Summer School in Mathematics 2013

Institute of Mathematics
Eötvös University
Budapest, Hungary, 2013

Contents

1 Ágnes Backhausz: Markov chains: from random walks to simulations **1**
- 1.1 Introduction . . . 1
- 1.2 Markov chains with finite state space . . . 1
- 1.3 Stationary distribution . . . 3
- 1.4 Algorithms based on Markov chains . . . 6
- 1.5 Exercises . . . 7
- 1.6 References . . . 8

2 Péter Csikvári and Tamás Héger: Strongly regular graphs – combinatorics and eigenvalues **11**
- 2.1 Combinatorics... . . . 11
- 2.2 ...and eigenvalues . . . 14
- 2.3 References . . . 21

3 István Faragó: Numerical methods for initial value problems **23**
- 3.1 Basic of the of the theory of initial-value problems . . . 23
- 3.2 Introduction into the one-step numerical methods . . . 24
- 3.3 Runge-Kutta method . . . 31
- 3.4 Multistep numerical methods . . . 35
- 3.5 Consistency of the general linear multistep methods . . . 36
- 3.6 Numerical solution of initial value problems with Matlab . . . 38
- 3.7 In-built Matlab programs . . . 40

4 Ferenc Izsák: Computer aided simulation of time dependent phenomena **47**
- 4.1 Introduction . . . 47
- 4.2 The basic principles in the numerical solution . . . 47
- 4.3 Motivation: a model problem and the chief questions . . . 48
- 4.4 Tools for the analysis . . . 51
- 4.5 Numerical methods for diffusion problems . . . 52
- 4.6 Numerical methods for advection problems . . . 53
- 4.7 Numerical methods for the one-dimensional wave-equations . . . 54
- 4.8 References . . . 55

5 András Frank: The graph orientation problem **57**
- 5.1 Introduction . . . 57
- 5.2 Degree-constrained orientations . . . 58
- 5.3 Applications . . . 64

5.4 References 68

6 Péter E. Frenkel: Algebraic inequalities and sums of squares 71
6.1 Inequalities between means 71
6.2 Positive semi-definite matrices 75
6.3 References 77

7 Tibor Jordán: Location and localization problems in networks 79
7.1 Introduction 79
7.2 Rigidity and global rigidity of graphs 81
7.3 Rigidity matrices and matroids 81
7.4 Warm up exercises 82
7.5 Appendix 83
7.6 References 84

8 Márton Naszódi: A Glimpse of Discrete Geometry 87
8.1 Borsuk's partitioning problem 87
8.2 Covering by translates of a convex body 87
8.3 Rigidity of Polyhedra 89
8.4 References 91

9 Dömötör Pálvölgyi: Algorithmic problems 93
9.1 Embedding planar graphs on a small grid 93
9.2 Online competitive algorithms 97
9.3 Probabilistically checkable proofs 100

10 Gergely Zábrádi: p-adic numbers and applications 103
10.1 Why p-adic numbers? 103
10.2 Solving equations in $\mathbb{Q}_p$ 105
10.3 Precise definition of $\mathbb{Q}_p$ 107
10.4 Towards irreducibility criteria for polynomials over $\mathbb{Q}$ 111
10.5 Applications, research directions, and further reading 116
10.6 References 119

11 András Zempléni: Extreme Value Modelling 121
11.1 Introduction 121
11.2 Extreme Value Theory 121
11.3 Statistical inference 128
11.4 References 130

©The copyright of each chapter rests with its respective author(s).

Chapter 1

Ágnes Backhausz: Markov chains: from random walks to simulations

1.1 Introduction

Imagine the following situation. A tourist is wandering around in a foreign city. At each corner he chooses a street at random to continue his walk along it. If he is so absent-minded that he always forgets where he has been before, then his random walk is a so-called Markov chain. We can ask about the probability that he returns to the starting place after some time, or the places that are the most likely to find him after a long time.

Loosely speaking, Markov chains are stochastic processes that forget their past: it is sufficient to know the actual state of the process to calculate the probabilities of future events. Since Markov chains have very weak memory, they are easier to handle from a theoretical point of view (especially when the number of possible states of the chain is finite); however, they are still flexible enough to have various applications both in real life examples (e.g. insurance) and in more complex problems of probability theory or combinatorics for example.

