form a field because they do not contain multiplicative inverses except for $\pm 1$).

In the following, we will use those labels for both the actual number sets as well as for the associated fields with the usual arithmetic operations. We will use the same approach for other fields or rings when the relevant operations can be inferred easily from context.

We will elaborate on finite fields in the final chapter of this appendix.

Note that, for a field $F$, $(F \setminus \{0\}, \cdot)$ is, per the above definition, an Abelian group. Together with the Abelian group $(F, +)$, this means that fields allow for all the four basic arithmetic operations. Subtraction is defined by adding the additive inverse of an element, and division by multiplying its multiplicative inverse, respectively.

### A.2.3 Vector Spaces

Fields allow for basic arithmetic between what we can represent as numbers. The following introduces vector spaces, which combine such numbers with new entities – called vectors – that may carry additional structure.

In anticipation of the Euclidean space we write most vectors as Latin letters with an arrow, as is common in Physics or Engineering (many mathematical textbooks instead use boldface Latin letters or the Fraktur script).

**Definition A.7** *Given a field $\mathbb{F}$, a triple $(V, +, \cdot)$, where $V$ is a set of objects, and the operations $+$ and $\cdot$ connect via $+ : V \times V \to V$ and $\cdot : \mathbb{F} \times V \to V$, is called a* vector space over $\mathbb{F}$ *if it satisfies the following conditions:*

*V1:* $(V, +)$ *is an Abelian group, with $\vec{0}$ as the identity element (the* zero vector*)*

*V2:* $\forall a, b \in \mathbb{F}, \vec{v}, \vec{w} \in V$*:*

$$\begin{aligned} (a+b)\cdot\vec{v} &= (a\cdot\vec{v})+(b\cdot\vec{v}) \\ a\cdot(\vec{v}+\vec{w}) &= (a\cdot\vec{v})+(a\cdot\vec{w}) \\ (ab)\cdot\vec{v} &= a\cdot(b\cdot\vec{v}) \\ 1\cdot\vec{v} &= \vec{v} \end{aligned}$$

*The operation* $+$ *is called* vector addition*, while* $\cdot$ *is called* scalar multiplication*. The elements of* $\mathbb{F}$ *are called* scalars*.*

Examples: Beside the Euclidean space which we will introduce in the next subsection, there are many spaces where the vectors are functions, like the space of continuous functions, or the Schwartz spaces, which are studied e.g. in functional analysis or harmonic analysis.

Note that V2 explains why the elements of the field $\mathbb{F}$ are called scalars – they can be used to scale the vectors in $V$. Plugging $1 \in \mathbb{F}$ in for both $a, b$ in the first equation (and assuming that the element representing $1+1$ in $\mathbb{F}$ is 2), we obtain $2\cdot\vec{v} = (1\cdot v)+(1\cdot v)$, which yields $\vec{v}+\vec{v}$ according to the fourth equation. So, if we scale the vector with a number 2, the result is the vector added to itself.

Plugging 1 and 0 from $\mathbb{F}$ in for $a$ and $b$ in the first equation, we obtain (again using the fourth equation): $\vec{v} = \vec{v} + (0\cdot\vec{v})$. This equation can only hold if $0\cdot\vec{v}$ equals the zero vector mentioned in V1.

We should point out that the $+$ in the left-hand side of the first equation in V2 is the addition operation of the field $\mathbb{F}$, not the one specified in the triple $(V, +, \cdot)$. Also, the product inside the brackets on the l.h. side of the third equation uses the field's multiplication operation.

Since V2 is postulated in order to achieve consistency between the two multiplication operations, we will label them with the same symbol from here on (the specific operation can be inferred from context if desired), or dispense with the $\cdot$ altogether in products, e.g., write $a\vec{v}$ instead of $a\cdot\vec{v}$.

In addition, note that any field $\mathbb{F}$ is also a vector space over itself. The operations $+$ and $\cdot$, in this case are the same for the field and for the vector space, and the zero vector equals $0 \in \mathbb{F}$. The conditions in V2 bring no new information here, because they translate into the distributive laws of any ring, the associativity of multiplication and the existence of a multiplicative identity in a ring with unity, which $\mathbb{F}$ satisfies by itself.

However, we will presently encounter vector spaces that are (in a simple way) constructed from fields while not being fields themselves. The main reason for this is that while fields demand a commutative multiplication operation (with full group structure) amongst their member elements, vector spaces do not. The product $\cdot$ of the vector space is only used for scaling vectors.

### A.2.4 Euclidean Space

As mentioned above, vector spaces per se do not stipulate any product between vectors. Many vector spaces do, however, allow for products of various kinds (e.g. tensor product, cross product). There is even a whole class of vector spaces called "inner product spaces" (a.k.a. pre-Hilbert spaces) that feature, not surprisingly, an *inner product*, which maps a pair of vectors to a number from the field that the space is defined over.

For our purposes, we restrict ourselves to real vector spaces and make use of the following concept:

**Definition A.8** *A vector space* $V$ *over the field* $\mathbb{R}$ *is called* Euclidean space *if it is combined with a* scalar product*, a map* $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{R}$ *with the properties* $(\forall \vec{u}, \vec{v}, \vec{w} \in V; a, b \in \mathbb{R})$*:*

*S1:* $\langle\vec{v}, \vec{w}\rangle = \langle\vec{w}, \vec{v}\rangle$ *(Symmetry)*

*S2:* $\langle a\vec{u} + b\vec{v}, \vec{w}\rangle = a\langle\vec{u}, \vec{w}\rangle + b\langle\vec{v}, \vec{w}\rangle$ *(Linearity in the first argument)*

*S3:* $\vec{v} \neq \vec{0} \Rightarrow \langle\vec{v}, \vec{v}\rangle > 0$ *(Positive-Definiteness)*

See the following section for examples.

Note that S1 and S2 can be combined to show that the scalar product is also linear in the second argument: $\langle\vec{u}, a\vec{v} + b\vec{w}\rangle = \langle a\vec{v} + b\vec{w}, \vec{u}\rangle = a\langle\vec{v}, \vec{u}\rangle + b\langle\vec{w}, \vec{u}\rangle = a\langle\vec{u}, \vec{v}\rangle + b\langle\vec{u}, \vec{w}\rangle$. The scalar

product is in fact bilinear.

We should point out that the arrow notation suggests vectors in the sense of physical coordinates; this was again done in anticipation of the next section on linear algebra. However, it is also possible to define a scalar product for other spaces, e.g. between the wave functions in quantum mechanics, or the sine/cosine functions in harmonic analysis.

## A.3 Linear Algebra Recap

In this section we focus on vectors in $n$-dimensional Euclidean space. We will also introduce matrices and examine their interplay with vectors, which is important groundwork for the later chapter on the Eigenvalue problem.

### A.3.1 The Euclidean space $\mathbb{R}^n$

$\mathbb{R}^n$ ($n \in \mathbb{N}$) is the $n$-fold Cartesian product of real numbers; its elements are $n$-tuples of reals.

In order to build this set into a vector space, the first thing we require is the addition operation. For reasons that will become apparent when we introduce matrices, we denote the tuples not horizontally but vertically.

**Definition A.9** *For $\vec{v}, \vec{w} \in \mathbb{R}^n$, with components $v_1, \cdots, v_n$ and $w_1, \cdots, w_n$, respectively, the vector $\vec{v} + \vec{w}$ is defined per*

$$\forall j \in \{1, \cdots, n\} : (\vec{v} + \vec{w})_j := v_j + w_j,$$

*where the $+$ operation on the right-hand side is just the standard addition of real numbers.*

*The zero vector $\vec{0}$ of $\mathbb{R}^n$ is the vector comprising $n$ components of value $0$.*

(When the number of components can be easily inferred from context and where it does not cause confusion, we will henceforth omit the interval specification and only write $\forall j$.)

Written out a bit more completely, this reads:

$$\vec{v} + \vec{w} = \begin{pmatrix} v_1 \\ v_2 \\ \cdots \\ v_n \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \\ \cdots \\ w_n \end{pmatrix} := \begin{pmatrix} v_1 + w_1 \\ v_2 + w_2 \\ \cdots \\ v_n + w_n \end{pmatrix}$$

With the component-wise addition, the vectors in $\mathbb{R}^n$ constitute an Abelian group, as demanded in the axiom V1 of the vector space definition, because the various single components form an Abelian group in the field $\mathbb{R}$.

In the same fashion, we define the scalar multiplication component-wise:

**Definition A.10** *For $\vec{v} \in \mathbb{R}^n, a \in \mathbb{R}$, the scaled vector $a\vec{v}$ is defined per*

$$\forall j : (a\vec{v})_j := a \cdot v_j$$

Or, alternatively:

$$a\vec{v} = \begin{pmatrix} (a\vec{v})_1 \\ (a\vec{v})_2 \\ \cdots \\ (a\vec{v})_n \end{pmatrix} := \begin{pmatrix} a \cdot v_1 \\ a \cdot v_2 \\ \cdots \\ a \cdot v_n \end{pmatrix}$$

With this component-wise scaling and the addition operation from above, all the requirements stated in axiom V2 of the vector space definition are easily met, because the equations can be evaluated component-wise.

Like we understand $\mathbb{R}$ as either the set or the field of real numbers, respectively, depending on the context, we will in the following understand $\mathbb{R}^n$ as either the $n$-fold Cartesian product of real numbers or the vector space (over $\mathbb{R}$) constructed with the component-wise addition and scalar multiplication.

In order to justify this subsection's heading, we still need to define a scalar product (not to be confused with the scalar multiplication!):

**Definition A.11** *For $\vec{v}, \vec{w} \in \mathbb{R}^n$, the* canonical scalar product *is defined per*

$$\langle \vec{v}, \vec{w} \rangle := \sum_{j=1}^{n} v_j \cdot w_j$$

(We will omit the summation range in the future, whenever the intended range can be easily inferred.)

Since the multiplication of reals is commutative, this product is symmetric, satisfying S1 of the above axioms.

We demonstrate the linearity demanded in S2:

$$\langle (a\vec{u} + b\vec{v}), \vec{w} \rangle = \sum_j (a\vec{u} + b\vec{v})_j \cdot w_j = \sum_j (au_j + bv_j)w_j = \sum_j [(au_j w_j) + (bv_j w_j)]$$

We split the sum into two (which is allowed because addition of reals is associative) and pull the factors $a$ and $b$ out of the sums:

$$\cdots = \sum_j (au_j w_j) + \sum_j (bv_j w_j) = a \sum_j (u_j w_j) + b \sum_j (v_j w_j) = a\langle \vec{u}, \vec{w} \rangle + b\langle \vec{v}, \vec{w} \rangle$$

For the third axiom S3 of scalar products, we note that $\langle \vec{v}, \vec{v} \rangle$ evaluates to the sum of the squares of the vector's components. For real numbers, such a square is never negative, and the whole sum can only be zero if all the squares (and therefore the components) are zero.

Thus, $\mathbb{R}^n$ with the canonical scalar product is indeed a Euclidean space.

Note that this Euclidean space is not a field, because there is no notion of vector multiplication that involves inverses, or of a "unit vector" that could serve as identity of multiplication. While we can apply the scalar product between a vector $\vec{v}$ and the vector consisting of ones, this would only yield the sum of $\vec{v}$'s components, a real number; the structure of the Cartesian product space is lost in the process. This serves as a belated example to our remark on vector spaces in the previous section.

### A.3.2 Cartesian Coordinate Systems

We can, however, imagine the $\mathbb{R}^n$ vectors as multi-dimensional arrows in a *Cartesian Coordinate System*. This is a representation of $\mathbb{R}^n$ with $n$ mutually perpendicular straight axes and a fixed origin where all axes intersect. Any point in such a coordinate system can be understood as synonymous with the arrow leading from the origin to that particular point.

Adding two vectors amounts to drawing the second arrow from the tip of the first one, or vice versa (this is commutative) in order to get a resulting arrow corresponding to the vector sum.

The scalar product also induces a norm, (in this case, the *Euclidean Norm*: the square root of the scalar product of a vector with itself), by which we can measure the length of such an arrow: The Euclidean norm is the generalization of the Pythagorean Theorem in $n$ dimensions.

Scalar multiplication neatly amounts to scaling the length of an arrow, because the squares and square root cancel each other out for the scaling factor. Scaling with a negative number flips the arrow's direction.

In addition to providing the notion of arrow length, the scalar product also allows us to define an angle between two arrows/vectors. In order to ensure consistency with two-dimensional geometry,

**Definition A.12** *The* angle *between two vectors $\vec{v}, \vec{w} \in \mathbb{R}^n \setminus \{\vec{0}\}$ is defined by*

$$\cos \angle(\vec{v}, \vec{w}) := \frac{\langle \vec{v}, \vec{w} \rangle}{\sqrt{\langle \vec{v}, \vec{v} \rangle}\sqrt{\langle \vec{w}, \vec{w} \rangle}}.$$

*$\vec{v}, \vec{w}$ are called* perpendicular *if $\langle \vec{v}, \vec{w} \rangle = 0$.*

In the Cartesian coordinate system, we may label the axes with numbers 1 to $n$. If we draw a vector from the origin, and project its tip on the various axes, we obtain the components of the vector in the *standard basis* $E$. We will elaborate on this shortly.

### A.3.3 Linear Combinations

**Definition A.13** *Given $m$ vectors $\vec{v}_1, \cdots, \vec{v}_m \in \mathbb{R}^n$, together with $m$ corresponding scaling factors $a_1, \cdots, a_m \in \mathbb{R}$, the sum*

$$\sum_{j=1}^{m} a_j \vec{v}_j$$

*is called a* linear combination *of those vectors.*

This sum is itself a vector from $\mathbb{R}^n$. (Note that $\vec{v}_j$ means the $j$-th vector, not the $j$-th component of some vector $\vec{v}$ (which we would denote $v_j$).)

If we take not only one set of scaling factors, but instead allow all possible combinations, we obtain a whole set of linear combinations:

**Definition A.14** *Given $m$ vectors $\vec{v}_1, \cdots, \vec{v}_m \in \mathbb{R}^n$, the set*

$$\left\{ \sum_{j=1}^{m} a_j \vec{v}_j \,\middle|\, a_1, \cdots, a_m \in \mathbb{R} \right\}$$

*is called the* span *of those vectors, denoted by* $\text{span}(\vec{v}_1, \cdots, \vec{v}_m)$.

Examples:

- The span of a single vector defines a line through the origin of the Cartesian coordinate system.
- The span of two vectors defines a plane containing the origin, if the two vectors do not point in the same or opposite direction (i.e. they cannot be transformed into each other by scaling). We will formalize this caveat in the shortly.
- The span of the zero vector is just the origin of the coordinate system.

### A.3.4 The Euclidean Unit Vectors

We now return to the standard basis of Euclidean space. This is a set of vectors that each contain $(n-1)$ zeros and one 1 component.

**Definition A.15** *For $j \in \{1, \cdots, n\}$, the* unit vector $\vec{e}_j$ *of Euclidean space $\mathbb{R}^n$ has the following components:*

$$(\vec{e}_j)_k := \delta_{jk} := \begin{cases} 1 & j = k \\ 0 & \textit{otherwise} \end{cases}$$

*$\delta_{jk}$ is called the* Kronecker symbol.

(In the following, the Kronecker symbol will be used frequently; it is particularly useful when evaluating sums.)

We calculate the scalar product of two Euclidean unit vectors:

$$\langle \vec{e}_j, \vec{e}_k \rangle = \sum_p (\vec{e}_j)_p (\vec{e}_k)_p = \sum_p \delta_{jp} \delta_{kp}$$

Since the first Kronecker symbol can only be 1 (and non-zero) for $p = j$, all other summands vanish, and the remaining term satisfies $p = j$. Thus:

$$\langle \vec{e}_j, \vec{e}_k \rangle = \delta_{kj} = \delta_{jk}$$

With the length and angle definitions from above, we see that any unit vector has unit length (by calculating the scalar product of the vector with itself, $\delta_{jj} = 1$) and that different unit vectors are perpendicular to each other.

We now project a vector $\vec{v}$ onto one of the unit vectors by calculating the scalar product:

$$\langle \vec{e}_j, \vec{v} \rangle = \sum_k (\vec{e}_j)_k v_k = \sum_k \delta_{jk} v_k = v_j$$

Now we can write $\vec{v}$ as a linear combination of the Euclidean unit vectors, where the scaling factors are just the $v_j$ obtained by projection:

$$\vec{v} = v_1\vec{e}_1 + \cdots + v_n\vec{e}_n = \sum_j v_j\vec{e}_j$$

And indeed, taking the $k$-th component of this yields

$$v_k = (\vec{v})_k = \sum_j v_j(\vec{e}_j)_k = \sum_j v_j\delta_{jk} = v_k \quad \checkmark$$

**Definition A.16** *The linear span of a set of vectors $\vec{v}_1, \cdots, \vec{v}_k$ is called a* generating system *of $V$ if any vector $\vec{v} \in V$ satisfies $\vec{v} \in \text{span}(\vec{v}_1, \cdots, \vec{v}_k)$.*

As we have demonstrated, $E := \{\vec{e}_1, \cdots, \vec{e}_n\}$ is a generating system of $\mathbb{R}^n$.

### A.3.5 Linear (In)Dependence and Bases

**Definition A.17** *A set of vectors $\{\vec{v}_1, \cdots, \vec{v}_k\}$ is called* linearly independent *if the only linear combination that yields the zero vector $\vec{0}$ is the trivial one, where each scaling factor is 0.*

*If there is a non-trivial linear combination of $\vec{0}$, the vectors are called* linearly dependent*.*

Examples:

- The zero vector is linearly dependent because $a\vec{0} = \vec{0}$ for non-zero $a$.
- Two vectors that are connected by scaling are linearly dependent. If $\vec{v} = a\vec{w}$ with non-zero $a$, the linear combination $\vec{v} + (-a)\vec{w}$ yields $\vec{0}$ non-trivially.

**Lemma A.18** *The Euclidean unit vectors $E$ are linearly independent.*

Proof: We have established that $E$ is a generating system of $\mathbb{R}^n$. Thus, we can write a linear combination of the vectors in $E$ that evaluates to $\vec{0} \in \mathbb{R}^n$. With scaling factors $a_.$, we obtain:

$$\vec{0} = \sum_j a_j\vec{e}_j$$

Taking the $k$-th component of that, yields:

$$0 = (\vec{0})_k = \sum_j a_j(\vec{e}_j)_k = \sum_j a_j\delta_{jk} = a_k$$

Evidently, we can only combine the vectors in $E$ to $\vec{0}$ if we scale each $\vec{e}_j$ with a factor of 0 – the trivial linear combination of $\vec{0}$. ■

We now can define the basis of a vector space:

**Definition A.19** *A set of vectors $B \subset V$ of a vector space $V$ is called a* basis *of $V$ if it is linearly independent and if it is a generating system of $V$. If the order $|B|$ of the basis is finite with value $n$, $V$ is called $n$-*dimensional*, otherwise $V$ is called* infinite-dimensional*.*

Since the vectors in $E$ are satisfy both those conditions, we can call $E$ a base of $\mathbb{R}^n$, an $n$-dimensional Euclidean space – in fact, $E$ is called the *standard basis* of $\mathbb{R}^n$.

It can be shown (cf. [FS20], ch. 2.5) that for a vector space with a finite basis of $n$ vectors, all other possible bases contain $n$ vectors as well (that is why the above definition of vector space dimension is meaningful), and that those bases contain the maximum amount of linearly independent vectors, i.e. $(n+1)$ vectors of an $n$-dimensional space must be linearly dependent. Also, if a vector space has a finite basis of $n$ vectors, any set of $n$ linearly independent vectors is also a basis of that space.

### A.3.6 Linear Maps in Euclidean Space, Matrices

A linear function $f$ in $\mathbb{R}$ satisfies two conditions: that scaling factors can be pulled outside the function application ($f(ax) = af(x)$), and that $f$ applied to a sum of arguments equals the sum of $f$ applied to the individual arguments, respectively.

We now extend this concept to vector spaces:

**Definition A.20** *A map $\vec{f} : \mathbb{R}^n \to \mathbb{R}^m$ is called a* linear map *if for any $a, b \in \mathbb{R}; \vec{v}, \vec{w} \in \mathbb{R}^n$:*

$$\vec{f}(a\vec{v} + b\vec{w}) = a\vec{f}(\vec{v}) + b\vec{f}(\vec{w})$$

We proceed by evaluating the components of all vectors. For that, we use the standard bases $E_m$, $E_n$ of the respective spaces. For any $\vec{v} \in \mathbb{R}^n$ we obtain:

$$\vec{v} = \sum_{k=1}^{n} v_k \vec{e}_k \qquad\qquad \vec{f}(\vec{v}) = \sum_{j=1}^{m} f_j(\vec{v})\vec{e}_j$$

And if we demand linear map behavior from all the components $f_j$, we can plug in the first equation into the $j$-th component of $\vec{f}$:

$$f_j(\vec{v}) = f_j\left(\sum_{k=1}^{n} v_k \vec{e}_k\right) = \sum_{k=1}^{n} v_k f_j(\vec{e}_k) = \sum_{k=1}^{n} \left(f_j(\vec{e}_k)\right) v_k =: \sum_{k=1}^{n} M_{jk} v_k$$

Note that the application of the function can be constrained to the (constant) vectors of the standard base, while the components of $\vec{v}$ that make up the concrete shape of the vector are multiplied, due to the linear behavior of $\vec{f}$ in all its components, which we utilized in the third equality.

Thus, a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ may be calculated by evaluating the map's behavior regarding the Cartesian unit vectors of $\mathbb{R}^n$; this yields the coefficients $M_{jk}$, where $j$ runs from 1 to $m$ and $k$ from 1 to $n$, respectively. Each of the $M_{jk} = f_j(\vec{e}_k)$ is a real number. It is sufficient to calculate those $m \cdot n$ numbers once; after that, they can be applied to any vector from $\mathbb{R}^n$ by evaluating the sum in the last equality of the above equation.

Also note that, for a fixed $j$, the sum over $M_{jk}v_k$ has the structure of the canonical scalar product in the Euclidean space $\mathbb{R}^n$.

We can organize the map coefficients into a rectangular array $M$, which we call a *Matrix*, and employ a suggestive product notation that we will shortly define properly:

$$\vec{f}(\vec{v}) = \begin{pmatrix} f_1(\vec{v}) \\ f_2(\vec{v}) \\ \cdots \\ f_m(\vec{v}) \end{pmatrix} = \begin{pmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,n} \\ M_{2,1} & M_{2,2} & \cdots & M_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ M_{m,1} & M_{m,2} & \cdots & M_{m,n} \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ \\ \cdots \\ \\ \\ v_n \end{pmatrix} =: M \cdot \vec{v}$$

Note that the rectangular array has as many rows as the resulting vector $\vec{f}(\vec{v})$ and as many columns as the vector $\vec{v}$ that it is multiplied to.

**Definition A.21** *A rectangular array of $m$ rows and $n$ columns, containing numbers from $\mathbb{R}$, is called a* real matrix *$M \in \mathbb{R}^{m \times n}$. Its components are denoted $M_{jk}$, where $j \in \{1, \cdots, m\}$ specifies the row index and $k \in \{1, \cdots, n\}$ the column index, respectively. Such a matrix can be multiplied to a vector $\vec{v} \in \mathbb{R}^n$ using the* matrix product*, to obtain a vector $\vec{w} \in \mathbb{R}^m$, via*

$$\vec{w} = M \cdot \vec{v} \qquad :\Leftrightarrow \qquad \forall j \in \{1, \cdots, m\} : w_j := \sum_{k=1}^{n} M_{jk} v_k$$

*Matrices in $\mathbb{R}^{m \times n}$ are exactly the representations of all possible linear maps $\mathbb{R}^n \to \mathbb{R}^m$.*

We will explore the full range of the matrix product in the next subsection, and conclude this subsection with observations on the structure of the matrix space.

**Lemma A.22** *The algebraic structure $\mathbb{R}^{m \times n}$ is a vector space over the field $\mathbb{R}$, with the component-wise vector addition and scalar multiplication exactly as defined for vectors in Euclidean spaces.*

Proof: We exploit the behavior of linear maps and of the direct correspondence between a linear map and its associated matrix.

Consider two maps $\vec{f}, \vec{g} : \mathbb{R}^n \to \mathbb{R}^m$. Since the resulting vectors are both in $\mathbb{R}^m$ they may readily be added using the familiar vector addition. Let the matrix $P \in \mathbb{R}^{m\times n}$ represent the resulting map:

$$\sum_{k=1}^{n} P_{jk} v_k = \left(\vec{f} + \vec{g}\right)_j (\vec{v}) = f_j(\vec{v}) + g_j(\vec{v})$$

Now, if $\vec{f}$ is represented by the matrix $M$, and $\vec{g}$ by $N$, respectively, we can plug in the definitions for $f_j$ and $g_j$, combine the sums and factor out the vector components of $\vec{v} \in \mathbb{R}^n$:

$$f_j(\vec{v}) + g_j(\vec{v}) = \sum_{k=1}^{n} M_{jk} v_k + \sum_{k=1}^{n} N_{jk} v_k = \sum_{k=1}^{n} (M_{jk} + N_{jk}) v_k$$

Thus, we obtain:

$$P = M + N \qquad \text{per} \qquad P_{jk} = M_{jk} + N_{jk}$$

Similarly, consider another map $\vec{h} := a\vec{f}$, where $a \in \mathbb{R}$ scales the map $\vec{f}$. Because $\mathbb{R}^m$ is a vector space, $\vec{h}$ is a bona fide vector of $\mathbb{R}^m$, and we may represent the map by a matrix $Q \in \mathbb{R}^{m\times n}$.

$$\sum_{k=1}^{n} Q_{jk} v_k = \vec{h}_j(\vec{v}) = \left(a\vec{f}(\vec{v})\right)_j = a f_j(\vec{v}) = a \sum_{k=1}^{n} M_{jk} v_k = \sum_{k=1}^{n} (aM_{jk})\, v_k$$

And thus:

$$Q = aM \qquad \text{per} \qquad Q_{jk} = aM_{jk}$$

Now, with the matrix addition explained component-wise as shown, the matrices in $\mathbb{R}^{m\times n}$ constitute an Abelian group, with the zero matrix as its identity (Condition V1). The component-wise scalar multiplication satisfies the equations in condition V2 of the vector space definition. Therefore, together with those operations, the set $\mathbb{R}^{m\times n}$ does indeed constitute a vector space (cf. definition A.7, p. 68). ■

### A.3.7 Matrix Multiplication

In order to formulate the full matrix multiplication, we revisit the visualization of the linear map $\vec{f}(\vec{v})$ with column vectors $\vec{f} \in \mathbb{R}^m$ and $\vec{v} \in \mathbb{R}^n$ and a matrix $M \in \mathbb{R}^{m\times n}$.

Apparently there is no difference in notation between the column vector $\vec{v}$ and a matrix from $\mathbb{R}^{n\times 1}$.

We now expand the column vector $\vec{v}$ into a matrix $V \in \mathbb{R}^{n\times p}$ with $p$ columns, which we write as

$$V := [\vec{v}_1, \cdots, \vec{v}_p] \qquad \text{per} \qquad V_{ks} = (\vec{v}_s)_k,$$

where the indexes are vector labels and do not indicate vector components. We write square brackets instead of round ones because we want to reserve the latter for row vectors (which will be introduced in the next subsection). The square brackets here do not indicate an interval.

For the multiplication, we can imagine that the linear map represented by $M$ treats each of those columns $\vec{v}_s$ ($s \in \{1, \cdots, p\}$) separately, yielding one vector $\vec{f}(\vec{v}_s)$. Consequently, the results now form a matrix, too, with $m$ lines and $p$ columns:

$$F(V) := [\vec{f}(\vec{v}_1), \cdots, \vec{f}(\vec{v}_p)] \qquad \text{per} \qquad F_{js} = (\vec{v}_s)_j,$$

where we write $F$ to indicate the matrix structure of the result, similar to when we used $\vec{f}$ with an arrow to indicate the vector structure.

This means that we can explain a multiplication between matrices by applying the underlying linear map represented by the left-hand matrix $M$ column by column on the right-hand matrix $V$. Before the formal definition, we visualize the operation:

$$F(V) = \begin{pmatrix} (\vec{f}(\vec{v}_1))_1 & \cdots & (\vec{f}(\vec{v}_p))_1 \\ \vdots & \ddots & \vdots \\ (\vec{f}(\vec{v}_1))_m & \cdots & (\vec{f}(\vec{v}_p))_m \end{pmatrix} = \begin{pmatrix} M_{1,1} & \cdots & M_{1,n} \\ \vdots & \ddots & \vdots \\ M_{m,1} & \cdots & M_{m,n} \end{pmatrix} \cdot \begin{pmatrix} (\vec{v}_1)_1 & \cdots & (\vec{v}_p)_1 \\ \vdots & \ddots & \vdots \\ (\vec{v}_1)_n & \cdots & (\vec{v}_p)_n \end{pmatrix}$$

**Definition A.23** *Given a matrix $M \in \mathbb{R}^{m\times n}$ and a matrix $V \in \mathbb{R}^{n\times p}$, the* matrix product $F = M \cdot V$ *is a matrix $F \in \mathbb{R}^{m\times p}$ with components*

$$F_{js} := \sum_{k=1}^{n} M_{jk} V_{ks} \qquad (j \in \{1, \cdots, m\}, s \in \{1, \cdots, p\})$$

Note that the summation is over the inner indexes in the product terms, running over the columns of $M$ and the rows of $V$. This product is only defined if the left-hand matrix as exactly as many columns as the right-hand one has rows.

Also note that the matrix product is *not* commutative. If $m \neq p$, the product $V \cdot N$ is not even defined (because the column and row count relation does not match). If $m = p$, but $m \neq n$, $M \cdot V$ yields an $m \times m$ matrix whereas $V \cdot M$ yields an $n \times n$ matrix. Only the products of square matrices with the same size can potentially commute – but since the left-hand and right-hand matrices are treated differently, this would depend on the values of the matrix components.

The product is, however, associative. Let $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$ and $C \in \mathbb{R}^{p\times q}$, then:

$$\begin{aligned}
[A \cdot (B \cdot C)]_{js} &= \sum_{k=1}^{n} A_{jk} (B \cdot C)_{ks} = \sum_{k=1}^{n} A_{jk} \left( \sum_{r=1}^{p} B_{kr} C_{rs} \right) = \sum_{k=1}^{n} \sum_{r=1}^{p} A_{jk} B_{kr} C_{rs} \\
&= \sum_{r=1}^{p} \sum_{k=1}^{n} A_{jk} B_{kr} C_{rs} = \sum_{r=1}^{p} \left( \sum_{k=1}^{n} A_{jk} B_{kr} \right) C_{rs} = \sum_{r=1}^{p} (A \cdot B)_{jr} C_{rs} \\
&= [(A \cdot B) \cdot C]_{js},
\end{aligned}$$

where we employed the distributive laws and the associativity of multiplication among real numbers, which enabled us to rearrange the sums.

For the purposes of this thesis, only square matrices $\mathbb{R}^{n\times n}$ operating on other such matrices or vectors from $\mathbb{R}^n$ are relevant – therefore we conclude this subsection with some remarks on those.

First, we observe that the set $\mathbb{R}^{n\times n}$ is closed under the matrix product: Multiplying two such matrices will yield another matrix of size $n \times n$.

Secondly, there is an identity of multiplication, namely the unit matrix $\mathbb{1}_n$, where $(\mathbb{1}_n)_{jk} = \delta_{jk}$. For $A \in \mathbb{R}^{n\times n}$:

$$\begin{aligned}
(A \cdot \mathbb{1}_n)_{js} &= \sum_{k} A_{jk} (\mathbb{1}_n)_{ks} = \sum_{k} A_{jk} \delta_{ks} = A_{js} \\
\text{and} \quad (\mathbb{1}_n \cdot A)_{js} &= \sum_{k} (\mathbb{1}_n)_{jk} A_{ks} = \sum_{k} \delta_{jk} A_{ks} = A_{js}, \\
\text{thus} \qquad & A \cdot \mathbb{1}_n = \mathbb{1}_n \cdot A = A
\end{aligned}$$

Together with the associativity of the matrix product, $(\mathbb{R}^{n\times n}, \cdot)$ is a monoid. (If a matrix is not square, the unit matrices with of the proper size can be multiplied from the left-hand or right-hand side, respectively, yielding the original matrix – this means there are left and right identities, but not a two-sided one which would be required for a monoid structure.)

Because the matrices in $\mathbb{R}^{n\times n}$ also satisfy the vector space conditions, $(\mathbb{R}^{n\times n}, +)$ is an Abelian group with the component-wise addition. It can be shown that the matrix product distributes over the component-wise addition, therefore $(\mathbb{R}^{n\times n}, +, \cdot)$ is a ring with unity.

When the kind of product (scalar multiplication or matrix multiplication) is evident from context, we will omit the $\cdot$ sign from here on, except where it adds clarity.

**Definition A.24** *Two square matrices $M, V \in \mathbb{R}^{n\times n}$* commute *if $MV = VM$*

Examples:

- If $M = \mathbb{1}_n$, then $MV = VM = V$.
- If $M = 0$ (the zero matrix), then $MV = VM = 0$.

### A.3.8 Transposed Matrices and Vectors

If a column vector from $\mathbb{R}^n$ can be expressed as a matrix from $\mathbb{R}^{n\times 1}$, we can also conceive of a row vector, where the elements are written horizontally, i.e. a matrix from $\mathbb{R}^{1\times n}$.

While we observe that such a structure must behave differently under matrix multiplication than a column vector would, we will ignore the underlying conceptual differences arising from the

functional analysis perspective. In fact, we already tacitly did so when introducing the scalar product in Euclidean space: The left-hand argument of the product $\langle \cdot, \cdot \rangle$ is in fact an element from the dual space; this dual space is also where the row vectors originate.

This hand-waviness is possible here because we are dealing with simple vectors that can be expressed as arrays of numbers from the same field the vector spaces are built on[1].

**Definition A.25** *Given a matrix $M \in \mathbb{R}^{m \times n}$, the* transposed matrix $M^T$ *is a matrix from $\mathbb{R}^{n \times m}$, with*

$$\left(M^T\right)_{jk} := M_{kj}.$$

*A square matrix $N \in \mathbb{R}^{n \times n}$ is called a* symmetric matrix *if $N^T = N$.*

We can obtain $M^T$ by taking the columns of $M$ from left to right and writing them down as rows of $M^T$, from top to bottom (or, alternatively, making $M$'s rows into columns of $M^T$).

Per the definition, we observe that $\left(M^T\right)^T = M$ for any matrix.

Thus, if $\vec{v}$ is a (column) vector from $\mathbb{R}^n$, $\vec{v}^T$ is its associated transposed vector, a row vector.

We now examine how a matrix product behaves under transposition. Given $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, we calculate:

$$\begin{aligned} \left((AB)^T\right)_{js} &= (AB)_{sj} = \sum_{k=1}^{n} A_{sk} B_{kj} = \sum_{k=1}^{n} \left(B^T\right)_{jk} \left(A^T\right)_{ks} = \left(B^T A^T\right)_{js}, \\ \text{thus} \qquad & (AB)^T = B^T A^T \end{aligned}$$

Note that the matrix product $B^T A^T$ is well-defined if $AB$ is; there is no mismatch of column/row lengths.

### A.3.9 The Canonical Scalar Product Revisited

In this short subsection we examine the behavior of the scalar product in Euclidean space (cf. definition A.11, p. 71). For $\vec{v}, \vec{w} \in \mathbb{R}^n$, we calculate the matrix product of $\vec{v}^T$ and $\vec{w}$. We will, for this calculation, view both those vectors as matrices whose components carry two indexes each. The resulting product is a matrix from $\mathbb{R}^{1 \times 1}$:

$$\left(\vec{v}^T \cdot \vec{w}\right)_{1,1} = \sum_j \left(\vec{v}^T\right)_{1,j} (\vec{w})_{j,1} = \sum_j (\vec{v})_{j,1} (\vec{w})_{j,1}$$

Since we can identify each $1 \times 1$ matrix trivially with its single component, we can rewrite this equation (which we only stated to demonstrate a proper matrix product), leaving out the middle equality and viewing $\vec{v}, \vec{w}$ as vectors again:

$$\vec{v}^T \cdot \vec{w} = \sum_j v_j w_j = \langle \vec{v}, \vec{w} \rangle$$

This also allows us to draw an important conclusion that we will later need in chapter C about the eigenvalue problem, namely the behavior of the scalar product when one of the vectors is the result of a linear mapping. Given vectors $\vec{v}, \vec{w} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$:

$$\langle \vec{v}, A\vec{w} \rangle = \vec{v}^T (A\vec{w}) = (\vec{v}^T A)\vec{w} = \left(A^T \vec{v}\right)^T \vec{w} = \langle A^T \vec{v}, \vec{w} \rangle$$

We could have obtained the same result without using matrix product notation, but then we would have had to write out the sums, re-arrange them and use the distributive laws. Here we just used the associativity of matrix multiplication (in the third equality) and the rules about the transpose of a matrix product (in the fourth equality, applied backwards).

Regardless of the method, we observe that the scalar product of two vectors yields the same value, whether we apply a map on one of the vectors, or its transposed map on the other vector.

[1] In other cases, e.g. Quantum Mechanics, greater care must be taken: the vectors there are complex wave functions (from a Hilbert space), and the dual vectors are linear functionals operating on such functions.

### A.3.10 Obtaining matrix components

We show that, for a square matrix $A \in \mathbb{R}^{n\times n}$, the components $A_{js}$ can be obtained per

$$A_{jk} = \vec{e}_j^T \cdot A \cdot \vec{e}_k$$

First, we define the vector $\vec{b} := A\vec{e}_k$:

$$b_r = \sum_s A_{rs}(\vec{e}_k)_s = \sum_s A_{rs}\delta_{ks} = A_{rk}$$

Then we multiply the left-hand row vector with the column vector $\vec{b}$, which yields a single real number:

$$\sum_r (\vec{e}_j)_r b_r = \sum_r \delta_{jr} b_r = \sum_r \delta_{jr} A_{rk} = A_{jk}$$

### A.3.11 Matrix Inversion

We return to the topic of square matrices. As per our previous remarks, the set $\mathbb{R}^{n\times n}$, together with the component-wise addition $+$ and the matrix product $\cdot$, constitutes a ring with unity.

While $\mathbb{1}_n \in \mathbb{R}^{n\times n}$ is the two-sided identity of the matrix product, it is not evident that every square matrix should also have a multiplicative inverse. In order for the square matrices to form a group under matrix multiplication, such an inverse $A^{-1}$ would have to be unique for any $A$, and two-sided.

**Definition A.26** *A matrix $A \in \mathbb{R}^{n\times n}$ is called* invertible *if (and only if) there is a matrix $X \in \mathbb{R}^{n\times n}$ such that*

$$A \cdot X = X \cdot A = \mathbb{1}_n$$

*The matrix $X$ is called the* Inverse *of $A$, and written as $A^{-1}$.*

Such invertible matrices do in fact exist, and make up a group, the *General Linear Group*, which we will briefly mention in the first algebra chapter, after we have discussed determinants (cf. the following chapter for those).

At this point, it is already possible to observe some general behavior of inversion. To start with, the inverse of an inverse matrix $A^{-1}$ is $A$, because the above definition is symmetrical in $A$ and $A^{-1}$.

Assuming that $A$ has an inverse, we can use the symmetry of $\mathbb{1}_n$ and transpose the above equation:

$$\left(A^{-1}\right)^T \cdot A^T = A^T \cdot \left(A^{-1}\right)^T = \mathbb{1}_n^T = \mathbb{1}_n$$

Thus, if $A$ is invertible, so is $A^T$, and its inverse is the transpose of $A^{-1}$.

Also, for a product $AB$ of two matrices, to be invertible, we must have:

$$(AB) \cdot (AB)^{-1} = (AB)^{-1} \cdot (AB) = \mathbb{1}_n$$

To obtain the form of $(AB)^{-1}$, we use the associativity of the matrix product. Ignoring the middle equality, we can rewrite:

$$A \cdot B \cdot (AB)^{-1} = \mathbb{1}_n$$

In order to extract the third factor, we multiply the inverses $A$ and $B$ (in this order) from the left-hand side. If those inverses exist, we obtain:

$$(AB)^{-1} = B^{-1} \cdot A^{-1} \cdot \mathbb{1}_n = B^{-1}A^{-1}$$

If we had used the middle equality, we would have had to multiply the inverses of $B$ and $A$ (in this order) from the right-hand side, which yields the same result.

In the next chapter, we will develop a way to decide on the existence of an inverse matrix using the determinant. This is not strictly necessary, because, we can give a method to calculate the inverse. If the calculation fails, the inverse does not exist.

We recall the subsection on linear maps (cf. definition A.20, p. 74) from above and relabel the objects in its central statement: For a matrix $A \in \mathbb{R}^{n\times n}$ and vectors $\vec{x}, \vec{b}$:

$$A\vec{x} = \vec{b}$$

If we have given coefficients $A_{jk}$ and a given resulting vector $\vec{b}$, we can obtain the unknown vector $\vec{x}$ by inverting $A$:

$$A^{-1} \cdot A\vec{x} = \left(A^{-1} \cdot A\right) \cdot \vec{x} = \mathbb{1}_n \cdot \vec{x} = \vec{x} = A^{-1} \cdot \vec{b}$$

This is just what happens when we solve a system of $n$ linear equations with $n$ (unknown) variables. Using Gaussian elimination[2], we transform the extended matrix $(A|\vec{b})$ into $(\mathbb{1}_n|A^{-1}\vec{b})$ and implicitly calculate the inverse of $A$.

Note that such a system of linear equations need not be uniquely solvable: If Gaussian elimination leads to a zero row in the left-hand part of the extended matrix, and a non-zero value in the right-hand part of the same row, the system has no solution. Or the solution could exist, but not be unique – if Gaussian elimination leads to a complete zero row in the extended matrix, corresponding to the trivial equation $\vec{0}^T\vec{x} = 0$. Each such row introduces one free variable.

Zero rows in the left-hand part of the extended matrix can (only!) occur when $A$'s rows are linearly dependent (then there is a non-trivial linear combination of $\vec{0}^T$, which can be obtained by Gaussian elimination).

Those considerations lead to the concept of *matrix rank*, which is determined by the number of linearly independent row vectors of a matrix (or column vectors, which yields the same value). If the rank equals $n$, the (square) matrix has full rank and is invertible.

From the topics laid out up to now, it is not evident that this (necessary) criterion is also sufficient. In order to show the latter, we would have to expand further on null spaces, bases and dimension, which we will omit in this appendix. Suffice it to say that the determinant calculation will yield an equivalent necessary criterion (because the determinant is sensitive to linear dependence of the row/column vectors of a matrix) and suffer from the same deficit.

For an explicit calculation of $A^{-1}$, we can adapt this method to a matrix $X$ of unknowns, and instead of some given vector $\vec{b}$, we put the matrix $B := \mathbb{1}_n$. So we solve the equation $AX = \mathbb{1}_n$ for $X$, and we can do this with Gaussian elimination in the same way, transforming the extended matrix $(A|\mathbb{1}_n)$ into $(\mathbb{1}_n|A^{-1}\mathbb{1}_n) = (\mathbb{1}_n|A^{-1})$. If the transformation is successful, the inverse of $A$ has been obtained.

We close this section (and chapter) with two definitions that will be useful later on, when dealing with determinants and the eigenvalue problem.

### A.3.12 Orthogonal Matrices

**Definition A.27** *A square matrix $A \in \mathbb{R}^{n\times n}$ is called* orthogonal *if (and only if) it satisfies*

$$A^T A = AA^T = \mathbb{1}_n$$

For orthogonal matrices, their respective inverses can be obtained by simple transposition.

### A.3.13 Similar Matrices

**Definition A.28** *Two square matrices $A, B \in \mathbb{R}^{n\times n}$ are called* similar *if there is an invertible matrix $S \in \mathbb{R}^{n\times n}$ such that*

$$B = S^{-1}AS$$

Note that, if the above holds, the matrix $A$ can be obtained from $B$ per

$$SBS^{-1} = S \cdot S^{-1}AS \cdot S^{-1} = (SS^{-1})A(SS^{-1}) = \mathbb{1}_n A \mathbb{1}_n = A$$

We observe an important fact:

**Corollary A.29** *If two square matrices $A, B \in \mathbb{R}^{n\times n}$ are similar via an orthogonal matrix $O \in \mathbb{R}^{n\times n}$, their transposed matrices are similar via the same matrix $O$.*

*If $A$ is symmetric, then so is $B$*

Proof: Since $O$ is orthogonal, $B = O^{-1}AO = O^T AO$. But then, using the remarks in subsection A.3.8 (p. 76), we may infer that

$$B^T = (O^T AO)^T = O^T A^T (O^T)^T = O^T A^T O$$

If $A$ is symmetric, then $A^T = A$, and thus $B^T = O^T AO = B$. ■

[2]Gaussian elimination does not influence the existence of the inverse matrix, as we will demonstrate in section B.2 of the chapter on permutations and determinants.

# Appendix B

# Permutations and Determinants

Determinants can be used to ascertain whether a square matrix is invertible, and will also be of great importance for solving the eigenvalue problem (cf. chapter C). We will characterize axiomatic conditions for the determinant function(a map $\mathbb{R}^{n\times n} \to \mathbb{R}$) and establish several properties that we need either in the thesis or for subsequent chapters here.

The determinant will be sensitive to permutations of a matrix's columns (or rows): swapping two different ones will make the determinant change its sign. In order to express this cleanly, we will look at permutations formally first. We will rely on [FS20, KM21] (chapters 4 and 9, respectively) but present some calculations of our own as well.

## B.1 Permutations

Given $n$ numbered objects (counting from 1 to $n$), there are several ways to arrange them into a single chain, and the number of possibilities rapidly increases with $n$.

**Lemma B.1** *There are* $n! = n(n-1)(n-2)\cdot\cdots\cdot 2\cdot 1$ *different ways to order* $n$ *numbered objects.*

Proof: By induction. For $n = 1$, there is evidently only one way. For the induction step, we assume $n > 1$ and that the statement holds for $(n-1)$ objects. For each of those $(n-1)!$ orderings, we now add the $n$-th object. We can place this at any of the $(n-2)$ positions between the $(n-1)$ objects, or before the first one, or behind the last one. Each of the $(n-1)!$ orderings therefore offers $n$ possibilities to achieve a new ordering of the $n$ objects; this makes for $n\cdot(n-1)! = n!$ orderings. ■

### B.1.1 Definition and Representation

More formally:

**Definition B.2** *A total bijective map* $\sigma : \{1,\cdots,n\} \to \{1,\cdots,n\}$ *is called a* permutation. *Two such maps* $\sigma, \pi$ *are equal if* $\forall j \in \{1,\cdots,n\} : \sigma(j) = \pi(j)$

We note that, because of bijectivity, for any $k$ in $\{1,\cdots,n\}$ there is exactly one $j$ in the same set that satisfies $k = \sigma(j)$. There are $n!$ different such permutations on the set $\{1,\cdots,n\}$.

One way to represent a permutation is by listing the element connections (an alternative way is presented in the subsection after next). The following example is a permutation of the numbers 1 to 10:

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma(j)$ | 5 | 1 | 8 | 4 | 2 | 10 | 9 | 6 | 7 | 3 |

The inverse permutation can be derived directly from that table, by swapping the lines. It is customary, but not necessary, to sort them so that the upper line is a linear progression again.

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma^{-1}(k)$ | 2 | 5 | 10 | 4 | 1 | 8 | 9 | 3 | 7 | 6 |

**Definition B.3** *A permutation that swaps only two numbers $j, k \neq j$ and leaves every other number unchanged is called a* transposition $\tau_{jk}$*:*

$$\tau_{jk}(r) := \begin{cases} k & r = j \\ j & r = k \\ r & otherwise \end{cases}$$

We observe that $\tau_{jk}^{-1} = \tau_{jk}$, and that $\tau_{jk} = \tau_{kj}$.

### B.1.2 The Symmetric Group

Permutations can be chained together; the result is another permutation. The operation is the standard function composition $\circ$, which is associative. For $\pi, \sigma$ permutations on the set $\{1, \cdots, n\}$ and any $1 \leq j \leq n$:

$$(\pi \circ \sigma)(j) := \pi(\sigma(j))$$

$\pi \circ \sigma$ is a bona fide permutation itself and can be calculated by drawing up two tables like in the above example. Its inverse permutation is given by $\sigma^{-1} \circ \pi^{-1}$, applying the inverse single permutations in reverse order.

The composition operation has an identity:

**Definition B.4** *The* identical permutation $\mathrm{id}_n$ *on the set* $\{1, \cdots, n\}$ *is defined by:*

$$\forall j \in \{1, \cdots, n\} : \mathrm{id}_n(j) = j$$

Thus, for any $\sigma$: $\sigma \circ \mathrm{id} = \mathrm{id} \circ \sigma = \sigma$.

Since permutations are by definition bijective, every permutation $\sigma$ has a unique inverse, and $\sigma^{-1} \circ \sigma = \sigma \circ \sigma^{-1} = \mathrm{id}$.

Together, this satisfies the requirements G1 to G4 of a group (cf. definitions A.1 to A.3, p. 66):

**Corollary B.5** *The permutations on the set* $\{1, \cdots, n\}$*, together with the function composition* $\circ$*, form a finite group with* $n!$ *elements, the* symmetric group $S_n$*.* ■

(The first algebra chapter D (p. 104) will explain the reference to symmetry.)

**Lemma B.6** *Only* $S_1$ *and* $S_2$ *are Abelian. For* $n > 2$*,* $S_n$ *is not Abelian.*

Proof: $S_1$ is Abelian, because it only contains $\mathrm{id}_1$, and $S_2$ is, too, because it only contains $\mathrm{id}_2$ and $\tau_{1,2}$ – and $\mathrm{id}_2$ is the identity.

We give an example of $S_3$ to prove the assertion:

| $j$ | 1 | 2 | 3 |
|---|---|---|---|
| $\sigma(j)$ | 1 | 3 | 2 |
| $\pi(j)$ | 2 | 3 | 1 |
| $(\pi \circ \sigma)(j)$ | 2 | 1 | 3 |
| $(\sigma \circ \pi)(j)$ | 3 | 2 | 1 |

Evidently, $\pi \circ \sigma \neq \sigma \circ \pi$.

This example serves for any $S_n$ with $n > 3$, too, if $\forall j > 3 : \sigma(j) := \pi(j) := j$. ■

### B.1.3 Cycles

An alternative way to define a permutation is by writing down its various *cycles*. From the above table notation, it is not immediately obvious that this is always possible. We will demonstrate that for any $\sigma \in S_n$ there is a unique minimal decomposition into cycles that lists each number in $\{1, \cdots, n\}$ at exactly once (or at most once, depending on taste). In such a decomposition, one can infer (or read directly) all the pairings $j \mapsto \sigma(j)$, and therefore it serves as an equivalent representation of $\sigma$ (as compared to the permutation table).

We will also show that there are infinitely more decompositions of $\sigma$ into cycles. While this may be useful for other problems (like determining the sign of a permutation; see the following subsections for that), it will usually increase the number of cycles, and the mappings of $\sigma$ cannot be extracted directly anymore – those decompositions, therefore, are not as useful to represent $\sigma$.

Since the symmetric group features only one operation, we will omit the $\circ$ symbol from here on where this does not reduce clarity.