We start with some toy examples which lead to the concept of a Markov chain in the finite state case. We continue with asking what happens after a long time, what is stationary distribution and mixing time. Finally we have a look at the methods of computer simulations that are based on Markov chains.

1.2 Markov chains with finite state space

Before giving the definition of a Markov chain, let us see some other examples.

Exercise 1.2.1 (Gambler's ruin)**.** *Alice has A pounds, Bob has B. They repeat a fair game with 1 pound at stake. If one of them loses all his or her money, then the other one wins, and the game is finished. What is the probability that Alice is the winner if*

a) $A = 2, B = 2$*;* *b)* $A = 1, B = 3$*;* *c)* $A = 2, B = 3$*;* *d)* $A = 10, B = 5$*?*

Note that even in the last case, the probability that Alice wins depends always only on the actual state of the game.

Exercise 1.2.2 (Simple model for weather forecast)**.** *If a day is sunny, then the next day will be sunny with probability* $2/3$*; otherwise it is raining. Moreover, after a rainy day, it will rain on the following day with probability* $1/4$*; otherwise it will be sunny. The sun is shining today. What is the probability that it will rain a) tomorrow; b) the day after tomorrow; c) on Thursday; d) on the next Monday.*

Note again that according to this simple model, the probabilities of rain and sunshine depend only on the weather of the previous day. This is the basic concept of a Markov chain. We will use the following definition.

1.2.1 Definition and transition matrix

Definition 1.2.3 (Markov chain)**.** *Let* $(\Omega, \mathcal{A}, P)$ *be a probability space and let* I *be a finite set:* $I = \{i_1, \dots, i_s\}$*. The sequence of random variables* $(X_n)_{n=0}^{\infty}$ *with* $X_n : \Omega \to I$ *is a Markov chain if*

$$\mathbb{P}(X_{n+1} = i_{n+1} | X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = \mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n)$$

holds for all $n = 1, 2, \dots$ *and* $i_0, \dots, i_{n+1} \in I$ *for which the conditional probabilities exist.*

Example 1.2.4. *In Exercise* 1.2.1*,* $I = \{0, 1, \dots, 15\}$*,* X_n *denotes the amount of money of Alice after* n *turns. Then* $X_0 = A$ *and* (X_n) *is a Markov chain.*

Example 1.2.5. *In Exercise* 1.2.2*,* $I = \{\text{sun}, \text{rain}\}$*,* X_n *represents the weather of day* n*,* $X_0 = \text{sun}$*.*

Exercise 1.2.6 (Conditional independence)**.** *Show that for a Markov chain* (X_n) *the following holds:*

$$\begin{aligned}\mathbb{P}(X_{n+1} = i_{n+1}, \dots, X_{n+m} = i_{n+m} | X_n = i_n) \mathbb{P}(X_0 = i_0, \dots, X_{n-1} = i_{n-1} | X_n = i_n) \\ = \mathbb{P}(X_{n+1} = i_{n+1}, \dots, X_{n+m} = i_{n+m}, X_0 = i_0, \dots, X_{n-1} = i_{n-1} | X_n = i_n).\end{aligned}$$

That is, the future and the past are conditionally independent with respect to the present.

We will suppose that the behaviour of the chain is constant in time in terms of conditional probabilities.

Definition 1.2.7 (Homogenity in time)**.** *We say that a Markov chain* (X_n) *is time-homogeneous, if for all* $n \geq 1$ *and* $i, j \in I$ *such that* $\mathbb{P}(X_n = i) > 0$*, the conditional probability*

$$\mathbb{P}(X_{n+1} = j | X_n = i)$$

does not depend on n*. That is,* $\mathbb{P}(X_{n+1} = j | X_n = i) = p_{ij}$ *holds for some* p_{ij} *for all* n*.*

>From now on we always consider time-homogeneous Markov chains. Then p_{ij} is the probability that from state i we move to state j in one step. In Exercise 1.2.2, we have $p_{\text{sun,sun}} = 2/3$ for example. It may be also interesting to look further and ask what is the probability that from state i we arrive at state j in two steps, three or seven steps. For a time-homogeneous Markov chain this depends only on i, j and the number of steps.