**Definition B.7** *A permutation in $\zeta \in S_n$ is called a* cycle*, written $\zeta = (a_1\ a_2\ \cdots\ a_k)$, if it describes a circular shift of a non-empty subset of $\{1, \cdots, n\}$ containing the numbers $\{a_1, \cdots, a_k\}$, while leaving all the other numbers unchanged:*

$$\zeta(a_1) = a_2; \zeta(a_2) = a_3; \cdots; \zeta(a_{k-1}) = a_k; \zeta(a_k) = a_1; \qquad \zeta(j) = j \text{ otherwise}$$

*The number of affected elements $k$ is called the* length of $\zeta$*, or $|\zeta|$. Each element $a.$ occurs exactly once in the sequence.*

*Two cycles $\zeta_1, \zeta_2$ are* identical *if the elements of $\zeta_2$ can be shifted circularly until they show the same sequence of numbers as $\zeta_1$.*

*Cycles of length 1 are* trivial *because they map every number onto itself, i.e. such cycles equal* $\mathrm{id}_n$*.*

*Cycles are called* disjoint *if the the subsets of numbers making up their sequence are disjoint.*

Examples:

- (1 2 3) is a cyclic shift $1 \mapsto 2$, $2 \mapsto 3$, $3 \mapsto 1$. If $n > 3$, all other numbers are mapped onto themselves.
- (3 1 2) is identical to the previous cycle. Identical cycles are always equal permutations, but the reverse is not true.
- (1)(4)(3)(2) is a composition of trivial cycles. If viewed as a permutation in $S_4$, it is one of the unique representations of $\mathrm{id}_4$.
- (1) is the shortest way to express $\mathrm{id}_n$ as a (trivial) cycle. It is also a valid way to express $\mathrm{id}_4$ in $S_4$. In fact, if a composition of cycles contains non-trivial cycles, all trivial ones can be omitted. It is a matter of convention if all trivial cycles are listed or not.
- A transposition $\tau_{jk} \in S_n$ equals the cycle $(j\ k)$.
- (1 2 3 2 3) is *not* a cycle. Not only are elements occurring more than once in the sequence, but the represented permutation cannot be defined because it would contain the mappings $3 \mapsto 2$ and $3 \mapsto 1$.
- (1 2 3 1 2 3) is also not a proper cycle. While the mappings are fine, there is redundant information, which was forbidden in the definition.
- (1 2 3)(1 2 3) is a permutation composed of two identical cycles; we can also write it as $(1\ 2\ 3)^2$. The effect equals two counter-clockwise circular shifts by 1 in the sequence.
- $(1\ 2\ 3)^3$ can be shortened to $\mathrm{id}_n$ (for $n \geq 3$): see the following corollary.
- $(3\ 7\ 2)^{22}$ can be shortened to (3 7 2): there are seven threefold cyclic shifts that amount to $\mathrm{id}_n$ ($n \geq 7$), and only one remaining shift that actually changes elements.

**Corollary B.8** *For a cycle $\zeta = (a_1\ \cdots\ a_k) \in S_n$, $\zeta^{-1} = (a_k\ \cdots\ a_1)$. Also, $\zeta^k = \mathrm{id}_n$.*

Proof: The inverse cycle can be obtained by writing the sequence of $\zeta$ in reverse. This undoes all the changes of $\zeta$ in the above definition. If $\zeta(a_j) = a_{j+1}$, then $\zeta^{-1}(a_{j+1}) = a_j$.

The product of $k$ times $\zeta$ involves $k$ circular shifts of the $k$ numbers in $\zeta$; this means that the sequence, in effect, remains unchanged; any of its elements is mapped onto itself. ■

From the above definition, we can conclude that, for any $j \in \mathbb{N}; r \in \{1, \cdots, k\}$:

$$\zeta^j(a_r) = a_s \qquad \text{with } s = 1 + (((r-1) + j) \bmod k)$$

(The shifting by 1 was necessary because the residues go from 0 to $(k-1)$, not from 1 to $k$.)

**Lemma B.9** *Disjoint cycles commute.*

Proof: Let $\zeta_1, \zeta_2 \in S_n$ disjoint cycles, and $j \in \{1, \cdots, n\}$. If at least one of the two cycles is trivial (length 1), there is nothing to show because $\mathrm{id}_n$ commutes with any permutation in $S_n$. Thus, we only need to consider non-trivial cycles. Because the cycles are disjoint, we know that any element $j$ occurring in the sequence of $\zeta_1$ will not appear in the sequence of $\zeta_2$, nor will its predecessor

or successor (which both belong to $\zeta_1 1$'s sequence). $\zeta_2$ will therefore map such an element, its successor and its predecessor, onto themselves, respectively. Thus:

$$\begin{array}{rcll} (\zeta_1 \circ \zeta_2)(j) & = & \zeta_1(j) & (\zeta_2 \text{ does not change } j) \\ (\zeta_2 \circ \zeta_1)(j) & = & \zeta_1(j) & (\zeta_2 \text{ does not change successor of } j) \\ (\zeta_1^{-1} \circ \zeta_2)(j) & = & \zeta_1^{-1}(j) & (\zeta_2 \text{ does not change } j) \\ (\zeta_2 \circ \zeta_1^{-1})(j) & = & \zeta_1^{-1}(j) & (\zeta_2 \text{ does not change predecessor of } j) \end{array}$$

The same argument can be used for the sequence elements of $\zeta_2$, which are not affected by $\zeta_1$. Thus, $\zeta_1$ and $\zeta_2$ commute. ■

We now present the main claim of this subsection. In the course of its proof, there will be several observations on the ways to compose cycles into a permutation (or, v.v., decompose a permutation into cycles).

**Claim B.10** *Any permutation $\sigma \in S_n$ can be decomposed into $c$ disjoint "canonical" cycles ($1 \leq c \leq n$):*

$$\sigma = \zeta_1 \circ \cdots \circ \zeta_c$$

*This decomposition is unique, up to circular shifts of the sequences of each $\zeta_\cdot$, and up to the ordering of the $\zeta_\cdot$ in the composition.*

*The cycle count $c$ is minimal (but varies due to whether trivial cycles are omitted or not).*

Before the proof, we present a working example for illustration: In $S_6$, consider

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\sigma(j)$ | 2 | 3 | 1 | 5 | 4 | 6 |

It can be easily verified that the following decomposition into three cycles represents $\sigma$:

$$\sigma = (1\ 2\ 3)(4\ 5)(6)$$

(We opt to count the trivial cycles for the time being.)

Its cycles are indeed disjoint, which means every number occurs exactly once. In the proof we will see that any decomposition into different cycles introduces a multiple occurrence of at least one number and will possibly consist of more cycles. Also, permutations do not generally commute. When the cycles are no longer disjoint, ordering becomes relevant (execution is always from right to left).

The cycle (1 2 3) maps the sequence $(1, 2, 3)$ onto $(2, 3, 1)$. The following two decompositions of that cycle achieve the same: (1 2)(2 3) and (1 3)(1 2). In both cases, one number occurs twice, so the cycles are no longer disjoint. Also, for the whole $\sigma$ decomposition, the cycle count would increase to 4.

We can also give an alternative decomposition of the product (1 2 3)(4 5). This maps the sequence $(1, 2, 3, 4, 5)$ onto $(2, 3, 1, 5, 4)$. The same is achieved by (1 4)(1 2 3 4 5). This time the cycle count of $\sigma$'s decomposition remains at three, but we incur a double occurrence of both the numbers 1 and 4.

Now for the proof: We recall that any permutation in $S_n$ is a total bijective map of a finite set of $n$ numbers onto itself. Each of the $n$ numbers is therefore directly connected with one predecessor and one successor. If we use the permutation table to line up numbers connected by mappings, i.e. if we construct a directed graph of the mappings, we will therefore never encounter any forking (in either direction). Since we can never have a situation where a number does not have a predecessor or successor (i.e. a graph node with only an outgoing or incident edge, respectively), the only structure we can build is a set of cycles (understood here in a graph-theoretical sense). There may be several such cycles, but those must be separate because there can be no forking from one cycle to another. Cycles consisting of a single element are allowed.

The graph of mappings, therefore has $n$ nodes and $n$ edges, and is structured into $c$ disjoint cycles. If there is any mapping at all in $\sigma$, there must be a cycle, so $c \geq 1$. Because there are only $n$ mappings, there can be no more cycles than $n$, so $c \leq n$. Those extreme cases are circular shifts between all elements and the identical permutation $\mathrm{id}_n$, respectively.

Because the graph contains nothing more or less than the complete mappings of the permutation $\sigma$, it is a unique representation.

We can translate the cycles of $\sigma$'s mapping graph into cycle permutations directly, by creating one cycle permutation $\zeta_k$ for each graph cycle. For each of those $\zeta_k$, we pick one number belonging

to the graph cycle as the leading number of $\zeta_k$'s sequence. From there, we define each further member of the sequence as the next node/number on the cycle graph. This process will always terminate because of the above general considerations.

Because the graph cycles are disjoint, the cycle permutations constructed from them are, too. Each mapping is used in exactly one cycle permutation, and each number is part of exactly one such permutation's element sequence.

The $\zeta_k$ constructed this way are commutative (see the above lemma on disjoint cycle permutations). Also, they are only defined up to circular shifts of their respective sequence elements, because picking the starting element was arbitrary (yet we recall that the definition calls all those cycle permutations identical).

To conclude the proof, we argue that the cycle count $c$ obtained from this construction is minimal for any $\sigma$. As demonstrated before, it is possible to compose $\sigma$ of more than $c$ cycle permutations. But the cycle graph of $\sigma$ is irrespective of the concrete decomposition; it can never have fewer than $c$ cycles.

Only by using the above construction can we ensure that for each $j \in \{1, \cdots, n\}$ there is exactly one $k$, $1 \leq k \leq c$ with $\sigma(j) = \zeta_k(j)$ and where $j$ appears in the defining sequence of $\zeta_k$. If there were any cycle $\zeta$ with $\zeta(j) \neq \sigma(j)$, this inequality would have to be compensated. This can happen trivially by having $\zeta(j) = j$, i.e. $\zeta$ a trivial cycle permutation of length 1, which amounts to $\mathrm{id}_n$, and increases the cycle count by one without any effect on the mapping. Or it must happen in another cycle, because the number $j$ can only appear once in $\zeta$. This other cycle cannot be one of the cycles obtained by the above construction because there sequence elements are defined by $\sigma$ alone. While we have demonstrated above that there are cases where this can be achieved without increasing the overall cycle count, it can never serve to reduce the number of cycle permutations to fewer than $c$, and (if $|\zeta| > 1$) it will introduce cycle permutations that are no longer disjoint. ■

Note that the above construction yields as many trivial cycles as the mapping graph of $\sigma$ contains, so $\mathrm{id}_n$ becomes decomposed into $(1)(2)\cdots(n)$. Per the cycle permutation definition, the identical permutation can also be written as just one trivial cycle containing a single element.

Example: For the ten-element permutation from the beginning of this chapter, the cycle decomposition (including trivial cycles) reads: (1 5 2)(3 8 6 10)(4)(7 9). We have rotated each cycle so that its first element is its smallest one, and we have ordered the cycles by the values of their first elements. This is always possible for disjoint cycles.

We can obtain this unique disjoint representation directly from the permutation table. The first cycle starts with 1. We strike out all columns visited in this cycle. For the next cycle, we take the leftmost (smallest) remaining $j$ and repeat.

**Corollary B.11** *Any permutation $\sigma \in S_n$, $\sigma \neq \mathrm{id}_n$, contains at least one non-trivial cycle $\zeta$ with $|\zeta| > 1$.*

Proof: If $\sigma \neq \mathrm{id}_n$, there must be a $j \in \{1, \cdots, n\}$ with $\sigma(j) \neq j$. Since any node on the cycle graph of $\sigma$ is part of one cycle, $j$ is part of a cycle with at least the elements $j$ and $\sigma(j)$. Such a cycle cannot be expressed by trivial cycle permutations alone. ■

**Corollary B.12** *If $\sigma \in S_n$ is decomposed into disjoint cycles by the technique described in the main proof, and if trivial cycles are not omitted, the sum of the lengths of the cycles is $n$.*

Proof: We write all disjoint cycles as determined from the mapping graph, trivial or not. Because any element belongs to exactly one of those disjoint cycles, all $n$ elements are written exactly once. ■

**Corollary B.13** *While the disjoint cycles of the mapping graph of any $\sigma \in S_n$ are unique (up to circular shifting), $\sigma$ may also be written (if $n > 1$) as the composition of cycles that are not disjoint.*

Proof: Such a composition exists: For some $k, j \in \{1, \cdots, n\}$

$$\sigma \circ (j\ k)(j\ k) = \sigma \quad \blacksquare$$

There exist other such compositions of non-disjoint cycles, which will be the next subsection's topic.

### B.1.4 Decomposing a Permutation into Transpositions

The following is an significant intermediary result because it will help us calculate an important property – the *sign* of a permutation – in the next subsection.

**Claim B.14** *Any $\sigma \in S_n$, $n > 1$, can be written as a composition of only transpositions, i.e. cycles of length 2.*

Proof: Since we already have established that $\sigma$ can be decomposed into disjoint cycles, it suffices to show that the statement applies to any cycle permutation.

Cycles of length 1 are trivial because they involve zero transpositions.

Cycles of length 2 are just single transpositions.

For larger cycles, we will give two formulas. We want to map the numbers $(a_1, a_2, \cdots, a_{k-1}, a_k)$ (without the commas, this defines the corresponding cycle permutation of length $k$) onto the numbers $(a_2, a_3, \cdots, a_k, a_1)$. For the course of this proof, it may be helpful to imagine the numbers as an array of length $k$, and the permutation will indicate the new array position of an element (although in fact the permutation only acts on numbers, however they might be arranged).

We first give a decomposition found in [KM21] (Lemma 9.1):

$$(a_1\ a_2\ a_3\ \cdots\ a_{k-1}\ a_k) = (a_1\ a_2)(a_2\ a_3)\cdots(a_{k-1}\ a_k),$$

of which the authors write that it is "evidently correct" – and it is, as we will demonstrate a bit later.

If we read this decomposition from left to right, it looks suggestively like the positions are swapped one by one – first, $a_1$ takes the place of $a_2$ (which moves into first position), then the positions 2 (containing $a_1$ now) and 3 are swapped, and so on – only that this is not what actually happens, because composed permutations are executed from right to left.

The rightmost of the transpositions is, therefore, the first to be executed, and it swaps the positions of the two last elements, putting $a_k$ in its proper new position. However, $a_{k-1}$ is now in last position, which should, in the end, hold $a_1$. This is indeed (successively) accomplished by the following transpositions.

Before we return to this decomposition, we want to motivate an alternative, that in fact does what we described might be naively imagined when one reads the previous decomposition from left to right, namely, swap the element $a_1$ successively through the following positions until it reaches the end (of the array).

The first transposition, then, swaps $a_1$ and $a_2$ per $(a_1\ a_2)$, putting $a_2$ in the correct position. After that, $a_1$ is where $a_3$ should go, so we swap those with $(a_1\ a_3)$, putting $a_1$ in the old place of $a_3$. The next transposition therefore would be $(a_1\ a_4)$, and so on, until $a_1$ has been swapped to occupy the position before $a_k$. The last swap exchanges those two, and we obtain for the complete decomposition:

$$(a_1\ a_2\ a_3\ \cdots\ a_{k-1}\ a_k) = (a_1\ a_k)(a_1\ a_{k-1})\cdots(a_1\ a_3)(a_1\ a_2)$$

We demonstrate this with an array of six numbers; to the left of the separator we write the transposition used to obtain the array to the right, and we highlight the changed mappings/positions by capital letters:

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ |
|---|---|---|---|---|---|---|
| $(a_1\ a_2)$ | $A_2$ | $A_1$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ |
| $(a_1\ a_3)$ | $a_2$ | $A_3$ | $A_1$ | $a_4$ | $a_5$ | $a_6$ |
| $(a_1\ a_4)$ | $a_2$ | $a_3$ | $A_4$ | $A_1$ | $a_5$ | $a_6$ |
| $(a_1\ a_5)$ | $a_2$ | $a_3$ | $a_4$ | $A_5$ | $A_1$ | $a_6$ |
| $(a_1\ a_6)$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $A_6$ | $A_1$ |

We observe that each transposition moves the element with higher index into the correct position (i.e. establishes the right mapping to this element), and all but the last transpositions introduce an incorrect mapping onto $a_1$, which is corrected in the next step.

This can be determined from the formula, too, without the recourse to array positions: The first transposition introduces a mapping from $a_1$ to $a_2$ (which is correct), and one from $a_2$ to $a_1$. The next transposition corrects this mis-mapping by mapping $a_1$ to $a_3$, thus, transitively, $a_2$ now maps to $a_3$ via the intermediary $a_1$. The next step provides the transitive correction $a_3 \mapsto a_1 \mapsto a_4$, and so on.

Returning to the decomposition given in [KM21], we present the array view for six numbers, too:

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ |
|---|---|---|---|---|---|---|
| $(a_5\ a_6)$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $A_6$ | $A_5$ |
| $(a_4\ a_5)$ | $a_1$ | $a_2$ | $a_3$ | $A_5$ | $a_6$ | $A_4$ |
| $(a_3\ a_4)$ | $a_1$ | $a_2$ | $A_4$ | $a_5$ | $a_6$ | $A_3$ |
| $(a_2\ a_3)$ | $a_1$ | $A_3$ | $a_4$ | $a_5$ | $a_6$ | $A_2$ |
| $(a_1\ a_2)$ | $A_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $A_1$ |

This method has the advantage that no element but the last one is temporarily wrongly mapped. Each transposition introduces one correct mapping $a_j \mapsto a_{j+1}$, which can be read directly from the formula. However, the mapping of the last element is changed in every step, successively reducing the index down to one – this means that the necessary correction is cascaded over all the steps.

On the whole, the literature decomposition may be easier to read; even the mapping $a_k \mapsto a_1$ via the cascade can be determined by the practiced eye. Both methods, however take exactly $(k-1)$ transpositions to effect the cyclic shift. ■

Note that the ordering of the transpositions in both methods is important, and that none of the mini-cycles of two elements is disjoint with all the others.

We have now shown that any transposition can be decomposed into transpositions, which enables us to define the sign of a permutation.

### B.1.5 The Sign of a Permutation

Usually (cf. [FS20, KM21]), the sign of a permutation is defined by using the information in the permutation table, creating a product of fractions that evaluates to $\pm 1$. We would like to provide an equivalent recursive definition of the sign here that uses the decomposition into transpositions and is more suited to our aims[1].

**Definition B.15** *The* sign of a permutation *is a function obeying the following conditions:*

*SP1:* $\forall n \in \mathbb{N} : \operatorname{sign}(\mathrm{id}_n) := +1$ *(Normalization)*

*SP2:* $\forall \sigma, \tau_{jk} \in S_n : \operatorname{sign}(\tau \circ \sigma) := -\operatorname{sign}(\sigma) =: \operatorname{sign}(\sigma \circ \tau)$ *(Transpositions flip the sign)*

*A permutation $\sigma$ is called* even *if it has sign* $(+1)$, odd *otherwise.*

**Corollary B.16** *A transposition $\tau_{jk} \in S_n$ has sign* $(-1)$.

Proof: Use condition SP2, then plug in SP1: $\operatorname{sign}(\tau_{jk}) = \operatorname{sign}(\tau_{jk} \circ \mathrm{id}_n) = -\operatorname{sign}(\mathrm{id}_n) = -1$ ■

**Corollary B.17** *If a $\sigma \in S_n$ can be decomposed into $t$ transpositions, its sign is $(-1)^t$.*

Proof: Use condition SP2 and corollary B.16 repeatedly. ■

We would like to give an additional formula using the cycles directly:

**Lemma B.18** *For any cycle permutation $\zeta \in S_n$ with length $k = |\zeta|$, the sign is given per*

$$\operatorname{sign}(\zeta) = (-1)^{k-1}$$

Proof: We refer to the proof of claim B.14 in the preceding subsection, where we determined that any cycle of length $k$ can be expressed by $(k-1)$ transpositions. According to the above corollary, the stated equation follows directly. ■

Since any permutation $\sigma$ can be decomposed into cycles (either the disjoint canonical ones obtained directly from the mapping graph, or other constructions (not disjoint) where deviations from the canonical cycles are compensated accordingly), the sign of $\sigma = \zeta_1 \circ \cdots \circ \zeta_c$ can also be determined by

$$\operatorname{sign}(\sigma) = \prod_{j=1}^{c} \operatorname{sign}(\zeta_j) = \prod_{j=1}^{c} (-1)^{|\zeta_j|-1}$$

**Lemma B.19** *For any permutation $\sigma \in S_n$:* $\operatorname{sign}(\sigma^{-1}) = \operatorname{sign}(\sigma)$

[1] The definition is aesthetically unpleasant because it is not directly defined via the permutation table, and because decompositions are not unique. We omit the proof that all decompositions of a permutation yield the same sign.

Proof: We decompose $\sigma$ into $t$ transpositions:

$$\sigma = \tau_{j_t,k_t} \circ \cdots \circ \tau_{j_1,k_1}$$

Because transpositions are self-inverse, we can construct $\sigma^{-1}$ by applying all those transpositions in reverse, thus

$$\sigma^{-1} = \tau_{j_1,k_1} \circ \cdots \circ \tau_{j_t,k_t}$$

For both $\sigma \circ \sigma^{-1}$ and $\sigma^{-1} \circ \sigma$, the respective inner transpositions cancel each other out to yield $\mathrm{id}_n$. Since $\sigma^{-1}$ contains exactly as many transpositions as $\sigma$, its sign is identical, too. ■.

Proof alternative: Decompose $\sigma$ into the disjoint canonical cycles, replace every such cycle by its inverse, which is (see corollary B.8, p. 82) a cycle with the same respective length and, consequently, the same sign. In all, the sign of the permutation is unaffected under inversion.

**Corollary B.20** *For a composition $\sigma = \sigma_2 \circ \sigma_1$ of permutations in $S_n$, the sign of $\sigma$ is given per*

$$\mathrm{sign}(\sigma) = \mathrm{sign}(\sigma_2) \cdot \mathrm{sign}(\sigma_1)$$

Proof: Decompose $\sigma_1, \sigma_2$ into $t_1$ and $t_2$ transpositions, respectively. Thus, $\sigma$ can be decomposed into $t_1 + t_2$ transpositions. According to corollary B.17, the sign computes as

$$\mathrm{sign}\,\sigma = (-1)^{(t_1+t_2)} = (-1)^{t_1} \cdot (-1)^{t_2} = \mathrm{sign}(\sigma_1) \cdot \mathrm{sign}(\sigma_2) \quad ■$$

### B.1.6 Permutation Matrices

We consider the identity matrix

$$\mathbb{1}_n = [\vec{e}_1, \cdots, \vec{e}_n]$$

and define a new matrix composed of the Euclidean unit vectors:

**Definition B.21** *For a permutation $\sigma \in S_n$, the* permutation matrix $P_\sigma \in \mathbb{R}^{n\times n}$ *is defined by*

$$P_\sigma := [\vec{e}_{\sigma(1)}, \cdots, \vec{e}_{\sigma(n)}]$$

Evidently, $P_\sigma$ contains the Euclidean unit (column) vectors, in the order specified by $\sigma$.

**Lemma B.22** *For any $\sigma \in S_n$, the permutation matrix $P_\sigma$ is orthogonal.*

Proof: We consider the transpose $P_\sigma^T$ and observe that it has the following structure of stacked row vectors:

$$P_\sigma^T = \begin{bmatrix} \vec{e}_{\sigma(1)}^T \\ \cdots \\ \vec{e}_{\sigma(n)}^T \end{bmatrix}$$

Therefore, the components of the matrix product between $P_\sigma^T$ and $P_\sigma$ are obtained per

$$(P_\sigma^T \cdot P_\sigma)_{js} = \sum_k (P_\sigma^T)_{jk}(P_\sigma)_{ks} = \sum_k (P_\sigma)_{kj}(P_\sigma)_{ks} = \sum_k (\vec{e}_{\sigma(j)})_k (\vec{e}_{\sigma(s)})_k = \langle \vec{e}_{\sigma(j)}, \vec{e}_{\sigma(s)} \rangle = \delta_{js} \quad ■$$

Since $P_\sigma$ is orthogonal, it is also invertible. We will now demonstrate a way to transform a given square matrix $A$ into a similar matrix $\tilde{A}$ by permuting the coordinate axes.

**Lemma B.23** *For any given matrix $A \in \mathbb{R}^{n\times n}$ and any permutation $\sigma \in S_n$, the matrix $\tilde{A} \in \mathbb{R}^{n\times n}$ satisfying $\tilde{A}_{j,k} := A_{\sigma(j),\sigma(k)}$ can be obtained per*

$$\tilde{A} = P_\sigma^T A P_\sigma$$

Proof: We make use of the column and row vector notations. First, we look at the product $B := AP_\sigma$. The components $B_{jk}$ can be determined by multiplying the $j$-th row of $A$ with the $k$-the column of $P_\sigma$, which is $\vec{e}_{\sigma(k)}$.

We recall from the previous chapter (cf. subsection A.3.10, p. 78) that we can obtain matrix components by multiplying Euclidean unit vectors; it therefore turns out that

$$B_{j,k} = A_{j,\sigma(k)}$$

Thus, the columns of the product $B = AP_\sigma$ are just the columns of $A$, permuted according to $\sigma$.