Definition 1.2.8 (Transition matrix)**.** *For a time-homogeneous Markov chain* (X_n) *the* n *step transition probabilities are the following:*

$$p_{ij}^{(n)} = \mathbb{P}(X_n = j | X_0 = i) \qquad (i, j \in I,\ n \geq 1).$$

The n *step transition matrix consists of the* n *step transition probabilities:* $P^{(n)} \in \mathbb{R}^{s \times s}$ *with*

$$P_{uv}^{(n)} = p_{i_u, i_v}^{(n)} = \mathbb{P}(X_n = i_v | X_0 = i_u) \qquad (u, v = 1, \ldots, s,\ n \geq 1).$$

For $n = 0$,$P^{(0)}$ *is the identity matrix, and for* $P^{(1)}$ *we write simply* P*.*

Note that $P^{(n)}$ is a so-called stochastic matrix: the sums of elements in each row is equal to 1, and its entries are nonnegative. Moreover, time-homogenity implies that

$$p_{i_u, i_v}^{(n)} = \mathbb{P}(X_{n+m} = i_v | X_m = i_u)$$

holds for all $m \geq 0$ and $u, v = 1, \ldots, s$.

Exercise 1.2.9. *Give the* $1, 2$ *and* 3 *step transition matrices of the Markov chain of Exercise* 1.2.2*, and the (*1 *step) transition matrix of Exercise* 1.2.1*.*

1.2.2 Chapman–Kolmogorov equations

The following is a consequence of the law of total probability. When going from state i to j in $n + m$ steps, we ask where the chain is after n steps.

Proposition 1.2.10 (Chapman–Kolmogorov equations)**.** *For a time-homogeneous Markov chain* (X_n) *we have*

$$p_{ij}^{(n+m)} = \sum_{k \in I} p_{ik}^{(n)} p_{kj}^{(m)} \qquad (i, j \in I,\ n, m \geq 0,\ \mathbb{P}(X_0 = i) > 0).$$

This gives a simple way for calculating the n step transition matrix from the 1 step transition probabilities. Namely, the n step transition matrix is the nth power of the one step transition matrix.

Proposition 1.2.11. *For a time-homogeneous Markov chain* (X_n) *we have* $P^{(n)} = P^n$ *for every* $n \geq 0$*.*

1.3 Stationary distribution

We would like to describe the long-term behaviour of a Markov chain with finite state space. We ask whether the probability of being at state i is convergent, and if so, does the limit depend on the initial position of the chain. This will lead to the concept of the stationary distribution, which may be the limit of the distibution of X_n under certain conditions.

1.3.1 Existence and uniqueness

We consider a time-homogeneous Markov chain (X_n). We asked what is the probability that X_n is in state i after n steps, given X_0. As we have seen, the transition probabilities are easier to handle together. We do the same, and we build a row vector for each n that contains the probabilities of being at the possible states. This also allows us to describe the case when the Markov chain starts from a random position, that is, X_0 is a random variable.

Definition 1.3.1. *For a time-homogeneous Markov chain (X_n) we use the following notation. For each $n \geq 0$ we have $x^{(n)} \in \mathbb{R}^s$ with*

$$x_u^{(n)} = \mathbb{P}(X_n = i_u).$$

Exercise 1.3.2. *Show that $x_u^{(n)} = \sum_{v=1}^{s} x_v^{(n-1)} P_{vu}$, that is, $x^{(n)} = x^{(n-1)}P$, and hence $x^{(n)} = x^{(0)}P^n$.*

Suppose that the vectors $x^{(n)}$ converge elementwise as $n \to \infty$. This may give a motivation for the following definition.

Definition 1.3.3 (Stationary distribution)**.** *$\pi \in \mathbb{R}^s$ is a stationary distribution for a Markov chain with transition matrix P if (i) $0 \leq \pi_u \leq 1$ for $1 \leq u \leq s$; (ii) $\sum_{u=1}^{s} \pi_u = 1$; (iii) $\pi = \pi P$, that is,*

$$\pi_i = \sum_{k \in S} \pi_k P_{ki}.$$

To put it in another way, if the initial distribution of the Markov chain is the stationary distribution π, then the distribution of X_n is also π for all n.

Note that since the sum of elements in each row of P is equal to 1, we have $P\mathbf{1}^T = \mathbf{1}^T$, that is, the column vector with all entries equal to 1 is a right eigenvector of P with eigenvalue 1. The stationary distribution is a left eigenvector with eigenvalue 1.

Exercise 1.3.4. *Find a stationary distribution for the Markov chain of Exercise* 1.2.2.

Exercise 1.3.5 (Random walk on a graph)**.** *Let G be a simple connected graph (no loops and multiple edges) on s vertices. Let us consider the following Markov chain. The state space is the vertex set of G, and the transition probabilities are*

$$p_{uv} = \begin{cases} \frac{1}{\deg(u),} & \text{if } uv \text{ is an edge in } G; \\ 0; & \text{otherwise.} \end{cases}$$

Find a stationary distribution.

We will see that the following is a sufficient condition for the existence of the stationary distribution.

Definition 1.3.6 (Irreducible Markov chain)**.** *A Markov chain is irreducible if for all $i, j \in I$ there exists $n \geq 1$ such that $p_{ij}^{(n)} > 0$.*

That is, for two arbitrary states it is possible to go from one to the other one. The number of necessary steps may depend on the two states by definition.

Proposition 1.3.7 (Irreducible case, proposition 1.14. in [5]). *Let P be the transition matrix of an irreducible Markov chain. Then there exists a probability distribution π on I such that $\pi = \pi P$ and $\pi_i > 0$ for all $i \in I$. Moveover,*

$$\pi_i = \frac{1}{m_{ii}} \qquad (i \in J),$$

where m_{ii} is the expectation of $\tau_i = \inf\{n > 0 : X_n = i | X_0 = i)$.

To put it in another way, a finite irreducible Markov chain has a unique stationary distribution. We also see from this proposition that $\mathbb{E}(\tau_i)$ is finite. We remark that the uniqueness also follows from the fact that a harmonic function is constant in the irreducible case (see Section 1.5.4. of [5] for the details).

1.3.2 Convergence

Now we would like to see that the probability of being in state i after n steps converges to π_i as the number of steps goes to infinity. To state the theorem giving even the speed of convergence, we need another important definition for the following reason. Think of a Markov chain with $I = \{0, 1, \ldots, s-1\}$, where we always move one up or down, starting from 0. Then the probability of being at 0 after n steps is clearly 0 for odd n, while the limit of the sequence of probabilities for even n is positive. We exclude this kind of periodicity.

Definition 1.3.8 (Aperiodic Markov chain). *The period of state i is defined by*

$$d(i) = \gcd\{n \geq 1 : p_{ii}^{(n)} > 0\}.$$

A Markov chain is aperiodic if $d(i) = 1$ for all $i \in I$.

Theorem 1.3.9 (Theorem 4.9. in [5]). *Let X_n be a time-homogeneous, irreducible, aperiodic Markov chain with finite state space with transition matrix P and stationary distribution π. There exist constants $\alpha \in (0,1)$ and $C > 0$ such that*

$$\max_{i \in I} \left|\left| p_{i\cdot}^{(n)} - \pi \right|\right|_{TV} = \max_{i \in I} \frac{1}{2} \sum_{j \in I} \left| p_{ij}^{(n)} - \pi_j \right| \leq C\alpha^n.$$

Loosely speaking, this theorem states that the distribution after n steps converges exponentially fast in the total variation distance to the stationary distribution. This also means that a Markov chain "forgets" its initial distribution fast.

Exercise 1.3.10. *Consider a chess board consisting of 64 squares. A king is placed in one of the corners. Then at each step it moves one in a random direction. All possible directions are equally likely to be chosen. Let q_n be the probability that the king is a) in the same corner b) in an arbitrary corner after n steps. What is the limit of q_n as n tends to infinity?*

1.3.3 Mixing time

Theorem 1.3.9 states the existence of an upper bound on the distance of the distribution after n steps and the stationary distribution in terms of the number of steps. Now we consider the reverse problem. We fix the distance from the stationary distribution, and ask how many steps do we need to reach this. It will turn out that this is in connection with the second largest eigenvalue (in absolute value), or more precisely the spectral gap of the transition matrix.