Likewise, multiplying $P_\sigma^T$ with $B$ will permute the rows of $B$: The component $\tilde{A}_{rs}$ is obtained by multiplying the $r$-th row of $P_\sigma^T$ (which is $\vec{e}_{\sigma(r)}^T$) with the $s$-th column of $B$; this yields

$$\tilde{A}_{r,s} = B_{\sigma(r),s}$$

But then we can use the above equation for the components of $B$ and obtain:

$$\tilde{A}_{r,s} = B_{\sigma(r),s} = A_{\sigma(r),\sigma(s)} \quad \blacksquare$$

This concludes our examination of permutations. The next section deals with matrix determinants, and will rely on this groundwork.

## B.2 Determinants

The definitions and some of the proofs used here are taken from [FS20], but adapted to suit our needs.

### B.2.1 Definition and Basic Properties

**Definition B.24** *A map* $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$ *is called a* determinant*, if, for any matrix* $A = [\vec{a}_1, \cdots, \vec{a}_n] \in \mathbb{R}^{n\times n}$*, the following criteria are met (Weierstraß axioms):*

*D1:* $\det \mathbb{1}_n = 1$ *(Normalization)*

*D2: If* $\vec{a}_j = \vec{a}_k$ *for some* $j \neq k$*, then* $\det A = 0$ *(*det *is alternating)*

*D3: For* $\alpha, \beta \in \mathbb{R}$*, and for an arbitrary but fixed column position:*

$$\det[\cdots, \alpha\vec{a}_j + \beta\vec{a}_k, \cdots] = \alpha \det[\cdots, \vec{a}_j, \cdots] + \beta[\cdots, \vec{a}_k, \cdots]$$

*(*det *is multilinear in the matrix columns)*

From this definition, we infer some further properties of the determinant. The matrix $A$ is always understood to have the column vectors $\vec{a}_1, \cdots, \vec{a}_n$.

**Corollary B.25** *Scaling one column of the matrix* $A$ *with factor* $\alpha \in \mathbb{R}$ *will scale the determinant of* $A$ *with the same factor:*

$$\det[\cdots, \alpha\vec{a}_j, \cdots] = \alpha \det[\cdots, \vec{a}_j, \cdots] = \alpha \det A$$

Proof: Consider D3 at position $j$ with $\beta := 0$. $\blacksquare$

**Corollary B.26** *Adding the column* $\vec{a}_k$*, scaled with a factor* $\beta \in \mathbb{R}$*, to the column* $\vec{a}_j$ *(with* $j \neq k$*) does not change the determinant of* $A$*:*

$$\det[\cdots, \vec{a}_j + \beta\vec{a}_k, \cdots, \vec{a}_k, \cdots] = \det[\cdots, \vec{a}_j, \cdots, \vec{a}_k, \cdots] = \det A$$

Proof: Consider D3 at position $j$ with $\alpha := 1$. The left-hand side of the equation equals the sum of $(\det[\cdots, \vec{a}_j, \cdots, \vec{a}_k, \cdots])$ (which is $\det A$) and of $(\beta \det[\cdots, \vec{a}_k, \cdots, \vec{a}_k, \cdots])$. The latter determinant is zero, according to D2. $\blacksquare$

**Corollary B.27** *Swapping two columns* $\vec{a}_j, \vec{a}_k$*,* $j \neq k$ *in* $A = [\vec{a}_1, \cdots, \vec{a}_n]$ *flips the sign of the determinant.*

Proof: We consider the positions $j$ and $k$. First, we add column $j$ to column $k$, which does not change $\det A$ as per the previous corollary:

$$\det A = \det[\cdots, \vec{a}_j, \cdots, \vec{a}_k, \cdots] = \det[\cdots, \vec{a}_j, \cdots, \vec{a}_j + \vec{a}_k, \cdots]$$

We now add the negative of column $k$ to column $j$:

$$\cdots = \det[\cdots, \vec{a}_j - (\vec{a}_j + \vec{a}_k), \cdots, \vec{a}_j + \vec{a}_k, \cdots] = \det[\cdots, -\vec{a}_k, \cdots, \vec{a}_j + \vec{a}_k, \cdots]$$

We add column $j$ to column $k$ and use the first above corollary to extract the factor $(-1)$:

$$\cdots = \det[\cdots, -\vec{a}_k, \cdots, \vec{a}_j, \cdots] = -\det[\cdots, \vec{a}_k, \cdots, \vec{a}_j, \cdots]$$

Multiplying the whole equation with $(-1)$ yields:

$$\det[\cdots, \vec{a}_k, \cdots, \vec{a}_j, \cdots] = -\det[\cdots, \vec{a}_j, \cdots, \vec{a}_k, \cdots] = -\det A \quad \blacksquare$$

**Corollary B.28** *Permuting the columns of $A$ with $\sigma \in S_n$ yields a matrix whose determinant is* $\operatorname{sign}(\sigma) \cdot \det A$.

Proof: Any permutation can be expressed as a sequence of transpositions (cf. claim B.14, p. 85). Each transposition flips the sign of the permutation (cf. corollary B.16, p. 86), and corresponds to one column swap of the matrix. ■

**Corollary B.29** *For any permutation $\sigma \in S_n$, its permutation matrix $P_\sigma$ has the determinant*

$$\det P_\sigma = \operatorname{sign}(\sigma)$$

Proof: The permutation matrix can be obtained by permuting the columns of $\mathbb{1}_n$ with $\sigma$ (recall definition B.21, p. 87). We use the previous corollary and recall that $\det \mathbb{1}_n = 1$ as per D1. ■

We use this corollary, together with D1 and D2 to express an important determinant that we will need later:

**Corollary B.30** *For $j_1, \cdots, j_n \in \{1, \cdots, n\}$, and Euclidean unit vectors $\vec{e}_{j_1}, \vec{e}_{j_2}, \cdots, \vec{e}_{j_n} \in \mathbb{R}^n$:*

$$\det[\vec{e}_{j_1}, \vec{e}_{j_2}, \cdots, \vec{e}_{j_n}] = \begin{cases} \operatorname{sign}(\sigma), & \text{if } \exists \sigma \in S_n : \forall r \in \{1, \cdots, n\} : j_r = \sigma(r) \\ 0, & \text{otherwise, i.e. if at least two of the indices are equal} \end{cases}$$

Proof: The first case was dealt with in the previous corollary because the determinant is $\det P_\sigma$. The second case derives from axiom D2. ■

In order to express the value of the previous corollary's determinant more succinctly, we make use of the *Levi-Civita Symbol* from tensor algebra:

**Definition B.31** *In flat Euclidean $n$-dimensional space, the* totally anti-symmetric tensor *of rank $n$ is denoted by the Levi-Civita symbol $\varepsilon$ and takes the component form:*

$$\varepsilon_{j_1, j_2, \cdots, j_n} := \begin{cases} 1, & \text{if } (j_1, j_2, \cdots, j_n) \text{ is an even permutation of } (1, \cdots, n) \\ -1, & \text{if } (j_1, j_2, \cdots, j_n) \text{ is an odd permutation of } (1, \cdots, n) \\ 0 & \text{otherwise, i.e. if at least two of the indices have the same value} \end{cases}$$

Thus:

$$\det[\vec{e}_{j_1}, \vec{e}_{j_2}, \cdots, \vec{e}_{j_n}] = \varepsilon_{j_1, j_2, \cdots, j_n}$$

**Corollary B.32** *If a column of $A$ is $\vec{0}$, then* $\det A = 0$

Proof: Assume that $\vec{a}_j = \vec{0}$. We consider position $j$. For some $\alpha \in \mathbb{R}$, where $\alpha \neq 0$ and $\alpha \neq 1$, and because $\alpha\vec{0} = \vec{0}$, we obtain:

$$\det A = \det[\cdots, \vec{0}, \cdots] = \det[\cdots, \alpha\vec{0}, \cdots] = \alpha \det[\cdots, \vec{0}, \cdots] = \alpha \det A$$

Thus, the product $(\det A)(1 - \alpha)$ equals zero, and because $\alpha$ was chosen not to equal 1, the second factor is not zero – therefore, the first one must be. ■

**Corollary B.33** *If the columns of $A$ are linearly dependent, then* $\det A = 0$

Proof: If one of the columns is the zero vector, the determinant vanishes as per the previous corollary. If not, then there is some nontrivial linear combination of the column vectors that yiels $\vec{0}$:

$$\sum_k \alpha_k \vec{a}_k = \vec{0}$$

Some of the $\alpha_k$ may be zero, but this combination contains at least two column vectors with non-zero factors because single non-zero vectors are never linearly dependent. Choose one of those vectors, which is at some position $j$ of the matrix columns. We now scale the above linear combination with the inverse of the factor $\alpha_j$:

$$\vec{0} = \frac{1}{\alpha_j} \vec{0} = \vec{a}_j + \sum_{k \neq j} \frac{\alpha_k}{\alpha_j} \vec{a}_k =: \vec{a}_j + \sum_{k \neq j} \beta_k \vec{a}_k$$

At least one of the $\beta_k$ is non-zero. Now, if we add the columns ($k \neq j$) to the column vector at position $j$, scaled with $\beta_k$, respectively, the determinant of $A$ does not change (cf. corollary B.26, p. 88). But since this yields the zero vector at position $j$, the (unchanged) determinant of $A$ must be zero itself, according to the previous corollary. ■

**Lemma B.34** *If a matrix $A$ is diagonal, i.e. $A_{jk} = 0$ for $j \neq k$, its determinant is the product of all its diagonal elements, i.e.* $\det A = \prod_k A_{kk}$.

Proof: For each column $k$, use corollary B.25 to extract the factor $A_{kk}$, leaving the unit vector $\vec{e}_k$ inside. The result is:

$$\det A = \left( \prod_k A_{kk} \right) \det \mathbb{1}_n = \prod_k A_{kk},$$

where we used D1. ■

This also works if one of the $A_{kk}$ should be zero, i.e. if the column $k$ is the zero vector $\vec{0}$, because we may always write $\vec{0} = 0 \cdot \vec{e}_k$. Alternatively, if $A_{kk}$ is zero, $A$ contains a zero vector column and therefore has determinant zero; also, the product of all diagonal elements must be zero.

**Lemma B.35** *If a matrix $A$ is upper-triangular, i.e. if $A_{jk} = 0$ for $j > k$, the determinant of $A$ is just the product of all its diagonal elements. The same holds for lower-triangular matrices.*

Proof: The matrix has the following structure:

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & \cdots & A_{1n} \\ 0 & A_{22} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & A_{(n-1),n} \\ 0 & \cdots & \cdots & 0 & A_{nn} \end{pmatrix}$$

If $A_{11} = 0$, the determinant will be zero (because $A$ contains a null vector column), and so will the product of all the diagonal elements be – in this case, we may stop.

Otherwise, we use corollary B.25 (p. 88) to extract the (non-zero) factor $A_{11}$; the first column then is $\vec{e}_1$:

$$\det A = A_{11} \cdot \det \begin{pmatrix} 1 & A_{12} & \cdots & \cdots & A_{1n} \\ 0 & A_{22} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & A_{(n-1),n} \\ 0 & \cdots & \cdots & 0 & A_{nn} \end{pmatrix}$$

Now, the determinant will not change if we add the first column to all columns to the right, scaled with factors $(-A_{12}), \cdots, (-A_{1n})$. This changes nothing below the first line (only zeros are added there), and produces zeros in the first line in all those columns:

$$\det A = A_{11} \cdot \det \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & A_{22} & A_{23} & \cdots & A_{2n} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & A_{(n-1),n} \\ 0 & \cdots & \cdots & 0 & A_{nn} \end{pmatrix}$$

From here, we can proceed recursively: extract factor $A_{22}$ (if it is not zero, in which case the calculation would stop), (making the second column $\vec{e}_2$), eliminate the rest of the second row, etc. In the last step, only the factor $A_{nn}$ needs to be extracted. Then we have

$$\det A = \left( \prod_k A_{kk} \right) \det \mathbb{1}_n = \prod_k A_{kk},$$

exactly as in the previous lemma.

For lower-triangular matrices, we can proceed in a similar way, starting with the last column, and working our way leftwards. This extracts the diagonal elements in reverse order but leaves a unit matrix inside the determinant, too. ■

**Lemma B.36** *The determinant is invariant under matrix transposition:* $\det A^T = \det A$.

The proof will occur as a side-effect when we introduce Leibniz's formula (cf. the next subsection for that). However, anticipating that proof, we would already like to state here that all of the above remarks regarding columns of $A$ hold for the rows of $A$ as well, as the rows of $A$ are the columns of $A^T$.

**Claim B.37** *If the determinant of a matrix $A \in \mathbb{R}^{n\times n}$ vanishes, its columns are linearly dependent. The same holds for its rows.*

Proof: First, we observe that if $A$ contains a column vector $\vec{0}$, the columns are linearly dependent, and we are done. For the following, we assume that $A$ has no zero column.

Secondly, since $\det A = 0$, we may freely permute the columns and scale them with arbitrary non-zero factors, which, along with the permutation sign, are absorbed into the product value of zero.

Our aim now is to transform $A$ into a lower-triangular form, using only operations which leave the (zero!) determinant invariant.

We now select one of the columns whose first component is *not* zero, and move it into first position by a swap (if it is not there already). We scale the column, dividing it by its first component, and thus obtain an element 1 in the first column. We now proceed to eliminate all the remaining non-zero first components of the other column vectors, by adding the appropriately scaled first column vector.

From the columns currently in positions $\{2, \cdots, n\}$, we proceed in the same way: We select a column with non-zero second component (its first will be zero after the above), swap it into position 2 and then eliminate all the remaining non-zero second components of row 2 to the right; etc.

If we did not find a column with non-zero second component in the second step, the second diagonal element and all elements to its right-hand side are already zero, and we skip this step.

Eventually, this leads to a lower-triangular matrix; we note that none of those column operations have changed the value of the determinant, as stated above, and all are invertible.

Since the determinant of the lower-triangular matrix is still zero, and it must equal the product of that matrix's diagonal elements (see the lemma before the previous one), at least one of those must be zero, and correspond to a skipped step from above. The other diagonal elements now have value 1.

We use the non-vanishing diagonal elements to eliminate all the non-zero elements to their lefts, starting from the rightmost.

The resulting matrix still has determinant zero, and consists of two types of rows: Either (type (a)) a row has a 1 on the diagonal position, and zeros otherwise. Or the row has (type (b)) a zero on the diagonal position and zeros to the right (in all columns with higher index). The elements to the left of the diagonal position are unspecified but may still be non-zero. At least one such row must exist.

We now pick the rightmost column with a zero on its diagonal position. All components above the diagonal position are zero because the matrix is lower-triangular. But all components below must be zero, too, because the columns with higher index correspond to rows of type (a). Therefore, this column is a null vector.

Because all the above operations can be expressed as linear combinations, we have shown that the columns of the original matrix $A$ (with vanishing determinant) are linearly dependent: the zero vector can be expressed as a nontrivial linear combination of its columns.

Due to the previous lemma, and because the determinant of $A^T$ is also zero, the same method can be used to show that the rows of $A$ are linearly dependent. ■

**Corollary B.38** *A matrix $A \in \mathbb{R}^{n\times n}$ has vanishing determinant if and only if its columns (and rows) are linearly dependent.*

Proof: Combine the previous claim and corollary B.33, which express the two directions of implication in this statement. ■

The matrices with vanishing determinants, therefore, are exactly the matrices with non-full rank, i.e. with linearly dependent column/row vectors, and (as hinted at in subsection A.3.11, p. 78) thus the matrices which do *not* possess an inverse:

**Corollary B.39** *A matrix $A \in \mathbb{R}^{n\times n}$ is invertible if and only if its determinant satisfies* $\det A \neq 0$

If we recall subsection A.3.11 on matrix inversion from the previous chapter, we observe that, while the connection between linearly dependent matrix rows and the existence of an inverse still has

a gap (as indicated there), we have now provided an a posteriori justification for using Gaussian elimination to calculate the inverse matrix: the elimination operations (on rows) are exactly those that we used in the proof of the above claim (on columns). Since the existence of the inverse does not depend on the sign or scale of the determinant, scaling or permuting rows is permitted when employing Gaussian elimination.

### B.2.2 Leibniz's Formula

We develop a method for calculating the determinant of a square matrix $A \in \mathbb{R}^{n\times n}$, using only Weierstraß's axioms and their immediate corollaries; this is an adaptation of the proof in [FS20]. After that, we rewrite the formula in a more common way, and prove en route that determinants are invariant under transposition. In a later subsection we will re-formulate Leibniz's formula as a recursive operation: The Laplace expansion.

**Theorem B.40** *(Leibniz Formula) The determinant of a matrix $A \in \mathbb{R}^{n\times n}$ is given by*

$$\det A = \sum_{\sigma\in S_n} \operatorname{sign}(\sigma)\cdot A_{1,\sigma(1)}\cdot A_{2,\sigma(2)}\cdot \cdots \cdot A_{n,\sigma(n)}$$

Before the proof, we note that this formula can be equivalently expressed as a full sum with $n^n$ parts, using the Levi-Civita symbol (cf. definition B.31, p. 89), per:

**Corollary B.41** *The determinant of a matrix $A \in \mathbb{R}^{n\times n}$ is given by*

$$\det A = \sum_{j_1,j_2,\cdots,j_n} \varepsilon_{j_1,j_2,\cdots,j_n}\cdot A_{1,j_1}\cdot A_{2,j_2}\cdot \cdots \cdot A_{n,j_n} =: \sum_{\vec{\jmath}} \varepsilon_{\vec{\jmath}}\cdot A_{1,j_1}\cdot A_{2,j_2}\cdot \cdots \cdot A_{n,j_n}$$

*with the shorthand $\vec{\jmath} := (j_1, j_2, \cdots, j_n)$.*

Proof of the corollary: Where $\vec{\jmath}$ corresponds to a permutation of $(1, 2, \cdots, n)$, the Levi-Civita symbol yields the sign of that permutation. All other parts of the multi-sum are zero; thus, only (and all) the permutations contribute to $\det A$, as in the original Leibniz formula. ■

Now for the proof of the claim: We start from the familiar column notation:

$$A = [\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_n]$$

Now we employ a technique called "expansion along the first column" (which we will further elaborate upon when discussing the Laplace expansion in the next subsection). For this, we express $\vec{a}_1$ in the Euclidean unit vectors:

$$\det A = \det[\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_n] = \det\left[\left(\sum_{j_1}(\vec{a}_1)_{j_1}\vec{e}_{j_1}\right), \vec{a}_2, \cdots, \vec{a}_n\right]$$

We plug in the corresponding matrix components and employ the multi-linearity demanded in axiom D3:

$$\cdots = \det\left[\left(\sum_{j_1} A_{j_1,1}\vec{e}_{j_1}\right), \vec{a}_2, \cdots, \vec{a}_n\right] = \sum_{j_1} A_{j_1,1}\det[\vec{e}_{j_1}, \vec{a}_2, \cdots, \vec{a}_n]$$

We now expand along the other columns, too, and obtain a multi-sum with $n^n$ parts:

$$\cdots = \sum_{j_1,j_2,\cdots,j_n} A_{j_1,1}\cdot A_{j_2,2}\cdot \cdots \cdot A_{j_n,n}\cdot \det[\vec{e}_{j_1}, \vec{e}_{j_2}, \cdots, \vec{e}_{j_n}]$$

Of the $n^n$ parts, "only" $n!$ remain because the determinant will vanish for any pair of equal indices $j_r = j_s$ due to axiom D2. The only remaining contributions are those, where any pairs of indices are unequal. But this means that the $j_1, j_2, \cdots, j_n$ are a permutation of $(1, 2, \cdots, n)$.

Thus, we effectively can sum not over all index values but over the $n!$ permutations in $S_n$:

$$\cdots = \sum_{\sigma\in S_n} A_{\sigma(1),1}\cdot A_{\sigma(2),2}\cdot \cdots \cdot A_{\sigma(n),n}\cdot \det[\vec{e}_{\sigma(1)}, \vec{e}_{\sigma(2)}, \cdots, \vec{e}_{\sigma(n)}]$$

But this determinant of permuted unit vectors is just the sign of $\sigma$, as proven above:

$$\cdots = \sum_{\sigma\in S_n} \operatorname{sign}(\sigma)\cdot A_{\sigma(1),1}\cdot A_{\sigma(2),2}\cdot \cdots \cdot A_{\sigma(n),n}$$

This looks almost like the Leibniz formula but is in fact (as compared to the claimed statement) $\det A^T$. If we can demonstrate that the two expressions are equal, we will also have proved that $\det A^T = \det A$.

To that end, we observe that the product of $n$ matrix elements contains all the numbers from 1 to $n$ both as right-hand indices and as left-hand indices. We recall that permutations form a group, so there is a uniquely defined inverse of $\sigma$. If we pick the element $A_{\sigma(j),j}$ with some value $\sigma(j) = k$, then $j = \sigma^{-1}(k)$, and the same matrix element can be written as $A_{k,\sigma^{-1}(k)}$. Now, $k$ may not be equal to $j$, but if we rearrange the matrix elements to reflect that (which does not change their product), we still obtain:

$$\cdots = \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) \cdot A_{1,\sigma^{-1}(1)} \cdot A_{2,\sigma^{-1}(2)} \cdot \ \cdots \ \cdot A_{n,\sigma^{-1}(n)}$$

We use our results from the section on permutations to replace $\operatorname{sign}(\sigma)$ by $\operatorname{sign}(\sigma^{-1})$:

$$\cdots = \sum_{\sigma \in S_n} \operatorname{sign}(\sigma^{-1}) \cdot A_{1,\sigma^{-1}(1)} \cdot A_{2,\sigma^{-1}(2)} \cdot \ \cdots \ \cdot A_{n,\sigma^{-1}(n)}$$

Now, since the sum is evaluated over all possible permutations in $S_n$, and because $\sigma$ is a bijective map, we may as well sum over all the $\sigma^{-1}$ without changing anything: each permutation is also the inverse a permutation, namely of its own inverse:

$$\cdots = \sum_{\sigma^{-1} \in S_n} \operatorname{sign}(\sigma^{-1}) \cdot A_{1,\sigma^{-1}(1)} \cdot A_{2,\sigma^{-1}(2)} \cdot \ \cdots \ \cdot A_{n,\sigma^{-1}(n)}$$

And since there is no qualitative difference between a permutation and an inverse permutation, we may now replace $\sigma^{-1}$ by $\sigma$ in the whole expression, to obtain:

$$\det A = \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) \cdot A_{1,\sigma(1)} \cdot A_{2,\sigma(2)} \cdot \ \cdots \ \cdot A_{n,\sigma(n)}$$

While we have shown that the determinant of $A$ must have the stated form, it is not technically clear that any expression as stated in Leibniz's formula always is a determinant. For this, we still have to show that the formula satisfies the axioms D1,2,3. While this is not hard in the cases D1 and D3, the alternating quality in D2 is more involved to show. We refer the interested reader to the proof in [FS20] (theorem 4.2.5) for an alternative proof of the D2 case.

D1: Given $A = \mathbb{1}_n$, we observe that $A_{j,\sigma(j)} = (\mathbb{1}_n)_{j,\sigma(j)} = \delta_{j,\sigma(j)}$. There is exactly one permutation in $S_n$ for which every of those Kronecker deltas is 1, namely the identical permutation $\mathrm{id}_n$. For all other permutations, some of the deltas will yield zero. Leibniz's formula thus yields a single contribution with value $\operatorname{sign}(\mathrm{id}_n)$, which is 1 as per definition B.15, SP1 (p. 86).

D3: Given a matrix $A = [\vec{a}_1, \cdots, \vec{a}_n] \in \mathbb{R}^{n \times n}$, we consider the column at some position $r$, where we put the scaled columns $\alpha\vec{a}_j + \beta\vec{a}_k$. Here, we will use Leibniz's formula in the transposed form where the permutation operates on the row indices instead of the column indices, and thus, using $A_{\sigma(r),r} = (\vec{a}_r)_{\sigma(r)}$:

$$\det[\cdots, \alpha\vec{a}_j + \beta\vec{a}_k, \cdots] = \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) \cdot (\cdots)(\alpha\vec{a}_j + \beta\vec{a}_k)_{\sigma(r)}(\cdots)$$

Using distributive laws, every part of the sum splits into two, yielding

$$\cdots = \alpha \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) \cdot (\cdots)(\vec{a}_j)_{\sigma(r)}(\cdots) + \beta \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) \cdot (\cdots)(\vec{a}_k)_{\sigma(r)}(\cdots)$$

But this is exactly Leibniz's formula for

$$\cdots = \alpha \det[\cdots, \vec{a}_j, \cdots] + \beta \det[\cdots, \vec{a}_k, \cdots]$$

D2: We use Leibniz's formula in the transposed form and interchange notation between the full multi-sum with Levi-Civita symbol and the sum over permutations in $S_n$. We assume a given matrix with equal columns $\vec{a}_r = \vec{a}_s$ $(r \neq s)$ and consider only the positions $r$ and $s$:

$$\det A = \sum_{\sigma \in S_n} \operatorname{sign}\sigma \cdot (\cdots) \cdot A_{\sigma(r),r} \cdot (\cdots) \cdot A_{\sigma(s),s} \cdot (\cdots)$$

Plugging in $\vec{a}_s = \vec{a}_r$:

$$\cdots = \sum_{\sigma \in S_n} \operatorname{sign} \sigma \cdot (\cdots) \cdot A_{\sigma(r),r} \cdot (\cdots) \cdot A_{\sigma(s),r} \cdot (\cdots)$$

We now transpose, via the inverse permutations, moving the $\sigma$ application from the left-hand to the right-hand matrix indices:

$$\cdots = \sum_{\sigma \in S_n} \operatorname{sign} \sigma \cdot (\cdots) \cdot A_{r,\sigma(r)} \cdot (\cdots) \cdot A_{s,\sigma(r)} \cdot (\cdots)$$

We switch to Levi-Civita:

$$\cdots \text{“} = \text{”} \sum_{\vec{\jmath}} \varepsilon_{\vec{\jmath}} \cdot (\cdots) \cdot A_{r,j_r} \cdot (\cdots) \cdot A_{s,j_r} \cdot (\cdots)$$

This is, however, only an intermediary stage because it contains a sum over $j_s$ that is reflected in the $\varepsilon_{\vec{\jmath}}$, but not the following product. We can amend this by introducing a Kronecker delta that fixes $j_s$ to $j_r$:

$$\cdots = \sum_{\vec{\jmath}} \delta_{j_r,j_s} \cdot \varepsilon_{\vec{\jmath}} \cdot (\cdots) \cdot A_{r,j_r} \cdot (\cdots) \cdot A_{s,j_s} \cdot (\cdots)$$

But because

$$\delta_{j_r,j_s} \cdot \varepsilon_{\vec{\jmath}} = \delta_{j_r,j_s} \cdot \varepsilon_{\cdots,j_r,\cdots,j_s,\cdots},$$

the sum over $j_s$ yields $\varepsilon_{\cdots,j_r,\cdots,j_r,\cdots}$, which is always zero; thus, the determinant vanishes.