Definition 1.3.11 (Mixing time)**.** *Consider a Markov chain with finite state space I, transition matrix P and stationary distribution π. Let $\mathcal{Q}$ be the set of probability distributions on I. Then we define*

$$D(n) = \sup\left\{\mu \in \mathcal{Q} : \left\|\mu P^{(n)} - \pi\right\|_{TV}\right\},$$

and for all $\varepsilon > 0$ the mixing time for ε is as follows.

$$t_{\text{mix}}(\varepsilon) = \min\{n \geq 1 : D(n) \leq \varepsilon\}.$$

In a special case we can give an upper bound on the mixing time.

Definition 1.3.12 (Reversibility)**.** *Consider a Markov chain with transition matrix P and stationary distribution π. We say that the chain is reversible with respect to the stationary distribution if*

$$\pi_i P_{ij} = \pi_j P ji$$

holds for all $i, j \in I$.

Exercise 1.3.13. *Find a Markov chain defined in one of the previous exercises that is reversible with respect to its stationary distribution.*

First we define the relaxation time which is based on the second largest eigenvalue of P, and then we give an upper bound on the mixing time using this quantity.

Definition 1.3.14 (Relaxation time)**.** *For a reversible transition matrix P, the relaxation time is defined by*

$$t_{\text{rel}} = \frac{1}{1 - \max\{|\lambda| : \lambda \text{ is an eigenvalue of } P, \lambda \neq 1\}}.$$

Theorem 1.3.15 (Mixing time and relaxation time, Theorem 12.3. in [5])**.** *For a reversible, irreducible Markov chain let $\pi_{\min} = \min_{i \in I} \pi_i$. Then we have*

$$t_{\text{mix}}(\varepsilon) \leq \log\left(\frac{1}{\pi_{\min}\varepsilon}\right) t_{\text{rel}}.$$

1.4 Algorithms based on Markov chains

It may happen that we would like to sample from a well-defined distribution but the system is too large or complicated to do this directly. For example, we would like to pick a web page at random such that each of them has the same probability to be chosen. Note that there are more than 13 billion of web pages. Typically we can start from a given one and then we can follow the links between them. Suppose that we also know who is linking to a given web page. Then we can forget about the direction of links, and we are in the situation given by Exercise 1.3.5. We could calculate the stationary distribution of the Markov chain given there, but it is clearly not uniform (unless the graph is regular).

The idea is that we define another Markov chain on the vertex set of G whose stationary distribution is uniform on the vertices. Then, starting from an arbitrary vertex or initial distribution, according to Theorem 1.3.9, we get close to

the uniform distribution quite fast. There are several ways to do this in various settings, see Chapter 3 of [5].

For example, the so-called Metropolis–Hastings algorithms follow a kind of a random walk but the suggested moves are accepted with a certain probability; if a move is not accepted, then they stay at the actual state. Usually it is not preferred to move to states that have larger weight in the stationary distibution than the actual state. It is not hard to calculate the appropriate probability that a move accepted in certain situation. For example, in the case of simple graphs we use the following transition probabilities to choose a vertex uniformly at random.

$$q_{ij} = \begin{cases} p_{ij}\alpha_{ij} & i \neq j; \\ p_i i + \sum_{j \neq i} p_{ij}(1 - \alpha_{ij}) & \text{otherwise.} \end{cases}$$

with

$$\alpha_{ij} = \min\left(\frac{|N(i)|}{|N(j)|}, 1\right),$$

where $N(i)$ is the number of neighbours (the degree) of i in the graph.

Another possibility is setting

$$\alpha_{ij} = \frac{1}{2}\frac{1}{|N(j)|}.$$

We may also want to choose a permutation of n elements uniformly at random by transpositions. Some kinds of Metropolis–Hastings algorithms may also be used here.

Markov chains may also be used to generate a uniform spanning tree or proper coloring of a graph. The idea is again that we define a Markov chain whose stationary distribution is the uniform distribution. So called Glauber dynamics are used in the latter case.

We mention that there are also methods based on Markov chains to sample from continuous or high dimensional distributions (see Appendix B of [5]).

1.5 Exercises

1. Alice and Bob are rolling two fair dices several times. If the sum is at most 3, then Alice wins. If the sum is 7, then Bob wins. Otherwise they throw the dices again. This is repeated until one of them wins. What is the probability that Alice wins the game?

2. A basketball player is practicing shots. If two consecutive shots are succesful, then the following one is succesful with probability 2/3. If only one of two consecutive shots is succesful, the the following one is succesful with probabiliy 1/2. If none of two consecutive shots are succesful, then his next attempt is succesful only with probability 1/4.

   Define a Markov chain describing the attempts of the player, and determine its transition matrix. Try guessing the ratio of succesful shots on a long term average.

3. Alice and Bob toss a fair coin several times. If the sequence HHHT comes before the sequence sequence HTHT, then Alice wins, otherwise Bob wins. What is the probability that Alice wins? What is the expected number of coin tosses in this game? (The game is finished when somebody wins.)

4. Alice, Bob and Charles play tennis. First Alice plays against Bob. They do the following. After each match, the winner plays with the one who did not participate in the last match. If somebody wins two times in a row, then the game is finished and he or she is the final winner. What is the probability that Alice is the final winner? What is the expected number of matches?

5. [Non backtracking random walk] We have a finite simple graph on s vertices. We start from an arbitrary vertex. At each step we go to a neighbouring vertex; we choose one of them uniformly at random, but the one where we have been in the previous step can not be chosen. Is the sequence of visited vertices is a Markov chain? Can we define a Markov chain that describes this process?

6. A driver can be in one of three classes with respect to the liability insurance: M, A, and B. M is the worst and B is the best. At each year the number of accidents he caused a) is 0 with probability $0,98$, and 1 or 2 both with probability 0,01; b) has Poisson distribution with expected value 0,01.

   If he does not cause any accident, then for the next year he moves one class upwards, or stays in B if he was there. If he causes an accident, then he moves one class downwards. But if he causes at least to accidents in a year, then he goes to M even from B.

   Determine the 1, 2, 3 step transition matrices of the Markov chain described above, and find the stationary distribution. What is the limit of probability of being in B as the number of years tends to infinity?

7. Are the following Markov chains irreducible? Determine the period $d(i)$ for $i \in I$ and a stationary distribution if possible.

   a) $I = \{0, 1, \dots, 100\}$. We start from 0. From 0 we go always to 1, from 100 we go always to 99. Otherwise we toss a fair coin, and if it is heads, we move one to the right, otherwise one to the left.

   b) $I = \{0, 1, \dots, 100\}$. We start from 0. From 0 we go always to 0, from 100 we go always to 99. Otherwise we toss a fair coin, and if it is heads, we move one to the right, otherwise one to the left.

8. Consider the following Markov chain with $I = \{0, 1\}$. The probability of staying at the actual state is p for all steps; these choices are independent. What is the transition matrix for $1, 2, 4, 64$ steps? Is this Markov chain reversible? What is its relaxation time?

1.6 References

[1] N. Alon, I. Benjamini, E. Lubetzky, S. Sodin: Non-backtracking random walks mix faster, *Commun. Contemp. Math.,* **9** no. 4, 585–603, 2007.

[2] N. Berger, C. Kenyon, E. Mossel, Y. Peres. Glauber dynamics on trees and hyperbolic graphs. *Probability Theory Related Fields*, **131** (3), 311-340, 2005.

[3] K-L. Chung. *Markov chains with stationary transition probabilities.* Springer, 1967.

[4] J. R. Norris. *Markov chains.* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge, 1998.

[5] D. A. Levin, Y. Peres, E. L. Wilmer. *Markov chains and mixing times*, American Mathematical Society, 2009.

Chapter 2

Péter Csikvári and Tamás Héger: Strongly regular graphs – combinatorics and eigenvalues

2.1 Combinatorics...

Throughout this lecture $G = (V, E)$ denotes a simple, undirected graph (so there are no multiple edges or loops in G) with vertex set V and edge set E. The *complete graph* (where every pair of vertices are connected by an edge) on n vertices is denoted by K_n. A graph is *empty* if it has no edges. If two vertices, u and v are connected by an edge, we call them *adjacent* or *neighbors*, and we may write $u \sim v$. A graph G is called *k-regular*, if every vertex of G has precisely k neighbors.