This completes the proof of Leibniz's formula. ■

For example, we reproduce the rule of Sarrus, namely the determinant of a $3 \times 3$ matrix. For $n = 3$, the even permutations of the cycle $(1, 2, 3)$, taken with a power of zero, one or two – which amounts to $\mathrm{id}_3$, or one or two cyclic shifts of $\mathrm{id}_3$, respectively. The three odd permutations can be obtained by taking the three even ones, and swapping the mappings for their second and third respective arguments:

$$\begin{aligned} & \det \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \\ = \;& A_{11}A_{22}A_{33} + A_{12}A_{23}A_{31} + A_{13}A_{21}A_{32} - A_{11}A_{23}A_{32} - A_{12}A_{21}A_{33} - A_{13}A_{22}A_{31} \\ = \;& A_{11}(A_{22}A_{33} - A_{23}A_{32}) + A_{12}(A_{23}A_{31} - A_{21}A_{33}) + A_{13}(A_{21}A_{32} - A_{22}A_{31}) \end{aligned}$$

One can remember this rule by periodically extending the matrix columns by two in both directions, and then taking the products along the positive diagonals (to the right and downwards) starting with the three elements in the first row of the original matrix for the positive contributions. The negative contributions arise from the products on the negative diagonals (to the left and downwards), as visualized here:

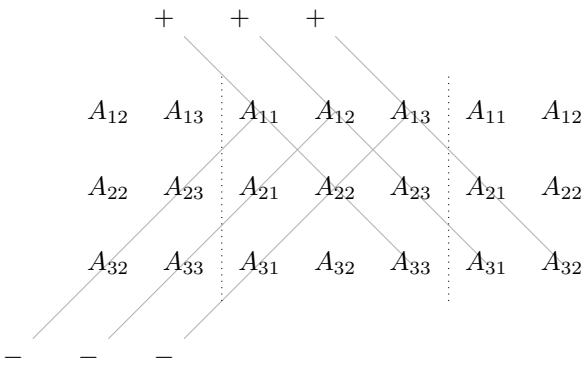


Figure B.1: Sarrus' rule – diagonal scheme

To conclude this subsection, we reiterate that the above proof also contains the proof for lemma B.36 (p. 90):

$$\det A^T = \det A \quad ■$$

### B.2.3 Laplace Expansion

A more practical way of calculating determinants than using Leibniz's formula is a recursive expansion according to Laplace. This is just a different notation of Leibniz, and does not reduce the number of elements that have to be added, but instead of listing all $n!$ permutations of $S_n$, this procedure replaces $\det A$ by a sum of $n$ scaled determinants of matrices from $\mathbb{R}^{(n-1)\times(n-1)}$, and so on. This approach is particularly helpful if $A$ contains lots of zeros. We have used theorem 4.3.2 in [FS20] (p. 222) as basis for our proof.

**Theorem B.42** *(Laplace Expansion) The determinant of a matrix $A \in \mathbb{R}^{n\times n}$ is given by*

$$\det A = \sum_j (-1)^{j+k} A_{jk} \cdot \det a(A,j,k) = \sum_k (-1)^{j+k} A_{jk} \cdot \det a(A,j,k),$$

*where the first equality is called "expansion along column $k$", and the second, "expansion along row $j$". The function $a(A,j,k)$ returns a matrix from $\mathbb{R}^{(n-1)\times(n-1)}$, which is derived from $A$ by cutting out its line $j$ and its column $k$.*

Proof: We recall the initial expansion along the first column in the proof of Leibniz's formula (p. 92). After the first expansion, we have the following (index $j_1$ renamed to $j$ here):

$$\det A = \sum_j A_{j,1} \det[\vec{e}_j, \vec{a}_2, \cdots, \vec{a}_n]$$

We first examine the case $j = 1$, and use Leibniz's formula to evaluate the determinant under the sum:

$$\det[\vec{e}_1, \vec{a}_2, \cdots, \vec{a}_n] = \sum_{\sigma\in S_n} \operatorname{sign}(\sigma)(\vec{e}_1)_{\sigma(1)} \cdot (\vec{a}_2)_{\sigma(2)} \cdot \cdots \cdot (\vec{a}_n)_{\sigma(n)}$$

The first vector component is in fact a Kronecker delta: $(\vec{e}_1)_{\sigma(1)} = \delta_{1,\sigma(1)}$. This fixes the permutations $\sigma$ to compositions $\sigma = (1) \circ \tilde{\sigma}$, where $\tilde{\sigma} : \{2, \cdots, n\} \to \{2, \cdots, n\}$ and $(1)$ is a trivial cycle.

Now, the sign of $\tilde{\sigma}$ is equal to $\operatorname{sign}(\sigma)$ because lemma B.18 (p. 86) that $\operatorname{sign}(\ (1)\ ) = (-1)^{1-1} = 1$.

If we re-labeled the numbers by subtracting 1 of each, $\tilde{\sigma}$ would be a bona fide permutation from $S_{(n-1)}$. Let $\tilde{S}_{(2,\cdots,n)}$ be the group of permutations on $\{2, \cdots, n\}$ (which is isomorphic to $S_{(n-1)}$). In this case:

$$\det[\vec{e}_1, \vec{a}_2, \cdots, \vec{a}_n] = \sum_{\tilde{\sigma}\in\tilde{S}_{(2,\cdots,n)}} \operatorname{sign}(\tilde{\sigma})(\vec{a}_2)_{\tilde{\sigma}(2)} \cdot \cdots \cdot (\vec{a}_n)_{\tilde{\sigma}(n)}$$

But this is, according to Leibniz's formula, just the determinant of the matrix $[\widetilde{\vec{a}_2}, \cdots, \widetilde{\vec{a}_n}]$, where $\widetilde{\vec{a}_k}$ is just $\vec{a}_k$ without its first component. This is because $\tilde{\sigma}$ never can select the first component of $\vec{a}_k$.

Differently put, and using the function $a$ declared in the above claim statement:

$$\det[\vec{e}_1, \vec{a}_2, \cdots, \vec{a}_n] = \det a(A,1,1)$$

For the next part of the sum, $j = 2$, we have to calculate $\det[\vec{e}_2, \vec{a}_2, \cdots, \vec{a}_n]$. But we can reduce this to the operations for case $j = 1$ if we just swap lines 1 and 2. Because of corollary B.27 (p. 88) and lemma B.36 (p. 90), the swap yields a factor of $(-1)$. Thus, we have

$$\det[\vec{e}_2, \vec{a}_2, \cdots, \vec{a}_n] = (-1)\det[\vec{e}_1, \vec{a}'_2, \cdots, \vec{a}'_n],$$

where $\vec{a}'_k$ is derived from $\vec{a}_k$ by swapping its first two components.

From here on, everything works as in case $j = 1$ – we only have to remember that the components $(2, 3, 4, \cdots, n)$ of $\vec{a}'_k$ are the components $(1, 3, 4, \cdots, n)$ of $\vec{a}_k$. If we relate this to the original matrix $A$, we find that, in this case, column 1 and row 2 are cut out, such that

$$\det[\vec{e}_2, \vec{a}_2, \cdots, \vec{a}_n] = (-1)\det a(A,2,1)$$

We now can generalize to an arbitrary $j$. In order to move row $j$ of the matrix to the top without changing the ordering of the rows in between, we need to perform a cycle permutation $(1, 2, \cdots, j)$ on the rows, which amounts to $(j-1)$ swaps and yields a sign factor of $(-1)^{(j-1)}$ (alternatively, cf. lemma B.18, p. 86). Thus, we have:

$$\det[\vec{e}_j, \vec{a}_2, \cdots, \vec{a}_n] = (-1)^{(j-1)}\det a(A,j,1)$$

And the whole determinant of $A$ as expanded along the first column, then, is

$$\det A = \sum_j (-1)^{(j-1)} A_{j,1} \det a(A, j, 1)$$

If we want to expand along column $k$, we can employ the same technique, moving column $k$ into first place with a cycle permutation on the columns, and incurring a sign factor $(-1)^{(k-1)}$ in the process. After that, everything can be done as laid out above, only that the function $a$ now cuts column $k$ out. The product of both sign factors is just $(-1)^{j+k}$:

$$\det A = \sum_j (-1)^{(j+k)} A_{j,k} \det a(A, j, k)$$

Because $\det A^T = \det A$, we may expand along a row in the same way. Now, $j$ is fixed and the sum is over the columns $k$, but other than that, the above formula's symmetry yields the same:

$$\det A = \sum_k (-1)^{(j+k)} A_{j,k} \det a(A, j, k)$$

This concludes our proof. ■

The recursion over all the sub-determinants always terminates because it eventually reaches determinants of matrices with single elements ($a$), for which we use corollary B.25 (p. 88) and axiom D1 in definition B.24 (p. 88): $\det(a) = a \det(1) = a$.

The sign factor $(-1)^{j+k}$ can easily be remembered as a checkered pattern of "+" and "−", starting with positive sign in the upper left-hand corner (coordinates $(1, 1)$).

For example, we may consider $n = 2$, and the matrix

$$A := \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Expanding along the first column yields:

$$\begin{aligned} \det A &= a \cdot \det(d) + (-1) \cdot c \cdot \det(b) \\ &= a \cdot d \cdot \det(1) - c \cdot b \cdot \det(1) \\ &= ad - bc \end{aligned}$$

### B.2.4 Determinant of a Product

**Claim B.43** *For $A, B \in \mathbb{R}^{n \times n}$, the determinant of $AB$ takes the form*

$$\det(AB) = \det(A) \det(B)$$

Proof: To begin with, we recall the determinant of $A$ in the Leibniz formulation with the Levi-Civita symbol (cf. corollary B.41, p. 92). Coming from the right-hand side of the equation, we want to manipulate the expression in a way that will allow us to facilitate the matrix product, which we will need to equate to the left-hand side – for this, we need a full sum over a set of indices, which is readily available in the Levi-Civita notation:

$$\det A = \sum_{\vec{\jmath}} \varepsilon_{\vec{\jmath}} \cdot A_{1,j_1} \cdot A_{2,j_2} \cdot \cdots \cdot A_{n,j_n}$$

We now start on the right-hand side of the claim's equation. For reasons that will become apparent later, we replace $B$ by $B^T$.

$$\begin{aligned} \det(A) \det(B) &= \det(A) \det(B^T) \\ &= \left( \sum_{\vec{\jmath}} \varepsilon_{\vec{\jmath}} \cdot A_{1,j_1} \cdot A_{2,j_2} \cdot \cdots \cdot A_{n,j_n} \right) \cdot \left( \sum_{\vec{k}} \varepsilon_{\vec{k}} \cdot B_{k_1,1} \cdot B_{k_2,2} \cdot \cdots \cdot B_{k_n,n} \right) \\ &= \sum_{\vec{\jmath},\vec{k}} \varepsilon_{\vec{\jmath}} \cdot \varepsilon_{\vec{k}} \cdot A_{1,j_1} \cdot A_{2,j_2} \cdot \cdots \cdot A_{n,j_n} \cdot B_{k_1,1} \cdot B_{k_2,2} \cdot \cdots \cdot B_{k_n,n} \end{aligned}$$

In order to reach the left-hand side of our claimed equation, we need to group the matrix components together so that they evaluate to the various components of $(AB)$. We want to use

the sum over $\vec{\jmath}$ for the matrix multiplications. Therefore we will have to eliminate the sum over $\vec{k}$, and we will want to manipulate the product of the Levi-Civita symbols in a way that connects the indices for the $B$ matrix elements with $\vec{\jmath}$. For any given combination of $\vec{\jmath}, \vec{k}$, we have:

$$\varepsilon_{\vec{\jmath}} \cdot \varepsilon_{\vec{k}} = \begin{cases} 1, & \text{if } \vec{\jmath} \text{ and } \vec{k} \text{ are both even permutations of } (1, 2, \cdots, n), \text{ or both odd permutations} \\ -1, & \text{if one of } \vec{\jmath}, \vec{k} \text{ is an even permutation and the other one an odd permutation} \\ 0, & \text{if at least one of } \vec{\jmath}, \vec{k} \text{ is not a permutation of } (1, 2, \cdots, n) \end{cases}$$

Now, for a given $\vec{\jmath}$ and $\vec{k}$, there is exactly one permutation that maps $\vec{\jmath}$ to $\vec{k}$, if both are permutations of $(1, 2, \cdots, n)$. Also, there is exactly one inverse permutation mapping $\vec{k}$ to $\vec{\jmath}$; we call this permutation $\sigma_{\vec{k}\to\vec{\jmath}}$. Both these permutations have the same sign.

If $\vec{\jmath}$ and $\vec{k}$ are both bona fide permutations, then the product $\varepsilon_{\vec{\jmath}} \cdot \varepsilon_{\vec{k}}$ evaluates exactly to the sign of $\sigma_{\vec{k}\to\vec{\jmath}}$. If both are odd or both are even, the permutation in between must be even. If one is odd and one is even, the permutation in between must be odd.

So, if $\varepsilon_{\vec{\jmath}} \cdot \varepsilon_{\vec{k}}$ is not zero, we can express it as $\text{sign}(\sigma_{\vec{k}\to\vec{\jmath}})$.

If $\vec{\jmath}, \vec{k}$ are not both permutations from $S_n$, there cannot be a permutation from $S_n$ mapping them on each other, either (for instance, if two of $\vec{k}$'s components were equal, then at least one other number in $\{1, 2, \cdots, n\}$ would not be mapped at all because there would be $(n-1)$ remaining numbers but only $(n-2)$ remaining components to be mapped). If they are, the permutation is unique. We may express this by the following statement:

$$\varepsilon_{\vec{\jmath}} \cdot \varepsilon_{\vec{k}} = \sum_{\sigma \in S_n} \text{sign}(\sigma) \cdot \delta_{\vec{\jmath}, \sigma(\vec{k})} = \sum_{\sigma \in S_n} \text{sign}(\sigma) \cdot \delta_{j_1, \sigma(k_1)} \cdot \delta_{j_2, \sigma(k_2)} \cdot \ \cdots \ \delta_{j_n, \sigma(k_n)}$$

We examine every possible permutation. If (and only if) $\sigma$ maps $\vec{k}$ to $\vec{\jmath}$, then all the Kronecker deltas evaluate to 1, and the sign of $\sigma$ has the correct value because $\sigma = \sigma_{\vec{k}\to\vec{\jmath}}$. All other permutations will cause at least one of the Kronecker deltas to be zero, so there can be only one contribution to the sum. If there is no permutation between $\vec{\jmath}$ and $\vec{k}$, then none of the $\sigma$ can make all the Kronecker deltas evaluate to 1; thus, the sum is zero. This happens exactly when at least one of $\vec{\jmath}, \vec{k}$ is not a permutation of $(1, 2, \cdots, n)$.

It turns out that this expression for the product $\varepsilon_{\vec{\jmath}} \cdot \varepsilon_{\vec{k}}$ satisfies all our initial demands. The $n$ Kronecker deltas can be used to eliminate one of the two multi-sums, and the Levi-Civita notation is replaced by a permutation notation for the Leibniz formula. Because we stated above that we intend to eliminate the sum over $\vec{k}$, we re-formulate equivalently:

$$\varepsilon_{\vec{\jmath}} \cdot \varepsilon_{\vec{k}} = \sum_{\sigma \in S_n} \text{sign}(\sigma) \cdot \delta_{\sigma^{-1}(\vec{\jmath}), \vec{k}}$$

We plug this identity in our equation from above:

$$\det(A) \det(B) = \sum_{\vec{\jmath}, \vec{k}} \sum_{\sigma \in S_n} \text{sign}(\sigma) \cdot \delta_{\sigma^{-1}(\vec{\jmath}), \vec{k}} \cdot A_{1,j_1} \cdot A_{2,j_2} \cdot \ \cdots \ \cdot A_{n,j_n} \cdot B_{k_1,1} \cdot B_{k_2,2} \cdot \ \cdots \ \cdot B_{k_n,n}$$

We evaluate the multi-sum over $\vec{k}$, which fixes $\vec{k}$ to $\sigma^{-1}(\vec{\jmath})$:

$$\cdots = \sum_{\vec{\jmath}} \sum_{\sigma \in S_n} \text{sign}(\sigma) \cdot A_{1,j_1} \cdot A_{2,j_2} \cdot \ \cdots \ \cdot A_{n,j_n} \cdot B_{\sigma^{-1}(j_1),1} \cdot B_{\sigma^{-1}(j_2),2} \cdot \ \cdots \ \cdot B_{\sigma^{-1}(j_n),n}$$

Since we sum over all permutations in $S_n$, and because of $\text{sign}(\sigma^{-1}) = \text{sign}(\sigma)$, we may employ the same switching we used for the Leibniz formula, and obtain:

$$\cdots = \sum_{\vec{\jmath}} \sum_{\sigma \in S_n} \text{sign}(\sigma) \cdot A_{1,j_1} \cdot A_{2,j_2} \cdot \ \cdots \ \cdot A_{n,j_n} \cdot B_{j_1,\sigma(1)} \cdot B_{j_2,\sigma(2)} \cdot \ \cdots \ \cdot B_{j_n,\sigma(n)}$$

We observe that we can now rearrange the matrix components to highlight the matrix multiplication:

$$\cdots = \sum_{\vec{\jmath}} \sum_{\sigma \in S_n} \text{sign}(\sigma) \cdot A_{1,j_1} \cdot B_{j_1,\sigma(1)} \cdot A_{2,j_2} \cdot B_{j_2,\sigma(2)} \cdot \ \cdots \ \cdot A_{n,j_n} \cdot B_{j_n,\sigma(n)}$$

We execute the sum over $\vec{\jmath}$:

$$\cdots = \sum_{\sigma \in S_n} \text{sign}(\sigma) \cdot (AB)_{1,\sigma(1)} \cdot (AB)_{2,\sigma(2)} \cdot \ \cdots \ \cdot (AB)_{n,\sigma(n)}$$

But this is exactly what Leibniz's formula yields for the left-hand side of our claim. ■

### B.2.5 Further Properties of Determinants

Here we mention several properties of determinants that will be needed later on, or in the main part of this work.

**Corollary B.44** *If (and only if) a matrix $A \in \mathbb{R}^{n\times n}$ is invertible, determinant of its inverse is given per*

$$\det(A^{-1}) = \frac{1}{\det A}$$

*In particular,* $\det A$ *cannot be zero.*

Proof: Use the determinant product formula B.43 from the preceding subsection, and axiom D1 of definition B.24 (p. 88):

$$1 = \det(\mathbb{1}_n) = \det(A^{-1} \cdot A) = \det(A^{-1}) \cdot \det(A) \quad \blacksquare$$

**Corollary B.45** *If a matrix $A \in \mathbb{R}^{n\times n}$ is orthogonal, its determinant is $\pm 1$.*

Proof: Since $A^T A = \mathbb{1}_n$ as per definition A.27 (p. 76), we may employ the product formula B.43. Because the transposed matrix has the same determinant (cf. lemma B.36, p. 90):

$$1 = \det(\mathbb{1}_n) = \det(A^T \cdot A) = \det(A^T) \cdot \det(A) = (\det(A))^2 \quad \blacksquare$$

**Corollary B.46** *If a matrix $\tilde{A} \in \mathbb{R}^{n\times n}$ is similar to a matrix $A \in \mathbb{R}^{n\times n}$, then* $\det \tilde{A} = \det A$.

Proof: We recall definition A.28 (p. 79) for similar matrices to write $\tilde{A} = S^{-1}AS$ with some invertible matrix $S$. The product formula B.43 yields:

$$\det \tilde{A} = \det(S^{-1}) \cdot \det(A) \cdot \det(S) = \det(A) \cdot [\det(S^{-1}) \cdot \det(S)]$$

Now we can use the corollary B.44:

$$\cdots = \det(A) \cdot \left[\frac{1}{\det(S)} \cdot \det(S)\right] = \det(A) \quad \blacksquare$$

**Corollary B.47** *If a matrix $\tilde{A} \in \mathbb{R}^{n\times n}$ is a permutation of a matrix $A \in \mathbb{R}^{n\times n}$ with some coordinate permutation $\sigma \in S_n$, then* $\det \tilde{A} = \det A$.

Proof: As per lemma B.23 (p. 87), $\tilde{A} = P_\sigma^T A P_\sigma$. But the permutation matrix $P_\sigma$ is orthogonal (cf. lemma B.22, p. 87), thus $P_\sigma^T = P_\sigma^{-1}$. Therefore $\tilde{A}$ and $A$ are similar, and we may use the corollary B.46. $\blacksquare$

**Lemma B.48** *Let $M \in \mathbb{R}^{n\times n}$ be block-upper-triangular matrix with square matrices $A \in \mathbb{R}^{j\times j}$ and $B \in \mathbb{R}^{(n-j)\times(n-j)}$, i.e.,*

$$M = \left(\begin{array}{c|c} \begin{matrix}1 & & j\\ & A & \\ j & & \end{matrix} & \begin{matrix}j{+}1 & & n\\ & C & \\ & & \end{matrix} \\ \hline \begin{matrix}j{+}1 & & \\ & 0 & \\ & & \\ n & & \end{matrix} & \begin{matrix} & & \\ & B & \\ & & \end{matrix} \end{array}\right)$$

*The matrices $A, B$ may be block-upper-triangular themselves. Then:* $\det M = (\det A)(\det B)$. *The same holds for block-lower-triangular matrices and for block-diagonal matrices.*

Proof: We use Leibniz's formula B.40 (p. 92):

$$\det M = \sum_{\sigma\in S_n} \operatorname{sign}(\sigma) \cdot M_{1,\sigma(1)} \cdot \cdots \cdot M_{j,\sigma(j)} \cdot M_{(j+1),\sigma(j+1)} \cdot \cdots \cdot M_{n,\sigma(n)}$$

Now, if for any $k \in \{j+1, \cdots, n\}$ the value $\sigma(k)$ were in $\{1, \cdots, j\}$, the contribution to the sum would be zero, because $M_{k,\sigma(k)} = 0$. So we only need consider permutations $\sigma$ where $\{j+1, \cdots, n\}$ is mapped on itself.

This immediately implies that $\{1, \cdots, j\}$ must be mapped on itself, too, because permutations are bijective maps. Thus, we may compose $\sigma$ of two permutations $\sigma = \sigma_A \circ \sigma_B$, where $\sigma_A$ operates non-trivially only on $\{1, \cdots, j\}$ and $\sigma_B$ on $\{j+1, \cdots, n\}$, respectively:

$$\forall r \in \{1, \cdots, j\}, s \in \{j+1, \cdots, n\} : \sigma_A(s) = s \wedge \sigma_B(r) = r$$

We use corollary B.20 (p. 87) to factorize $\operatorname{sign}(\sigma) = \operatorname{sign}(\sigma_A) \cdot \operatorname{sign}(\sigma_B)$, and rearrange, using the respective permutations that can yield non-zero contributions:

$$\det M = \sum_{\sigma_A \circ \sigma_B \in S_n} \left[\operatorname{sign}(\sigma_A) \cdot M_{1,\sigma_A(1)} \cdot \cdots \cdot M_{j,\sigma_A(j)}\right] \cdot \left[\operatorname{sign}(\sigma_B) \cdot M_{(j+1),\sigma_B(j+1)} \cdot \cdots \cdot M_{n,\sigma_B(n)}\right]$$

We may now restrict $\sigma_A, \sigma_B$ to their respective non-trivial sets of numbers without changing the overall result. In a similar way our proof of Laplace's expansion (claim B.42, p. 95), we write

$$\tilde{\sigma}_A := \sigma_A\Big|_{\{1,\cdots,j\}} \in S_j, \qquad \tilde{\sigma}_B := \sigma_B\Big|_{\{j+1,\cdots,n\}} \in \tilde{S}_{(j+1,\cdots,n)}$$

We then split the sum and apply Leibniz's formula B.40 (p. 92):

$$\begin{aligned}
\det M &= \left( \sum_{\tilde{\sigma}_A \in S_j} \operatorname{sign}(\tilde{\sigma}_A) \cdot M_{1,\tilde{\sigma}_A(1)} \cdot \cdots \cdot M_{j,\tilde{\sigma}_A(j)} \right) \\
&\quad \cdot \left( \sum_{\tilde{\sigma}_B \in \tilde{S}_{(j+1,\cdots,n)}} \operatorname{sign}(\tilde{\sigma}_B) \cdot M_{(j+1),\tilde{\sigma}_B(j+1)} \cdot \cdots \cdot M_{n,\tilde{\sigma}_B(n)} \right) \\
&= (\det A) \cdot (\det B)
\end{aligned}$$

Since this did not depend on the inner structure of $A, B$, we may apply the above recursively, if $A$ or $B$ are block-upper-triangular themselves.

Since the elements of the sub-matrix $C$ were never used, the formula is also correct if all of $C$'s components are zero, i.e. if $M$ is block-diagonal.

And because matrix transposition does not affect the determinant value, the formula also holds for block-lower-triangular matrices. ■

(If applied recursively, this formula also reproduces lemma B.35 (p. 90).)

# Appendix C

# The Eigenvalue Problem

For a given square matrix $A \in \mathbb{C}^{n\times n}$, there are some vectors that do not change direction[1] when multiplied with $A$, but are scaled with a factor $\lambda \in \mathbb{C}$. While we will deal only with real matrices, it is not clear a priori that all the possible scaling factors are real, too. In fact, rotation matrices in $\mathbb{R}^2$ and $\mathbb{R}^3$ do each feature two scaling factors $e^{\pm i\alpha}$ – the associated vectors are complex, too, because a matrix with real components could not create imaginary parts for a real vector's components by multiplication.

We will start by stating the problem formally and examining some general properties. In a second section, we deal with the special case of symmetric matrices. The proofs for the spectral theorem and the linear independence of certain eigenvectors can found in [FS20] and other linear algebra textbooks.

## C.1 General Properties

**Definition C.1** *(Eigenvalue Problem) For a given square matrix $A \in \mathbb{C}^{n\times n}$, a number $\lambda \in \mathbb{C}$ is called* eigenvalue *of $A$, if there is a vector $\vec{v} \in \mathbb{C}^n$, $\vec{v} \neq \vec{0}$, such that*

$$A \cdot \vec{v} = \lambda \cdot \vec{v}$$

*This vector is called* eigenvector *of $A$ to the eigenvalue $\lambda$.*

**Corollary C.2** *For an eigenvector $\vec{v}$ to the eigenvalue $\lambda$ of a matrix $A$, any vector $c\vec{v}$, $c \in \mathbb{C}$, is also an eigenvector to the same eigenvalue.*

Proof: Matrix multiplication is linear; thus, $A \cdot (c\vec{v}) = c \cdot (A\vec{v}) = c(\lambda\vec{v}) = \lambda \cdot (c\vec{v})$. ■

Thus, the unique property of an eigenvector always is its direction; its length may be chosen arbitrarily.

We now proceed to find non-trivial solutions for the eigenvalue problem. First, we may rewrite the equation, using $\mathbb{1}_n\vec{v} = \vec{v}$:

$$A\vec{v} = \lambda\vec{v} \qquad \Leftrightarrow \qquad (A - \lambda\mathbb{1}_n)\vec{v} = \vec{0}$$

Now, if $M_\lambda := A - \lambda\mathbb{1}_n$ were invertible, we might solve for $\vec{v}$, and obtain:

$$\vec{v} = \mathbb{1}_n \cdot \vec{v} = (M_\lambda^{-1} \cdot M_\lambda) \cdot \vec{v} = M_\lambda^{-1} \cdot (M_\lambda\vec{v}) = M_\lambda^{-1} \cdot \vec{0} = \vec{0}$$

Since we stipulated $\vec{v} \neq \vec{0}$, this is no nontrivial solution of our problem. Thus, $\vec{v} \neq \vec{0}$ implies that $M_\lambda$ must *not* be invertible. We refer to corollary B.39 (p. 91) and obtain:

**Corollary C.3** *The solutions to the eigenvalue problem for a given matrix $A \in \mathbb{C}^{n\times n}$ are numbers $\lambda \in \mathbb{C}$, for which*

$$\det(A - \lambda\mathbb{1}_n) = 0$$

**Definition C.4** *The expression $\det(A - \lambda\mathbb{1}_n)$ for a given matrix $A \in \mathbb{C}^{n\times n}$ is a polynomial in $\lambda$, called the* characteristic polynomial *of $A$ and denoted $\chi_A(\lambda)$.*

*The eigenvalues are the roots of this polynomial, i.e. the solutions for $\chi_A(\lambda) = 0$.*

*The set of eigenvalues of a given matrix $A$ is called the* spectrum *of $A$.*

*If an eigenvalue $\lambda$ corresponds to an $k$-fold root of $\chi_A$, it is said to have an* algebraic multiplicity *of $k$.*

[1] They may change to the opposite direction, though, because this equals scaling with a negative real number

**Lemma C.5** *For a matrix $A \in \mathbb{C}^{n\times n}$, the characteristic polynomial $\chi_A$ has $n$ roots, i.e. the sum of algebraic multiplicities of the different eigenvalues equals $n$.*

Proof: Using Leibniz's formula B.40 (p. 92), we observe that, for the permutation $\mathrm{id}_n \in S_n$, the contribution to $\det(M_\lambda)$ is

$$(M_\lambda)_{1,1} \cdot \cdots \cdot (M_\lambda)_{n,n} = (A_{1,1} - \lambda) \cdot \cdots \cdot (A_{n,n} - \lambda)$$

This permutation is the only one for which all the diagonal elements of $M_\lambda$ feature in the contribution to $\det(M_\lambda)$, and thus has the highest number of factors containing $\lambda$. The polynomial $\chi_A$ therefore has degree $n$, and will have $n$ roots in $\mathbb{C}$, according to the Fundamental Theorem of Algebra[2] (cf. [KM21], p. 395).

Therefore, $\chi_A$ has $n$ roots in $\mathbb{C}$. ■

**Definition C.6** *For an eigenvalue $\lambda$ of a matrix $A \in \mathbb{C}^{n\times n}$, the span of its associated eigenvectors is called* eigenspace *of $\lambda$, denoted $E_\lambda$. The dimension of this vector space is called $\lambda$'s* geometric multiplicity*.*

The eigenvectors to $\lambda$ can be determined by solving the homogeneous equation $M_\lambda \vec{v} = \vec{0}$ for this particular $\lambda$, i.e. by finding a basis of $M_\lambda$'s kernel, which can be done by Gaussian elimination. Since $M_\lambda$ is not invertible, the kernel is non-trivial and will therefore contain at least one vector unequal to $\vec{0}$. But, depending on the components of $M_\lambda$, the dimension of $E_\lambda$ might be less than the algebraic multiplicity of $\lambda$ (it cannot exceed that; cf. lemma 3.5.2 in [FS20]).

The eigenspaces of different eigenvalues are, however, linearly independent. We show this by adapting a proof of [FS20], p. 245:

**Theorem C.7** *For any matrix $A \in \mathbb{C}^{n\times n}$, eigenvectors to different eigenvalues are linearly independent.*

Proof: Assume there are $m \leq n$ different eigenvalues $\lambda_1, \cdots, \lambda_m$, and respective eigenvectors $\vec{v}_1, \cdots, \vec{v}_m$. We prove the statement by induction for $k \in \{1, \cdots, m\}$, recalling definition A.17 (p. 73). The case $k = 1$ is trivial because, since $\vec{v}_1$ is an eigenvector, it cannot be the zero vector, thus $\alpha_1 \vec{v}_1 = \vec{0}$ is only possible for $\alpha_1 = 0$ (non-zero single vectors are always linearly independent).

For $1 < k \leq m$, assume that the statement holds for $(k-1)$, such that $\{\vec{v}_1, \cdots, \vec{v}_{k-1}\}$ is a linearly independent set. We now add $\vec{v}_k$ to that set and consider the condition for linear independence:

$$\alpha_1 \vec{v}_1 + \cdots + \alpha_{k-1}\vec{v}_{k-1} + \alpha_k \vec{v}_k = \vec{0} \qquad \Rightarrow \qquad \alpha_1 = \cdots = \alpha_{k-1} = \alpha_k = 0 \tag{*}$$

We multiply the antecedent equation once with $A$ (from the left), and once with $\lambda_k$:

$$\begin{aligned} A \cdot (\cdots) \quad &\rightsquigarrow \quad \alpha_1\lambda_1\vec{v}_1 + \cdots + \alpha_{k-1}\lambda_{k-1}\vec{v}_{k-1} + \alpha_k\lambda_k\vec{v}_k = \vec{0} \\ \lambda_k \cdot (\cdots) \quad &\rightsquigarrow \quad \alpha_1\lambda_k\vec{v}_1 + \cdots + \alpha_{k-1}\lambda_k\vec{v}_{k-1} + \alpha_k\lambda_k\vec{v}_k = \vec{0} \end{aligned}$$

We now subtract both of these equations and obtain:

$$\alpha_1(\lambda_1 - \lambda_k)\vec{v}_1 + \cdots + \alpha_{k-1}(\lambda_{k-1} - \lambda_k)\vec{v}_{k-1} = \vec{0}$$

The vector $\vec{v}_k$ has vanished here because it had identical scaling factors.

Now, since we assumed the $\vec{v}_1, \cdots, \vec{v}_{k-1}$ to be linearly independent in the induction hypothesis, we may use definition A.17 and observe that all the scaling factors of the above vectors must vanish, i.e.,

$$\alpha_1(\lambda_1 - \lambda_k) = \cdots = \alpha_{k-1}(\lambda_{k-1} - \lambda_k) = 0$$

Because the eigenvalues $\lambda_1, \cdots, \lambda_m$ are all different, their differences cannot be zero; but then the respective $\alpha$ factors must be:

$$\alpha_1 = \cdots = \alpha_{k-1} = 0$$

Plugging this result back into the antecedent equation in (*), we obtain

$$\alpha_k \vec{v}_k = 0,$$

but this implies $\alpha_k = 0$, too, like in the induction base case.

Therefore, $\alpha_1 = \cdots = \alpha_k = 0$, which means that the implication in (*) is correct, and thus the set $\{\vec{v}_1, \cdots, \vec{v}_k\}$ is linearly independent, too. ■

To conclude the general observations on eigenvalues, we show that similar matrices are *co-spectral*:

[2] $\mathbb{C}$ is algebraically closed: Any complex polynomial has a complex root. Obtain the $n$ roots by dividing off the roots one by one, using polynomial division. Each resulting polynomial has a lesser degree (by 1) and as per the fundamental theorem, will have a complex root if it has degree $\geq 1$.

**Lemma C.8** *If two square matrices $A, B \in \mathbb{R}^{n\times n}$ are similar via a a matrix $S \in \mathbb{R}^{n\times n}$, they share the same spectrum of eigenvalues.*

Proof: The eigenvalues of $A$ are determined by solving $\det(A - \lambda\mathbb{1}_n) = 0$. Since $A$ and $B$ are similar via $S$, we know that $B = S^{-1}AS$ (cf. definition A.28, p. 79).

We take the matrix $M_\lambda := A - \lambda\mathbb{1}_n$ and observe that

$$S^{-1}M_\lambda S = S^{-1}AS - \lambda S^{-1}\mathbb{1}_n S = B - \lambda S^{-1}S = B - \lambda\mathbb{1}_n$$

Thus, $(B - \lambda\mathbb{1}_n)$ and $(A - \lambda\mathbb{1}_n)$ are similar, and have the same determinant, according to corollary B.46 (p. 98). Thus, if $\lambda$ is an eigenvalue for $A$, it also is one for $B$, and vice versa. ■

## C.2 Symmetric Matrices

For real symmetric matrices, we can show that not only are the various eigenspaces linearly independent, but all eigenspaces have maximum dimension, and the set of eigenvectors is a basis of $\mathbb{R}^n$. Also, all eigenvalues are real numbers. We will show the second implication first, and then adapt a proof in [FS20] (p. 361f) for real symmetric matrices to prove the first implication. The combination of both implications is known as *spectral theorem*.

**Lemma C.9** *Real symmetric matrices have real eigenvalues.*

Proof: We show a more general result, namely that hermitian matrices have real eigenvalues. A matrix $A \in \mathbb{C}^{n\times n}$ is hermitian if $A^\dagger := (A^T)^* = (A^*)^T = A$, where the star denotes complex conjugation. If a hermitian matrix is real, this reduces to $A^T = A$, which is just the symmetry condition for real matrices.

Now, we consider a hermitian matrix $A = A^\dagger \in \mathbb{C}^{n\times n}$ and write the defining equation for the characteristic polynomial:

$$\det(A - \lambda\mathbb{1}_n) = 0 \qquad (*)$$

Because $A = A^\dagger$, we may also write

$$\det(A^\dagger - \lambda\mathbb{1}_n) = 0$$

and infer that if $\lambda$ is an eigenvalue of $A$, it is also an eigenvalue of $A^\dagger$.

We now take the complex conjugate of that equation. Because that operation commutes with addition and multiplication in $\mathbb{C}$, this is equivalent to

$$\det((A^\dagger)^* - \lambda^*\mathbb{1}_n) = 0$$

But $A^\dagger = (A^T)^*$, so that $(A^\dagger)^* = ((A^T)^*)^* = A^T$, and thus

$$\det(A^T - \lambda^*\mathbb{1}_n) = 0$$

Taking the transpose of that yields:

$$\det(A - \lambda^*\mathbb{1}_n) = 0 \qquad (**)$$

We now solve equations (*) and (**) simultaneously and obtain (in the same ordering) solutions

$$\lambda_1 = \cdots, \ \cdots, \lambda_n = \cdots$$
$$\lambda_1^* = \cdots, \ \cdots, \lambda_n^* = \cdots,$$

where the numerical values of each $\lambda_j$ and $\lambda_j^*$ are equal. But then

$$\forall j \in \{1, \cdots, n\} : \lambda_j = \lambda_j^*,$$

which means that all the $\lambda_j$ are real numbers. Since a real symmetric matrix is hermitian, this concludes our proof. ■

**Theorem C.10** *For real symmetric matrices in $\mathbb{R}^{n\times n}$, there is an orthogonal basis of $\mathbb{R}^n$ consisting of normalized eigenvectors.*

Proof: In preparation, we recall subsection A.3.9 (p. 77), where we showed that for a matrix $A \in \mathbb{R}^{n\times n}$ and vectors $\vec{v}, \vec{w} \in \mathbb{R}^n$, $\langle \vec{v}, A\vec{w}\rangle = \langle A^T\vec{v}, \vec{w}\rangle$. If $A$ is also symmetric, this implies:

$$\langle \vec{v}, A\vec{w}\rangle = \langle A\vec{v}, \vec{w}\rangle$$

Let $(\lambda_1, \cdots, \lambda_n)$ be the spectrum of $A$, and $(\vec{v}_1, \cdots, \vec{v}_n)$ a tuple of associated eigenvectors. We now present an iterative method for obtaining an orthogonal set of normalized eigenvectors that shadows the Gram-Schmidt orthonormalization process (cf. theorem 6.5.5 (pp. 342ff) in [FS20]).

We start with the first eigenvector $\vec{v}_1$ and normalize it, scaling it with the square root of $\langle \vec{v}_1, \vec{v}_1\rangle$; this we call $\vec{u}_1$.

Evidently, $\vec{u}_1$ is an eigenvector of $A$ to the eigenvalue $\lambda_1$ (cf. corollary C.2, p. 100).

We now take $\vec{v}_2$, the eigenvector to $\lambda_2$, and project out its component parallel to $\vec{v}_1''$:

$$\vec{v}_2' := \vec{v}_2 - \alpha_1\vec{u}_1$$

We fix $\alpha_1$ by demanding that $\vec{v}_2'$ be orthogonal to $\vec{u}_1$:

$$0 \stackrel{!}{=} \langle \vec{v}_2', \vec{u}_1\rangle = \langle \vec{v}_2, \vec{u}_1\rangle - \alpha_1\langle \vec{u}_1, \vec{u}_1\rangle = \langle \vec{v}_2, \vec{u}_1\rangle - \alpha_1 \qquad \rightsquigarrow \qquad \alpha_1 = \langle \vec{v}_2, \vec{u}_1\rangle$$

Also, we define the ortogonal space fo $\vec{u}_1$:

$$W_1 := \{\, \vec{w} \in \mathbb{R}^n \,|\, \langle \vec{u}_1, \vec{w}\rangle = 0\}$$

But for all vectors $\vec{w} \in W_1$, we can use the scalar product equation from above:

$$\langle \vec{u}_1, A\vec{w}\rangle = \langle A\vec{u}_1, \vec{w}\rangle = \lambda_1\langle \vec{u}_1, \vec{w}\rangle = \lambda_1 \cdot 0 = 0$$

Thus, if $\vec{w}$ is in $W_1$, so is $A\vec{w}$.

Since $\vec{v}_2'$ clearly is in $W_1$, we may infer that $A\vec{v}_2'$ is also in $W_1$. We plug in the above definition for $\vec{v}_2'$ and use the eigenvalue equation for $\vec{v}_2$ and $\vec{u}_1$:

$$A\vec{v}_2' = A(\vec{v}_2 - \alpha_1\vec{u}_1) = \lambda_2\vec{v}_2 - \alpha_1\lambda_1\vec{u}_1 = \lambda_2(\vec{v}_2' + \alpha_1\vec{u}_1) - \alpha_1\lambda_1\vec{u}_1 = \lambda_2\vec{v}_2' + \alpha_1(\lambda_2 - \lambda_1)\vec{u}_1$$

Now, the part parallel to $\vec{u}_1$ must vanish we had already established that $A\vec{v}_2' \in W_1$. Thus:

$$\alpha_1(\lambda_2 - \lambda_1) = 0$$

If $\lambda_2 = \lambda_1$, both $\vec{v}_2$ and $\vec{u}_1$ are eigenvectors from the same eigenspace. Then, $\vec{v}_2'$ is a linear combination inside that eigenspace, it is perpendicular to $\vec{u}_1$, and it is a bona fide eigenvector of $A$ to that eigenvalue.

If, however, $\lambda_2 \neq \lambda_1$, the projection factor $\alpha_1$ must vanish. This means that the eigenvector $\vec{v}_2$ was already perpendicular to $\vec{u}_1$ in the first place, and $\vec{v}_2'$ is in fact equal to $\vec{v}_2$.

In both cases, $A\vec{v}_2' = \lambda_2\vec{v}_2'$, so $\vec{v}_2'$ is an eigenvector of $A$ to $\lambda_2$. We determine $\vec{u}_2$ by scaling $\vec{v}_2'$ with the inverse square root of its Euclidean norm, to obtain unit length.

We proceed accordingly for the next eigenvector (that is, if $n > 2$). From $\vec{v}_3$, we project out the parts along $\vec{u}_1$ and $\vec{u}_2$, so that $\vec{v}_3' \perp \vec{u}_1$ and $\vec{v}_3' \perp \vec{u}_2$. Clearly, $\vec{v}_3'$ is in $W_1$, but we now define

$$W_2 := \{\, \vec{w} \in W_1 \,|\, \langle \vec{u}_2, \vec{w}\rangle = 0\}$$

$W_2$ is an orthogonal subspace of $W_1$. All its elements are orthogonal to $\vec{u}_1$, but also to $\vec{u}_2$. By examining $\langle \vec{u}_2, A\vec{w}\rangle$ for any $\vec{w} \in W_2$, we see that $A\vec{w}$ will also be in $W_2$. Plugging in the definition of $\vec{'}_3$, and using the eigenvalue equation for $\vec{v}_3$, $\vec{u}_1$ and $\vec{u}_2$, we again see that $\vec{v}_3'$ is an eigenvector of A to $\lambda_3$. If $\lambda_3 = \lambda_1$, it belongs to the eigenspace of $\lambda_1$; if $\lambda_3 = \lambda_2$, it belongs to the eigenspace of $\lambda_2$ (which would be redundant, but not wrong, if $\lambda_1 = \lambda_2$). If it is not part of one of those eigenspaces, the projection factor must vanish, and $\vec{v}_3$ was already in $W_2$.

In this way we can construct, one by one, an orthogonal set of normalized eigenvectors for $A$, which concludes our proof. ■

**Corollary C.11** *For any real symmetric matrix $A \in \mathbb{R}^{n\times n}$, the eigenspaces of different eigenvalues are perpendicular to each other.*

Proof: This follows directly from our observations in the previous theorem's proof. If an eigenvector of $A$ does not belong to a certain eigenspace, it is not only linearly independent (by virtue of theorem C.7, p. 101) with that eigenspace, but it is already perpendicular to it, because no components along that space would be projected out in the Gram-Schmidt process as described above. ■

# Appendix D

# Algebra 1: Some Groups

This short chapter is only supposed to illustrate some widely used groups, and to give a graphical representation of the symmetric group $S_n$, which is the permutation group (cf. section B.1, pp. 80ff.).

Dihedral groups $D_n, n \geq 3$ are subgroups of $S_n$, respectively, and are only mentioned because $D_3$ is sometimes used as an example of a symmetric group. For $n = 3$, the symmetric group $S_3$ is in fact equal/isomorphic to $D_3$, but for higher $n$ the symmetric groups contain more elements than the respective $D_n$; thus, introducing dihedral groups may be helpful to avoid confusion.

The first section introduces some matrix groups because matrices feature prominently in this work.

The definitions are taken from [K$^+$88] but could be found in any textbook on linear algebra.

## D.1 Linear Maps / Matrices

**Definition D.1** *The group $GL(n, \mathbb{F})$ is called the* general linear group *and consists of invertible matrices from $\mathbb{F}^{n\times n}$, i.e. matrices with non-vanishing determinants. The group operation is the standard matrix multiplication.*

Usually, the fields considered are $\mathbb{R}$ or $\mathbb{C}$.

The unit matrix $\mathbb{1}_n$ is part of $GL(n, \mathbb{F})$, because it has determinant 1. We already mentioned that the product of two invertible matrices is invertible, too (cf. subsection A.3.11, p. 78) – this means that the group is closed under its operation (axiom G1 of definition A.1, p. 66).

**Definition D.2** *The group $SL(n, \mathbb{F}) \subset GL(n, \mathbb{F})$ is called the* special linear group *consists of matrices from $\mathbb{F}^{n\times n}$ with determinant 1 ("unimodular matrices").*

This group is closed as well, due to the determinant product law B.43 (p. 96).

**Definition D.3** *The group $O(n) \subset GL(n, \mathbb{R})$ is called the* orthogonal group *and consists of the matrices $A$ in $\mathbb{R}^{n\times n}$ that satisfy*

$$A^T A = AA^T = \mathbb{1}_n$$

Recalling the subsection A.3.8 (p. 76) on transposed matrices, we can easily verify that this group is closed. Let $A, B \in O(n)$, then

$$(AB)^T(AB) = B^T A^T AB = B^T B = \mathbb{1}_n; \qquad (AB)(AB)^T = ABB^T A^T = AA^T = \mathbb{1}_n$$

Thus, the product of two orthogonal matrices is itself orthogonal.

We recall from lemma B.22 (p. 87) that the permutation matrices are orthogonal and therefore belong to this group. Also, corollary B.45 (p. 98) tells us that all the matrices in $O(n)$ have determinant $\pm 1$.

**Definition D.4** *The group $SO(n) \subset O(n)$ is called the* special orthogonal group *and contains the matrices $A$ in $O(n)$ satisfying* $\det(A) = +1$.

$O(2)$ and $O(3)$ can be used to represent rotations in $\mathbb{R}^2, \mathbb{R}^3$, respectively.

Note that not every real matrix with determinant 1 belongs to $SO(n)$. Consider

$$A := \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$$

Clearly, $A \in SL(2,\mathbb{R})$, but we easily verify that $A \notin SO(2)$:

$$A^T A = AA^T = \begin{pmatrix} 4 & 0 \\ 0 & \frac{1}{4} \end{pmatrix} \neq \mathbb{1}_2$$

## D.2 Dihedral Groups

The dihedral groups contain symmetry transformations in two dimensions, namely rotations and reflections of regular $n$-gons, relative to their center points. We will omit the cases $n = 1$ and $n = 2$ and define:

**Definition D.5** *For $n \in \mathbb{N}, n \geq 3$, the group $D_n$ consists of the $n$ rotations and $n$ reflections that transform a regular $n$-gon in the two-dimensional plane on itself, with the composition of mappings as its group operation.*

One can imagine a piece of cardboard in the shape of such an $n$-gon with numbered corners centered at the origin of a 2D Cartesian coordinate system. It has a front side and a flip side (hence the name "dihedral", i.e. an object with two faces). Rotations never change the side facing upwards, while reflections flip the cardboard piece. Each reflection is self-inverse, but each composition of two reflections will leave the side facing up unchanged, i.e. amount to one of the rotations.

Rotations commute (in 2D) and constitute an Abelian subgroup of $D_n$. If we initially place the corner 1 at coordinates $(1, 0)$ by default, and number the corners counter-clockwise (front side facing up), there are $n$ rotations transporting corner 1 to the former location of corner $(1 + j)$, $(j \in \{0, 1, \cdots, (n-1)\})$. We name the rotations $r_j$ and identify the respective rotation angles as $j \cdot (2\pi)/n$. Evidently, $r_0$ is the identity operation of the group. Also, we observe that

$$r_j \circ r_k = r_k \circ r_j = r_{(j+k) \bmod n} \qquad \text{and} \qquad (r_j)^{-1} = r_{(n-j) \bmod n}$$

Therefore, the rotations constitute an Abelian subgroup of $D_n$.

As for the reflections, we have to differentiate between even and odd $n$.

- If $n$ is even, then the $n$-gon has $(n/2)$ distinct pairs of parallel edges. The line connecting the midpoints of such a pair splits the $n$-gon in two symmetric halves and therefore corresponds to one reflection operation. The $n$-gon also has $(n/2)$ distinct pairs of opposite vertices (corners). The lines connecting those also correspond to reflection operations.

  For $j \in \{1, \cdots, (n/2)\}$, we define $m_{2j-1}$ as the reflection for the symmetry line through vertex $j$, and $m_{2j}$ as the reflection for the symmetry line through the edge between vertices $j$ and $(j + 1)$.

- If $n$ is odd, the $n$-gon has $n$ distinct pairs of one vertex (corner) and its opposing edge. Connecting the vertex to the midpoint of that edge defines a reflection operation.

  For $j \in \{1, \cdots, n\}$, we define $m_j$ as the reflection for the symmetry line through vertex $j$.

We show the axes of the various reflection operations for $n = 5$ and $n = 6$:

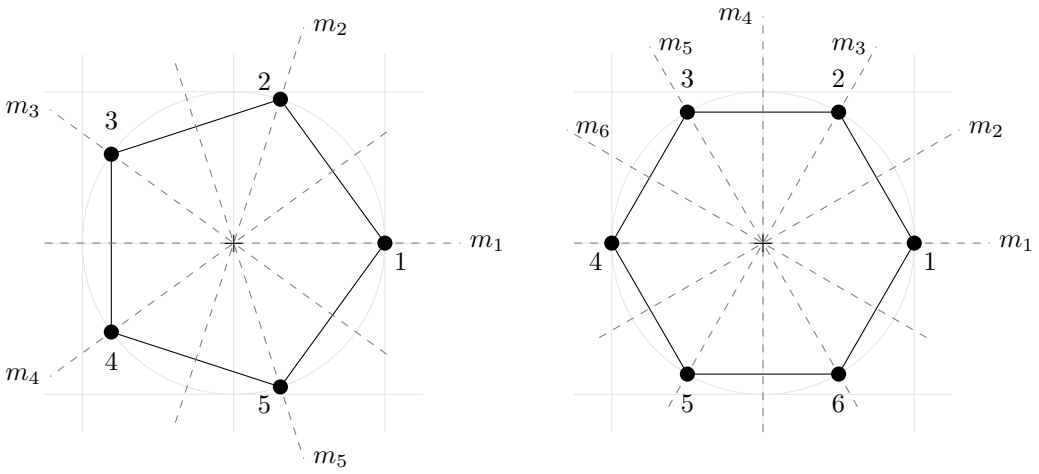


Figure D.1: Reflections in the dihedral groups $D_5$ and $D_6$

## D.3 Symmetric Groups Revisited

We already established that the permutations of $\{1, \cdots, n\}$ constitute the symmetric group $S_n$ (cf. subsection D.3, p. 106).

While the dihedral groups $D_n$ have $2n$ elements ($n$ rotations including id, $n$ reflections), the symmetric groups $S_n$ have $n!$ elements, respectively. Those numbers are equal in case $n = 3$ but not otherwise.

In fact, we may think of all the operations in $D_n$ as permutations of the corner labels in the respective $S_n$ ($n \geq 3$).

But if we follow a cyclic path

$$1 \to 2 \to \cdots \to (n-1) \to n \to 1,$$

this will only ever lead us along the edges of the $D_n$ polygon in one of two directions. However, the permutations in $S_n$ also allow us to create twisted shapes. We do not see any effects of this for the triangle, but for four vertices, there are already three different shapes (times two directions). One could follow the edges of a square (as in $D_4$), but one could also take first one such edge, then a diagonal, then another square edge, and return. Or one could first take a diagonal, then a square edge, then the another diagonal and the remaining square edge:

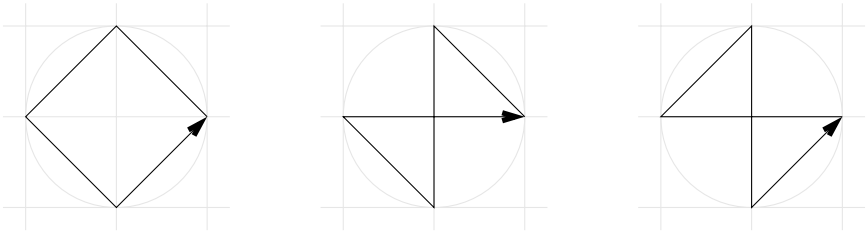

Figure D.2: Three basic path shapes in $S_4$

For $S_5$, there are 12 basic shapes (times two directions): one regular pentagon, one star, and two other shapes in five different flavors (depending on where the path starts). We show one of each kind:

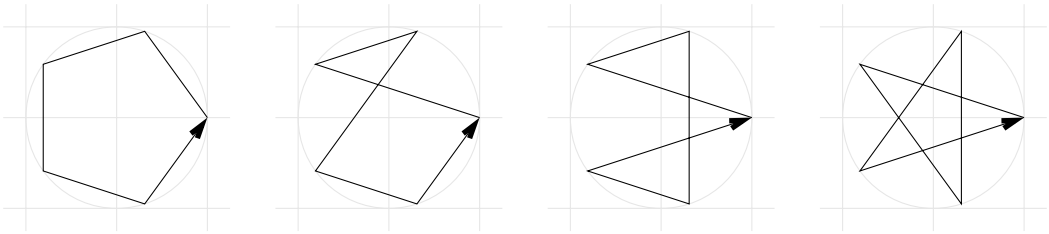

Figure D.3: Four of the 12 basic path shapes in $S_5$

Each of the basic paths can be rotated in $n$ ways, or traveled backwards. Thus, there are $n!/(2n)$ basic path shapes connecting the corners of a regular $n$-gon – only one of them would be the "cardboard" equivalent of $D_n$.

For $n = 3$, there is only one basic path, namely the regular triangle. A twist would amount to the same as a reflection, therefore $D_3$ already describes all the elements of $S_3$ – but as we have demonstrated, this is not the case for $n > 3$.

One can determine all the basic shapes (with both directions) by fixing one of the corners of the regular $n$-gon and applying all $(n-1)!$ permutations of the other corner labels. This will reveal $n!/(2n)$ pairs of basic shapes (i.e. the basic shape in its two directions). Rotating the $(n-1)!$ shapes in any one of $n$ ways yields all possible permutations. This may serve as a justification for the term "symmetric group" – apart from the twists, there are just the $D_n$ symmetry operations at work.

# Appendix E

# Algebra 2: Rings and Fields

We provide a short introduction to finite fields, relying mainly on [Hof14], chapter 2. We will, however, omit certain proofs from Number Theory and refer the interested reader to dedicated textbooks. Neither will we be able to cover all the proofs from Algebra in the scope of this work.

We begin with rings of polynomials, as a prerequisite for constructing finite fields. Then we examine certain finite rings and fields, from which we will generalize to Galois Fields (using polynomials).

## E.1 Polynomials

**Definition E.1** *The finite sum expression*

$$p(x) := \sum_{j=0}^{n} a_j x^j = a_n x^n + \cdots + a_1 x + a_0,$$

*where the coefficients $a_j$ are from a field $\mathbb{F}$ and $a_n \neq 0$ is called* polynomial of degree $n$ in $x$ over $\mathbb{F}$. *The set of all polynomials over $\mathbb{F}$ is denoted $\mathbb{F}[x]$.*

*Any polynomial in $\mathbb{F}[x]$ can be represented by (and identified with) a tuple of its coefficients, e.g. $(a_n, \cdots, a_1, a_0)$.*

**Definition E.2** *The function* $\deg : \mathbb{F}[x] \to \mathbb{N}_0, p(x) \mapsto \deg p(x)$, *is called* degree *and retrieves, for any $p(x) \in \mathbb{F}[x]$, the maximum exponent of $x$ with a non-zero coefficient occurring in $p(x)$.*

**Definition E.3** *The sum of two polynomials $p(x), q(x) \in \mathbb{F}[x]$ is a polynomial of degree*

$$\max\{\deg p(x), \deg q(x)\},$$

*whose coefficients are the sums of the respective coefficients in $p(x), q(x)$. If one in $\{p(x), q(x)\}$ has lesser degree than the other, leading zeros may be added to its tuple representation.*

If the coefficient tuples are brought to the same lengths, addition may be performed component-wise as in vector spaces.

**Corollary E.4** *The polynomials in $\mathbb{F}[x]$ constitute an Abelian group with the polynomial addition.*

Proof: Recalling definition A.3 (p. 67), we observe that the operation is closed as per the above definition E.3 (G1). Since it is defined by component-wise application of $\mathbb{F}$'s addition, it is also associative (G2). The polynomial (0) is the identity element (G3), and since any coefficient has its additive inverse in $\mathbb{F}$, the polynomial $(-a_n, \cdots, -a_1, -a_0)$ is the (well-defined) additive inverse of $(a_n, \cdots, a_1, a_0)$ (G4). The addition described above is also commutative; thus, the group is Abelian. ■

**Definition E.5** *The product of two polynomials $p(x), q(x) \in \mathbb{F}[x]$ is a polynomial with a degree of*

$$\deg p(x) + \deg q(x)$$

*If the coefficients of $p(x), q(x), (p \cdot q)(x)$ are named with $a, b, c$, respectively, then*

$$\forall j \in \{0, 1, \cdots, (\deg p(x) + \deg q(x))\} : c_j := \sum_{\substack{0 \leq \alpha \leq \deg p(x) \\ 0 \leq \beta \leq \deg q(x) \\ \alpha + \beta = j}} a_\alpha \cdot b_\beta$$

**Corollary E.6** *The polynomials in $\mathbb{F}[x]$ constitute a monoid with the polynomial multiplication.*

Proof: Recalling definition A.2 (p. 67), we observe that the operation is closed (G1) and associative (G2). The polynomial (1) is the identity element (G3). ■

However, since there is no way that the product of two polynomials can have lesser degree than its factors, only polynomials of degree zero can have multiplicative inverses – but this is trivial because such polynomials $(a_0) \neq (0)$ are identical with their coefficient $a_0$, for which $\mathbb{F}$ readily provides the inverse $(1/a_0)$.

We omit demonstrating that distributive laws do hold, and directly state, according to definition A.5 (p. 68):

**Corollary E.7** $(\mathbb{F}[x], +, \cdot)$ *with polynomial addition and multiplication is a commutative ring with unity.*

Note that, because $(\mathbb{F}[x] \setminus \{(0)\}, \cdot)$ is not a group, $(\mathbb{F}[x], +, \cdot)$ is not a field (cf. definition A.6, p. 68).

**Definition E.8** *A polynomial $p(x) \in \mathbb{F}[x]$ is called* reducible *if there are non-constant factors $f_1(x), f_2(x) \in \mathbb{F}[x]$ such that $p(x) = f_1(x) \cdot f_2(x)$. It is called* irreducible *otherwise.*

Evidently, while there may be no multiplicative inverses in the ring of polynomials, a reducible polynomial can still be *divided* by one of its factors, yielding a bona fide polynomial[1]. We can also expand this notion to

**Lemma E.9** *(Polynomial Division) Given polynomials $p(x), q(x) \in \mathbb{F}[x]$, where neither is the zero polynomial, there is a unique way to express $p(x)$ as*

$$p(x) = q(x) \cdot s(x) + r(x),$$

*with* $\deg r(x) < \deg q(x)$. *$r(x)$ is called the* residue polynomial.

Proof: Let $n := \deg p(x)$ and $m := \deg q(x)$. If we denote the coefficients of $p(x), q(x), s(x), r(x)$ with $a, b, c, d$, respectively, we may infer from the premise that $a_n$ and $b_m$ are non-zero. We consider three cases:

- If $n < m$, then $q(x) \cdot s(x)$ must be zero because of the remarks after corollary E.6, or the product would have too high a degree. Thus $s(x) := (0)$. This implies $r(x) := p(x)$, satisfying that $\deg r(x) < m$.
- If $n = m$, we can multiply $q(x)$ with a single number, namely the polynomial $s(x) := (c_0)$, with
$$c_0 := \frac{a_n}{b_n} \neq 0$$
This fixes the highest coefficient in $q(x) \cdot s(x)$ as $a_n$. For the lower coefficients, we observe that $a_j = b_j c_0 + (a_j - b_j c_0) =: b_j c_0 + d_j$. The first part of that sum is the coefficient $j$ for $q(x) \cdot s(x)$, and the second (in brackets) the necessary correction in $r(x)$. But since $j < n$, $\deg r(x) < n$, with a maximum value of $(n-1)$.
- If $n > m$, this is the classic case for polynomial division. We take $s(x)$ to be a polynomial with degree $(n-m)$. We now write $(n+1)$ equations for the coefficients of $p(x)$ – the first $(n-m+1)$ ones (from $j = n$ to $j = (n-(n-m)) = m$) will determine the coefficients of $s(x)$; the remaining $m$ ones (from $j = (m-1)$ to $j = 0$) will fix $r(x)$, which will have degree up to $(m-1)$.
  - For $j = n$ downwards to $j = m$:
$$a_j = \sum_{\substack{0 \leq \alpha \leq m \\ 0 \leq \beta \leq (n-m) \\ \alpha + \beta = j}} b_\alpha c_\beta$$
For $j = n$, there is only one contribution: $a_n = b_m c_{n-m}$, which fixes $c_{n-m}$.
For $j = (n-1)$ (if $m > 0$), there are two contributions: $a_{n-1} = b_{m-1} c_{n-m} + b_m c_{n-m-1}$. Since we know $c_{n-m}$, we can plug this in and can fix $c_{n-m-1}$.

[1] In fact, fractions of polynomials (rational functions) do constitute a field; polynomials are just rational functions with denominator 1.

And so forth: Each equation contains up to $(1 + \min\{m, (n-m)\})$ contributions (depending on the index ranges, and which of the polynomials $q(x), s(x)$ has lesser degree), including the product $b_m c_{j-m}$, and possibly coefficients of $s(x)$ with higher index that have already been fixed. $c_{j-m}$ is always the only unknown variable and can be fixed in the current equation. This is always possible because $b_m \neq 0$.

The last step fixes $c_{m-m} = c_0$; now, the polynomial $s(x)$ has been determined.

– For $j = (m-1)$ downwards to $j = 0$ (although the order does not matter here):

$$a_j = d_j + \sum_{\substack{0 \leq \alpha \leq m \\ 0 \leq \beta \leq (n-m) \\ \alpha+\beta=j}} b_\alpha c_\beta$$

Here, the $d_j$ are the unknowns, and there is also one per equation, allowing the $d_j$ to be fixed one by one. All the $c_\beta$ are already known from above.

After this (deterministic!) procedure, $s(x)$ and $r(x)$ have been uniquely determined.

The degree of $s(x)$ is fixed as $(n-m)$, but $r$ may have lesser degree than $(m-1)$, if higher $d$ coefficients are zero.

If $r(x) = (0)$, $p(x)$ is reducible and $q(x), s(x)$ are a pair of factors of $p(x)$.

■

As an example, we calculate the division of $p(x) = (1, 0, 0, 3, -2, 0, 1, 1)$ by $q(x) = (1, 0, -4, 3)$. Evidently, $\deg p(x) = 7$ and $\deg q(x) = 3$, which implies $\deg s(x) = (7-3) = 4$ and $\deg r(x) < 3$.

The non-zero coefficients in $p(x)$ and $q(x)$ are:

$$a_7 = 1, \quad a_4 = 3, \quad a_3 = -2, \quad a_1 = 1, \quad a_0 = 1, \qquad b_3 = 1, \quad b_1 = -4, \quad b_0 = 3$$

We give the eight equations that determine the coefficients $c_4, \cdots, c_0$ and $d_2, \cdots, d_0$.

$$\begin{array}{lrcll|lcr}
j=7: & 1 & = & b_3c_4 = c_4 & & \rightsquigarrow & c_4 & = & 1 \\
j=6: & 0 & = & b_2c_4 + b_3c_3 = c_3 & & \rightsquigarrow & c_3 & = & 0 \\
j=5: & 0 & = & b_1c_4 + b_2c_3 + b_3c_2 = -4 + c_2 & & \rightsquigarrow & c_2 & = & 4 \\
j=4: & 3 & = & b_0c_4 + b_1c_3 + b_2c_2 + b_3c_1 = 3 + c_1 & & \rightsquigarrow & c_1 & = & 0 \\
j=3: & -2 & = & b_0c_3 + b_1c_2 + b_2c_1 + b_3c_0 = -16 + c_0 & & \rightsquigarrow & c_0 & = & 14 \\
\hline
j=2: & 0 & = & d_2 + b_0c_2 + b_1c_1 + b_2c_0 = d_2 + 12 & & \rightsquigarrow & d_2 & = & -12 \\
j=1: & 1 & = & d_1 + b_0c_1 + b_1c_0 = d_1 - 56 & & \rightsquigarrow & d_1 & = & 57 \\
j=0: & 1 & = & d_0 + b_0c_0 = d_0 + 42 & & \rightsquigarrow & d_0 & = & -41
\end{array}$$

Thus:

$$(x^7 + 3x^4 - 2x^3 + x + 1) = (x^3 - 4x + 3) \cdot (x^4 + 4x^2 + 14) + (-12x^2 + 57x - 41)$$

## E.2 Residue Rings and Residue Fields

We already mentioned the fact that $(\mathbb{Z}, +, \cdot)$ is a commutative ring with unity as an example for the ring definition A.5 (p. 68). We now want to consider *finite* sets of numbers. The integers can be partitioned into a finite set of residue classes when operating modulo some natural number. For this, we introduce the concept of *residue rings*, for which we need only a few preparations:

**Definition E.10** *For $m \in \mathbb{N}, j \in \mathbb{Z}$, the set $\{j + km | k \in \mathbb{Z}\}$ is called* residue class of $j$ modulo $m$, *denoted $[j]_m$.*

(If $m$ is fixed, we omit the subscript "$_m$" for residue classes.)

**Definition E.11** *For $m \in \mathbb{N}$, the* residue system *of $m$, denoted $\mathbb{Z}_m$, is the set of possible residues modulo $m$:*

$$\mathbb{Z}_m = \{0, 1, \cdots, (m-1)\}$$

Strictly speaking, this is a bit hand-waving; the residue system formally should be the set of residue *classes*. But no information is lost here, because if we operate modulo $m$, we may pick any member of a residue class as its representative; and we opt to always choose the smallest non-negative member.

**Lemma E.12** *For $m \in \mathbb{N}$, $(\mathbb{Z}_m, +)$ is an Abelian group with identity 0 and $(\mathbb{Z}_m, \cdot)$ is a semigroup. For $m > 1$, $(\mathbb{Z}_m, \cdot)$ is a monoid with identity 1.*

Proof: We recall the definitions A.1, A.2 and A.3 on pp. 66f:

- Addition modulo $m$ in $\mathbb{Z}_m$ is commutative, associative and closed. The latter holds because taking the modulus after addition always will project any number into $\mathbb{Z}_m$.

  The identity element is $0 \in \mathbb{Z}_m$, and any element $j \in \mathbb{Z}_m$ has a unique additive inverse in $\mathbb{Z}_m$: If $j$ is zero, its inverse is $0 \in \mathbb{Z}_m$. Any other $j$ has inverse $(m - j)$, which is an integer between (exclusively) 0 and $m$, and thus $(m - j) \in \mathbb{Z}_m$.

- Multiplication modulo $m$ in $\mathbb{Z}_m$ is also commutative, associative and closed. Therefore, $(\mathbb{Z}_m, \cdot)$ is a semigroup.

- If $m > 1$, $\mathbb{Z}_m$ contains at least 0 and 1, the latter being the identity of "$\cdot$". In that case, $\mathbb{Z}_m$, therefore, is a monoid.

■

We omit the proof that multiplication distributes over addition in the familiar way in $\mathbb{Z}_m$ and conclude:

**Corollary E.13** *For $m \in \mathbb{N}$, $(\mathbb{Z}_m, +, \cdot)$ is a commutative ring (with unity, if $m > 1$).*

For the following, we want to fix $m > 1$, because operating modulo 1 puts all integers in a single residue class [0]; from there on, all previous numeric information is lost.

We now consider what restrictions we can place on $m$ so that $(\mathbb{Z}_m, +, \cdot)$ is not only a commutative ring with unity but a full field. Recalling definition A.6 (p. 68), we need to ensure that any $j$ in $\mathbb{Z}_m \setminus \{0\}$ has a multiplicative inverse modulo $m$.

In preparation, we consider co-primeness in a product of natural numbers:

**Lemma E.14** *$a, b \in \mathbb{N}$ are both co-prime with $m \in \mathbb{N}$ if and only if their product $(ab)$ is co-prime with $m$.*

Proof: We show this equivalence in its negated form.

- If $a$ or $b$ are *not* co-prime with $m$, one of them (without loss of generality, let this be $a$), shares a factor $d > 1$ with $m$ per

  $$a = \alpha d \quad \wedge \quad m = \beta d$$

  But then, $(ab)$ contains the factor $d$ too via $ab = \alpha db$. Thus, $(ab)$ and $m$ share the factor $d$, and cannot be co-prime.

- If $(ab)$ and $m$ are not co-prime, they share a factor $d > 1$ per

  $$ab = \alpha d \quad \wedge \quad m = \beta d$$

  We may assume that $d$ is prime. If it were not, any factors could be absorbed into both $\alpha$ and $\beta$. But, per Euclid's lemma, if $d$ is prime and divides $(ab)$, then it must divide at least one of its factors $a, b$: Since $d$ is prime, the only way to spread $d$ into a product such that $a$ and $b$ may take up one factor each is $d = 1 \cdot d$.

  Thus, $d$ is a factor af at least one of $a, b$, and not both $a$ and $b$ can be co-prime with $m$.

■

At this point we will have to take a closer look at gcd calculation, and establish the principles of the Euclidean algorithm.

**Lemma E.15** *For $a, b \in \mathbb{N}$, $a \geq b$:* $\gcd(a, b) = \gcd(a - b, b)$.

Proof: Since $a, b$ are (positive) natural numbers, they share a common divisor $d$, which is at least 1 (and exactly 1 if and only if they are co-prime). Of all common divisors, let $d$ be the biggest one. Then:

$$a = \alpha d \quad \wedge \quad b = \beta d$$

with co-prime $\alpha, \beta \in \mathbb{N}$, $\alpha \geq \beta$.

Then, the difference $(a - b)$ equals $(\alpha - \beta)d$. Since $\alpha, \beta$ are co-prime (or they would share a non-trivial factor that would need to be in $d$), no additional factor can be extracted from $(\alpha - \beta)$. But then, $d$ is also the gcd of $(a - b)$ and $b$. ■

**Corollary E.16** *For $a, b \in \mathbb{N}$, $a = qb + r$, $0 \leq r < b$:*

$$\gcd(a, b) = \gcd(a - qb, b) = \gcd(r, b) = \gcd(a \bmod b, b)$$

Proof: Apply the lemma E.15 repeatedly, as long as its condition $a \geq b$ is met. This amounts to a modulo operation. ■

The modulo rule is useful for gcd calculation because, in this case, it is always obvious which one of the gcd arguments is the greater one. If we take $a \bmod b$, the result will be in $\{0, 1, \cdots, (b-1)\}$ but definitely less than $b$. This allows us to formulate a concise version of Euclid's algorithm:

**Lemma E.17** *(Euclidean Algorithm) For $a, b \in \mathbb{N}$, $\gcd(a, b)$ can be calculated in the following way:*

- *Let $M_1 := \max\{a, b\}$, $m_1 := \min\{a, b\}$. Then, $r_1 := M_1 \bmod m_1$.*
- *For $k > 1$: Let $M_k := m_{k-1}$, $m_k := r_{k-1}$.*
  *Then, $r_k := M_k \bmod m_k$, and $\gcd(M_{k-1}, m_{k-1}) = \gcd(M_k, m_k)$.*
- *Terminate for any $k \in \mathbb{N}$ when $r_k = 0$. Then, $\gcd(M_k, m_k) = m_k$.*

Proof: The iteration of gcd functions is valid as per corollary E.16. If the maximum of the gcd arguments is a multiple of the minimum (including if they are equal, or if the minimum is 1), the new remainder would be zero, and the next iteration would calculate the gcd of the previous minimum and zero, which is the previous minimum because any natural number divides zero.

Each step of the algorithm will feature a minimum value $m_k$ that is less than in the previous step, because of the modulo operation. The sequence of $m_k$ therefore is strictly monotonically decreasing, but $m_k \geq 0$ for all steps. The value zero would be reached if we did the next iteration step after $r_k = 0$ has occurred. Thus, the algorithm must terminate after at most $m_1$ steps. ■

**Corollary E.18** *If a number $a \in \mathbb{N}$ is in the residue class $[1]_m$ for some $m \in \mathbb{N}$, $\gcd(a, m) = \gcd(1, m) = 1$.*

Proof: $a$ is either 1 (nothing further to do), or $a = (km + 1) > m$ for some $k \in \mathbb{N}$. Use the Euclidean algorithm:

$$r_1 = (km + 1) \bmod m = 1 \quad \blacksquare$$

With these preparations, we can prove the most important criterion we need to establish residue fields:

**Lemma E.19** *$j \in \mathbb{Z}_m \setminus \{0\}$ has a multiplicative inverse modulo $m$ if and only if $j$ and $m$ are co-prime, i.e. $\gcd(j, m) = 1$.*

Proof:

- If $j \in \mathbb{Z}_m \setminus \{0\}$ has a multiplicative inverse modulo $m$, let us call this inverse $\tilde{\jmath}$. Then, the product $j \cdot \tilde{\jmath}$ is 1 (modulo $m$). Use corollary E.18 to establish that $j \cdot \tilde{\jmath}$ is co-prime with $m$. Now we can use lemma E.14 from above and see that both $j$ and $\tilde{\jmath}$ are co-prime with $m$.
- If $j$ is co-prime with $m$, then $\gcd(j, m) = 1$. We could use Bezout's identity

  $$\exists \alpha, \beta \in \mathbb{Z} : \gcd(j, m) = \alpha j + \beta m$$

  to establish the existence of an inverse: Since $\gcd(j, m) = 1$, this means that $1 = \alpha j + \beta m$. If we take this equation modulo $m$, we obtain:

  $$1 \equiv (\alpha \cdot j) \mod m$$

  Thus, the inverse of $j$ exists and equals $\alpha$ (modulo $m$). We omit the formal proof of Bezout's identity here; it is short and can be found in any textbook on number theory. Instead, we give a construction using the extended Euclidean algorithm, which actually delivers the values of $\alpha, \beta$, so that we can calculate $j$'s inverse.

  For this, we can use an extension of the Euclidean algorithm E.17 from above. In any algorithm step, we may extend the calculation $r_k := M_k \bmod m_k$ by stating that $r_k = M_k - q_k m_k$. In step 1, this directly allows us to express $r_1$ in terms of (here) $j$ and $m$.

In step 2, we use $M_2 = m_1 = j$ and $m_2 = r_1$, which was expressed in terms of $j$ and $m$ in step 1. The factors for Bezout's identity emerge after collecting the terms.

For higher steps, $m_k = r_{k-1}$ and $M_k = r_{k-2}$, so $r_k$ can readily be expressed in terms of $j$ and $m$ by looking up the previous two remainders (provided those have been expressed as linear combinations of $j$ and $m$).

After this extended Euclidean algorithm has terminated, we have found the inverse of $j$ modulo $m$, which is sufficient to show that such an inverse exists.

■

As an example, we consider $\gcd(19, 17)$, which is 1 because both numbers are prime. We find the inverse of 17 (modulo 19), highlighting the $M$, $m$ and $r$ numbers:

$$\begin{array}{rclcrcl} \underline{19} & = & 1 \cdot \underline{17} + \underline{2} & \Leftrightarrow & \underline{2} & = & 1 \cdot \underline{19} - 1 \cdot \underline{17} \\ \underline{17} & = & 8 \cdot \underline{2} + \underline{1} & \Leftrightarrow & \underline{1} & = & \underline{17} - 8 \cdot \underline{2} \\ & & & & & = & \underline{17} - 8 \cdot (1 \cdot \underline{19} - 1 \cdot \underline{17}) \\ & & & & & = & (-8) \cdot \underline{19} + 9 \cdot \underline{17} \\ \underline{2} & = & 2 \cdot \underline{1} + \underline{0} & & \text{terminate: } r_3 = 0 & & \end{array}$$

So, if we take the equation for $r_2 = 1$ modulo 19, we obtain: $9 \cdot 17 \equiv 1 \mod 19$, so the inverse of 17 (modulo 19) is 9. And, in fact $9 \cdot 17 = 153$, and $8 \cdot 19 = 152$.

We can now collect the preparations from above and state the following:

**Corollary E.20** *For $m \in \mathbb{N}$, $m > 1$, $(\mathbb{Z}_m, +, \cdot)$ is a field if and only if $m$ is a prime number.*

Proof: For this, any $j \in \mathbb{Z}_m \setminus \{0\}$ must have a multiplicative inverse, so that $(\mathbb{Z}_m \setminus \{0\}, \cdot)$ is a group. The prime numbers are exactly the numbers in $\mathbb{N}$ that satisfy $\gcd(j, m) = 1$ for $1 \leq j < m$. ■

For example, we give the multiplication table of $\mathbb{Z}_5 \setminus \{0\}$:

| $\cdot$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 1 | 2 | 3 | 4 |
| 2 | 2 | 4 | 1 | 3 |
| 3 | 3 | 1 | 4 | 2 |
| 4 | 4 | 3 | 2 | 1 |

# E.3 Galois Fields

## E.3.1 Remarks

It can be shown (e.g. [KM21], ch. 26) that finite fields are possible with exactly $p^k$ elements, where $p$ is prime and $k \in \mathbb{N}$, and that all the various realizations with an equal number of elements are isomorphic, i.e. equal up to a permutation of the element labels. These finite fields are often called *Galois fields*, denoted $GF(p^k)$.

For the case $k = 1$, we can use the preceding section to argue that $(\mathbb{Z}_p, +, \cdot)$ is a finite field with $p$ elements, so $GF(p) \cong \mathbb{Z}_p$.

This will not work (cf. the preceding section) for $k > 1$, though, because $p^k$ is no longer prime. But we can give an alternative construction, using polynomials. We follow [Hof14] for the details. It is important to remember, that the numbers $\{0, 1, \cdots, p^k - 1\}$ cannot be taken literally, but as labels, or otherwise uniquely mapped to the actual elements. For instance, if 2 represents the polynomial $(2x + 1)$, the squared polynomial $(4x^2 + 4x + 1)$ may not be represented by $4 = 2^2$, but by some other number, maybe 7. Only for $k = 1$ can the labels from $\mathbb{N}_0$ represent themselves as numbers.

We will lay out the principle of obtaining a realization of $GF(p^k)$ for fixed but arbitrary $k > 1$ and $p$, and consider the case $p^k = 2^4$ as an example.

## E.3.2 Preparations

We already mentioned the notion of (ir-)reducible polynomials in definition E.8, p. 108. In the above section E.1, the field from where the coefficients are taken was kept unfixed deliberately: We now examine polynomials over $\mathbb{Z}_p$:

**Lemma E.21** *The ring of polynomials $\mathbb{Z}_p[x]$ contains $p^k$ different polynomials of lesser degree than $k$.*

Proof: We temporarily view polynomials of degree up to $(k-1)$ as full polynomials of degree $(k-1)$, i.e. we allow leading zeros in the coefficient tuple, which makes counting easier. We already stated in definition E.1 (p. 107) that we may represent each polynomial by its coefficient tuple.

Now, if the field is finite (with $p$ elements), a polynomial of degree $(k-1)$ has $k$ coefficients. Each of those $k$ coefficients can take any of $p$ values, making for $p^k$ different tuples. If we now remove the leading zeros again, we are left with proper representations of polynomials with lesser degree. ■

In anticipation of the later example, we list the polynomials with degree less than 4 over $\mathbb{Z}_2$. We also provide their tuple representations (with leading zeros, but without commas or brackets) and the numbers in $\mathbb{N}_0$ they code for:

| | | | | | |
|---|---|---|---|---|---|
| $0$ | 0000 | 0 | $x^3$ | 1000 | 8 |
| $1$ | 0001 | 1 | $x^3+1$ | 1001 | 9 |
| $x$ | 0010 | 2 | $x^3+x$ | 1010 | 10 |
| $x+1$ | 0011 | 3 | $x^3+x+1$ | 1011 | 11 |
| $x^2$ | 0100 | 4 | $x^3+x^2$ | 1100 | 12 |
| $x^2+1$ | 0101 | 5 | $x^3+x^2+1$ | 1101 | 13 |
| $x^2+x$ | 0110 | 6 | $x^3+x^2+x$ | 1110 | 14 |
| $x^2+x+1$ | 0111 | 7 | $x^3+x^2+x+1$ | 1111 | 15 |

Table E.1: Polynomials with degree less than 4 in $\mathbb{Z}_2[x]$, with tuple representations

If we concatenate the coefficients, we get binary (in general: $p$-ary) strings which code for numbers in $\mathbb{N}_0$ – here, 0 up to $(16-1)=15$.

In fact, if we take the product of the polynomials coded by 2 and 3, respectively (binary 10 and 11), the result is the polynomial coded by 6 (binary 110). This is possible because we labeled the polynomials in a consistent way: plugging in the number 2 (in general: $p$) for the argument $x$, the polynomial functions labeled this way return just the numbers that the binary ($p$-ary) coefficient strings encode.

However, this will not work in this simple way for all combinations: half the polynomials are degree three, and could only be multiplied with 1 or 0 to stay within the set. Multiplying them with any higher-degree polynomial yields degrees of up to 6.

When examining the multiplicative groups of the finite fields $\mathbb{Z}_p$, we always performed multiplications modulo $p$. It turns out that the same is possible for polynomials as well, by way of polynomial division as in lemma E.9 (p. 108), – only that the modulus needs to be a polynomial, not a number.

In fact, the polynomial division with remainder polynomials is the exact equivalent of arithmetic integer division with remainder. Both are called "Euclidean division".

For $\mathbb{Z}_p$ we then considered the remainders (and their residue classes) modulo $p$ as the elements of our finite fields – in the same way, we can view the remainders of polynomial division by a (fixed) modulus polynomial as elements of the Galois fields. Since Euclidean division is possible for numbers and for polynomials in the same way, the concepts of co-primeness, greatest common divisor, and the (extended) Euclidean algorithm can be transferred as well; thus, all the remarks from the previous section can be translated into equivalent concepts for polynomials.

We may therefore forgo a detailed repetition of our previous observations and concentrate more on the practical differences.

One of the key insights in the previous section was that in order for the ring to become a proper field requires that each element other than zero should have its unique multiplicative inverse (modulo $p$, or, in this case our modulus polynomial). We found that only the prime numbers (and all the prime numbers) could be suitable moduli for the integers. The equivalent notion of a prime number would be, in our case, an *irreducible* polynomial, i.e. one that has no factors of lesser degree except (1) – and therefore would be co-prime with all those polynomials of lesser degree. If we now find such a polynomial of degree $k$, we can use Bezout's identity via the extended Euclidean algorithm to determine each element's multiplicative inverse. Any irreducible polynomial of degree $k$ will do for that purpose.

Considering our example, we should point out that the polynomial $x^4$ corresponding to the number 16 would be of degree 4, but could *not* serve as a modulus polynomial because it has factors, e.g. $x$, and thus is not co-prime with all the polynomials of lesser degree than 4. We will, in the next subsection, find that there are polynomials of degree 4 that qualify as modulus – in fact, there is more than one. This is a difference to the $\mathbb{Z}_p$ case, because when dealing with numbers there always was exactly the modulus $p$.

**Lemma E.22** *The modulus polynomial $m(x)$ of degree $k$, which is used to restrict multiplication in $GF(p^k)$ to polynomials with degree less than $k$, must be irreducible.*

Proof: If $GF(p^k)$ is a field, all non-zero elements must have unique multiplicative inverses. Assume that $m(x) = f_1(x) \cdot f_2(x)$, then those factors have degree less than $k$. We consider their inverses, and use the fact that $f_1(x) \cdot f_2(x) \equiv 0 \mod m(x)$:

$$\begin{aligned} 0 &= f_1^{-1}(x) \cdot 0 = f_1^{-1}(x) \cdot f_1(x) \cdot f_2(x) = 1 \cdot f_2(x) = f_2(x) \\ 0 &= 0 \cdot f_2^{-1}(x) = f_1(x) \cdot f_2(x) \cdot f_2^{-1}(x) = f_1(x) \cdot 1 = f_1(x) \end{aligned}$$

So, if the inverse elements of $f_1(x), f_2(x)$ exist, this implies that $f_2(x), f_1(x)$ are zero, respectively, which cannot be unless the whole $m(x)$ were zero. Therefore, $m(x)$ needs to be irreducible. ■

**Corollary E.23** *The modulus polynomial $m(x)$ for $GF(p^k)$ may not have zero as its $x^0$ coefficient.*

Proof: If that coefficient were zero, the polynomial would contain a factor of $x$, and thus be reducible. ■

We omit the proof that one can always find an irreducible polynomial in $\mathbb{Z}_p[x]$ for any degree $k$. While we will present a method to obtain an irreducible polynomial, without this proof our method will necessarily stay optimistic.

### E.3.3 Construction Example

As indicated above, we present the case of $GF(2^4)$, the finite field with 16 elements. The elements consist of the 16 polynomials with degree less than 4, and the coefficients are from $\mathbb{Z}_2 = \{0, 1\}$ For the coefficients, this means that adding and subtracting amount to the same operation since $1 + 1 \equiv 0 \mod 2$; this also holds for polynomials in $\mathbb{Z}_2[x]$ because they are added component-wise.

**Definition E.24** *For $k \in \mathbb{N}$, $p$ prime, the Galois field $GF(p^k)$ is isomorphic to $\mathbb{Z}_p[x]_{m(x)}$, for a given (and fixed) irreducible modulus polynomial $m(x)$ with degree $k$. Addition in $\mathbb{Z}_p[x]_{m(x)}$ is executed component-wise, and modulo $p$ per component. Multiplication in $\mathbb{Z}_p[x]_{m(x)}$ is executed modulo $m(x)$ in order to obtain polynomials of degree less than $k$.*

#### Finding an Irreducible Polynomial for the Modulus

In order to establish a multiplication table, we first need to find one irreducible polynomial $m(x) = (m_4, m_3, m_2, m_1, m_0)$ of degree 4 (cf. lemma E.22). According to corollary E.23, the component $m_0$ may not be zero; in the case of $\mathbb{Z}_2$ this already fixes $m_0 = 1$. We can employ the same argument for $m_4$, because $m_4 \neq 0$ for $\deg m(x) = 4$; thus, $m_4 = 1$.

We now fix the remaining three coefficients (which afford eight possibilities) in a systematic way, by eliminating all the possible candidates that are in fact reducible polynomials.

In order to detect the reducible polynomials, we consider all products of lesser-degree polynomials that conform to our previous restrictions. In this case, we have to look at factor degree combinations $(1, 3)$ and $(2, 2)$ (only those will yield a product of degree 4), and we also can confine ourselves to factors ending with an $x^0$ coefficient of 1, because only those can yield $m_0 = 1$ for the product. (There would be more possibilities and fewer restrictions (relatively) in case $p > 2$.)

For the degree combination $(1, 3)$, there is only one factor with degree 1 to consider, namely $(x + 1)$, and four factors with degree 3, all of which we can obtain from table E.1 (p. 113). For the degree combination $(2, 2)$, we find two possible factors, which we can multiply in three ways: each with itself, and one with the other.

We execute the multiplications in tuple notation (without brackets or commas), like an integer multiplication done manually on paper. When summing the intermediate results, only the columns with odd parity (i.e. the number of '1' bits) yield 1; the others, zero, because of $\mathbb{Z}_2$ arithmetic.

$$11 \cdot 1001 = \begin{array}{rr} & 1001 \\ + & 10010 \\ \hline & 11011 \end{array} \qquad 11 \cdot 1011 = \begin{array}{rr} & 1011 \\ + & 10110 \\ \hline & 11101 \end{array}$$

$$11 \cdot 1101 = \begin{array}{rr} & 1101 \\ + & 11010 \\ \hline & 10111 \end{array} \qquad 11 \cdot 1111 = \begin{array}{rr} & 1111 \\ + & 11110 \\ \hline & 10001 \end{array}$$

$$101 \cdot 101 = \begin{array}{rr} & 101 \\ + & 10100 \\ \hline & 10001 \end{array} \qquad 101 \cdot 111 = \begin{array}{rr} & 101 \\ + & 1010 \\ + & 10100 \\ \hline & 11011 \end{array}$$

$$111 \cdot 111 = \begin{array}{rr} & 111 \\ + & 1110 \\ + & 11100 \\ \hline & 10101 \end{array}$$

This yields five distinct reducible polynomials: 11011, 11101, 10111, 10001 and 10101. The other three remaining polynomials of degree 4 therefore must be irreducible: 10011, 11001 and 11111.

**Creating the Multiplication Table**

For this example of $GF(2^4)$, we want to take the following irreducible polynomial as modulus:

$$m(x) = (x^4 + x + 1) = 10011$$

The multiplication table of $\mathbb{Z}_2[x]_{m(x)} \backslash \{0\}$ will have $15^2 = 225$ entries, but since it is symmetric, only $15 \cdot 16/2 = 120$ equations have to be solved. Of those, 15 are trivial because they are multiplications with 1, and another six are easy because multiplication with 10 amount to shifts for polynomials whose four-bit representation starts with a zero. Another five equations involve multiplication with 11 and polynomials with bit representations with a leading zero. Those 26 equations do not involve taking the modulus, because the products are still of degree less than four, and we recall the first case of lemma E.9 (p. 108) about polynomial division, which states that the remainder when dividing by $m(x)$ must then just be that product. For these cases, multiplication may be carried out exactly as above (when searching for an irreducible polynomial).

The remaining 94 non-trivial equations will involve taking the modulus after calculating the product, i.e. we calculate the product, then divide it by $m(x)$ in a polynomial division[2], and take the remainder polynomial (which will have a degree less than four) as result.

We give one example for such a multiplication, namely the product of 6 and 13, i.e. the polynomials 110 and 1101. First, the product:

$$110 \cdot 1101 = \begin{array}{rr} & 11010 \\ + & 110100 \\ \hline & 101110 \end{array}$$

This is a polynomial of degree five. We take the remainder modulo $m(x)$, i.e. we calculate

$$101110 = 10011 \cdot (?) + r(x),$$

where we ignore the quotient completely and just collect $r(x)$. In $\mathbb{Z}_2[x]$ we only ever can subtract the modulus, which amounts to adding it, which is an `xor` operation on the binary bits. For $p > 2$, the arithmetic would be a bit more involved.

$$\begin{array}{rll} & 101110 & = 10011 \cdot (?) + r(x) \\ - & 10011 & \\ \hline & 001000 & \end{array}$$

[2]In fact, this procedure can be simplified by using Horner's Rule for polynomials, as explained in [Hof14], subsection 2.4.3. This spreads the multiplication to successive additions and multiplications with $x = 10$, and taking the modulus several times and as soon as necessary. Thus, one would only have to deal with case 2 of the polynomial division (cf. lemma E.9, p. 108), and the remainder could be obtained by one subtraction.

Here, we only had to subtract the modulus once; the remainder is $r(x) = 1000 = x^3$, a polynomial of degree three, coding for the number 8 in our above table E.1 (p. 113).

In this way, the whole multiplication table can be calculated. The table does, however, depend on the choice of $m(x)$, the irreducible modulus polynomial.

If we identify the polynomials with their coefficient strings and the $p$-ary numbers from $\mathbb{N}_0$ associated with those codes, it is evident that the various multiplication tables can be reshuffled with permutations, and that all the $\mathbb{Z}_p[x]_{m(x)}$ are isomorphic to each other.

### E.3.4 The Special Case $k = 1$

It is possible to construct Galois fields with $p^1 = p$ elements with polynomials in the same way we described above – but we can quickly demonstrate that this yields no new information, other than that we could label the polynomials of degree less than 1 (i.e. numbers in $\{0, \cdots, (p-1)\}$) in some permuted way and obtain an isomorphic field with $p$ elements.

Let us consider the case $p = 5$. We now need an irreducible modulus polynomial of degree 1, whose components $m_1$ and $m_0$ are non-zero. There are sixteen different such polynomials, from $x+1$ to $4x+4$. Each of them is irreducible because a polynomial of degree 1 could not be factored by multiplying two numbers, and while e.g. $4x + 2$ equals $(2) \cdot (2x + 1)$, (2) is not a proper factor when considering reducibility, as the factors should be non-constant (cf. definition E.8, p. 108), i.e. at least of degree 1.

The polynomials that would be the elements of the field $GF(5)$ are none other than $(0), \cdots, (4)$, the five polynomials of lesser degree than 1.

Multiplying such polynomials yields just the product of two numbers, modulo $p$, so $(3)\cdot(3) = (4)$. The choice of modulus polynomial is irrelevant here because the modulus has higher degree (cf. the first case in lemma E.9, p. 108), and any of the sixteen moduli would yield $(4) \equiv (4) \mod m(x)$. All the various degree-1 irreducible polynomials therefore produce the same multiplication table.

Thus, $GF(p^1)$ is isomorphic with $\mathbb{Z}_p$, which at least demonstrates consistency – although we could not have done our $GF(p^k)$ construction at all if we had not already accepted $\mathbb{Z}_p$ as fields (for the polynomial coefficients), so this observation is tautological.

# Appendix F

# Java Algorithm for Isoperimetric Constants

For the isoperimetric constant of an undirected graph $G = (V, e)$, we need to consider all bi-partitions of its node set, and remember both the number of connecting edges between $S$ and $\bar{S} = V \setminus S$ and the node count of the smaller (or at least, not larger) partition. In order to avoid double calculations, one can restrict $S \leq |V|/2$. For the purpose of documentation, it is also helpful to remember the precise subset $S$ in some form.

We store this information in objects of type `IsoperimetricDataEntry` (lines 2ff.). While looping over all possible subsets, those entries are collected in a sorted set of type `TreeSet` (lines 13ff.), such that they are ordered ascending in the quotients of $e(S, \bar{S})/|S|$. In order to avoid floating point arithmetic, the comparator of two entries only evaluates the numerators of the two fractions after converting them to a common denominator (line 19).

For each partition of $V$, two sets are populated according to the bits in an `int` number – this is possible because the `Graph` objects have a maximum node count of $30 < 32$ (because of the exponential execution time, graphs with more than twenty nodes are rather time-consuming to process on a private computer, so this restriction is in fact acceptable in practice).

The loop is executed in order to consider all bit patterns from $00\ldots01$ to $11\ldots11$, avoiding the number 0 where no bit is set. If a bit is set, its corresponding node index is added to the selected `subset`; otherwise it belongs to the set `complement`. Because of the maximum node count, no unwanted effects due to 2-complement integer encoding have to be considered.

After the sets have been populated, the edges between $S$ and $\bar{S}$ are counted (lines 51ff.), which involves iterator calls because the Java `Set` are not originally intended for entry traversal.

The main function returns the smallest `IsoperimetricDataEntry` of the `TreeSet` structure, which corresponds to the smallest quotient of $e(S, \bar{S})/|S|$.

The following page shows an excerpt of the Java class used to calculate several isoperimetric constants for toy examples (with some boilerplate code removed). The type `Graph` is omitted here because it contains nothing algorithmically interesting – it is just a wrapper for an adjacency matrix with various access methods and a `String` output. Calls to the `Graph` type are in lines 10 (to retrieve the graph's node count) and 60f (to access the adjacency matrix element).

```
public class Isoperimetric {
  public static class IsoperimetricDataEntry {
    private int subsetSelectors;
    private int subsetSize;
    private int edgeCount;
    // getters, setters, toString
  }

  public static IsoperimetricDataEntry calculateIsoperimetricConstant(Graph g) {
    int nodeCount = g.getNodeCount();

    // prepare a sorted set of various subsets
    SortedSet<IsoperimetricDataEntry> isoperimetricEntries
      = new TreeSet<>(new Comparator<IsoperimetricDataEntry>() {

      @Override
      public int compare(IsoperimetricDataEntry o1,
          IsoperimetricDataEntry o2) {
        return (o2.subsetSize * o1.edgeCount - o1.subsetSize * o2.edgeCount);
      }

    });

    Set<Integer> subset = new HashSet<>();
    Set<Integer> complement = new HashSet<>();

    // loop over all possible subsets (represented as binary flags in an int)
    for (int subsetSelectors = 1;
             subsetSelectors < (1 << nodeCount); subsetSelectors++) {

      // initialize sets
      subset.clear();
      for (int j = 0; j < nodeCount; j++) {
        complement.add(j);
      }

      // populate sets
      int setSize = 0;
      int node = 0;
      while (node < nodeCount && setSize < (nodeCount / 2)) {
        if ((subsetSelectors & (1 << node)) != 0) {
          subset.add(node);
          complement.remove(node);
          setSize++;
        }
        node++;
      }

      // determine edge count between subset and complement,
      // and add data entry to the sorted set
      if (setSize > 0) {
        int edgeCount = 0;
        Iterator<Integer> subsetNodes = subset.iterator();
        Iterator<Integer> complementNodes;
        int subsetNode;
        while (subsetNodes.hasNext()) {
          subsetNode = subsetNodes.next();
          complementNodes = complement.iterator();
          while (complementNodes.hasNext()) {
            edgeCount += g.getAdjacencyInternal(subsetNode,
                complementNodes.next());
          }
        }
        IsoperimetricDataEntry dataEntry = new IsoperimetricDataEntry();
        dataEntry.subsetSelectors = subsetSelectors;
        dataEntry.subsetSize = setSize;
        dataEntry.edgeCount = edgeCount;
        isoperimetricEntries.add(dataEntry);
      }
    }

    return isoperimetricEntries.isEmpty() ? null : isoperimetricEntries.first();
  }
}
```

# Bibliography


[AR94] N. Alon and Y. Roichman. Random Cayley Graphs and Expanders. *Random Structures and Algorithms*, 5:271–284, 1994. Reprint online from 2002-02-22, `www.cs.tau.ac.il/~nogaa/PDFS/exp1.pdf`, retrieved September 21, 2022.

[ASS08] N. Alon, O. Schwartz, and A. Shapira. An Elementary Construction of Constant-Degree Expanders. *Combinatorics, Probability and Computing*, 17:319 – 327, 2008.

[BM11] M. Bläser and B. Manthey. Advanced Complexity Theory. Universität des Saarlandes. Draft, 2011.

[Bol88] B. Bollobás. The isoperimetric number of random regular graphs. *Europ. J. Combinatorics*, 9:241–244, 1988.

[BS80] I. N. Bronstein and K. A. Semendjajew. *Taschenbuch der Mathematik*. Harri Deutsch, Thun, 1980.

[Din05] I. Dinur. The PCP Theorem by Gap Amplification. In *Electronic Colloquium on Computational Complexity*. Weizmann Institute of Science, 2005. Revision 1 of Report No. 46.

[Fri21] A. Frieze. Edge Colouring, 2021. Lecture on Graph Theory at Carnegy Mellon University, Ch. 6, `https://www.math.cmu.edu/~af1p/Teaching/GT/CH6.pdf`, retrieved September 3, 2022.

[FS20] G. Fischer and B. Springborn. *Lineare Algebra*. Springer, Heidelberg, 2020.

[HLW06] S. Hoory, N. Linial, and A. Wigderson. Expander Graphs and their Applications. In *Bull. Amer. Math. Soc.*, volume 43(4), pages 439 – 561. American Mathematical Society, 2006.

[Hof14] D. W. Hoffmann. *Einführung in die Informations- und Codierungstheorie*. Springer, Heidelberg, 2014.

[K+88] O. Kerner et al. *Vieweg Mathematik Lexikon*. Vieweg, Braunschweig, 1988.

[KM21] C. Karpfinger and K. Meyberg. *Algebra*. Springer, Heidelberg, 2021.

[Moh89] B. Mohar. Isoperimetric Numbers of Graphs. *Journal of Combinatorial Theory*, Series B 47:274–291, 1989.

[Nic18] B. Nica. *A Brief Introduction to Spectral Graph Theory*. European Mathematical Society Publishing House, Zürich, 2018.

[Sil00] J. R. Silvester. Determinants of Block Matrices. *Mathematical Gazette*, 84(501):460–467, 2000.

[Sta17] Z. Stanic. *Regular Graphs. A Spectral Approach*. de Gruyter, Heidelberg, 2017.