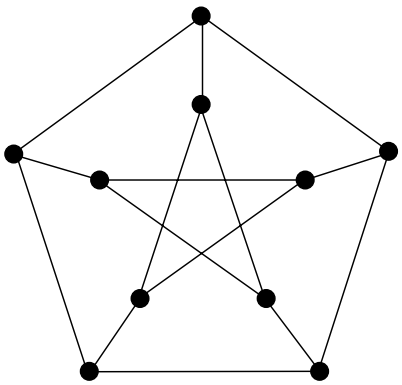


Figure 2.1: This is the famous Petersen-graph. It is 3-regular on 10 vertices. Moreover, if two vertices are adjacent, then they have no common neighbors; if two vertices are not adjacent, then they have exactly one common neighbor.

Definition 2.1.1. A graph G is called a *strongly regular graph* with parameters (n, k, λ, μ) (in notation: SRG(n, k, λ, μ)) if the following properties hold:

1. G has n vertices and G is k-regular;

2. if two distinct vertices are adjacent, then they have λ common neighbors;

3. if two distinct vertices are nonadjacent, then they have μ common neighbors;

4. G is not complete, nor empty (that is, $1 \leq k \leq n-2$).

For example, the Petersen-graph is an $\mathrm{SRG}(10,3,0,1)$. Sometimes the 4th point is omitted from the definition. Note that if we did not require this property, the parameters λ and μ would not be well defined; for example, the complete graph K_n would be an $\mathrm{SRG}(n, n-1, n-2, \mu)$ for arbitrary μ, since there are no non-adjacent vertices in K_n.

Exercise 2.1.2. *Determine which cycles are strongly regular, and determine their parameters.*

Exercise 2.1.3. *Show that if a k-regular bipartite graph is strongly regular, then either $k = 1$ (so the graph consists of independent edges) or it is isomorphic to $K_{k,k}$. ($K_{k,k}$ is the complete bipartite graph on $k+k$ vertices; that is, both vertex classes have k vertices and any two vertices from different classes are adjacent.)*

Exercise 2.1.4. *Show that the Petersen-graph is the unique* $\mathrm{SRG}(10,3,0,1)$.

Exercise 2.1.5. *Construct an* $\mathrm{SRG}(16,5,0,2)$ *and show that it is unique. (This graph is called the Clebsch-graph. Hint: the Petersen-graph is a subgraph of it.)*

Exercise 2.1.6. *Construct an* $\mathrm{SRG}(16,6,2,2)$.

Clearly there are some restrictions on the parameters of a strongly regular graph; for example, one must have $\lambda \leq k-1$ and $\mu \leq k$. Next we establish a connection among the parameters, which shows that any three of them determines the fourth.

Theorem 2.1.7. *Suppose that an* $\mathrm{SRG}(n,k,\lambda,\mu)$ *exists. Then*

$$k(k-1-\lambda) = (n-1-k)\mu.$$

Proof. Let u be an arbitrary vertex, and count the triplets $\{(u,v,w)\colon uv \in E,\ vw \in E,\ uw \notin E,\ u \neq w\}$. We may choose v in k different ways, and after that there are $k-1-\lambda$ suitable neighbors of v for the choice of w. Thus the number of such triplets is $k(k-1-\lambda)$. On the other hand, if we choose w first, then in a similar way we find that the number of such triplets is $(n-1-k)\mu$. □

Recall that the complement of a graph G has the same vertex set as G, and two vertices are adjacent in it if and only if they are not adjacent in G.

Theorem 2.1.8. *If G is an* $\mathrm{SRG}(n,k,\lambda,\mu)$, *then its complement, denoted by $\overline{G}$, is an* $\mathrm{SRG}(n,\bar{k},\bar{\lambda},\bar{\mu})$, *where $\bar{k} = n-k-1$, $\bar{\lambda} = n-2k+\mu-2$, $\bar{\mu} = n-2k+\lambda$.*

Proof. It is clear that $\overline{G}$ is $(n-k-1)$-regular. Let u and v be two adjacent vertices in $\overline{G}$. Then the number of vertices not adjacent to both u and v in G is $n-2k+\mu-2$, which is just the number of common neighbors of u and v in $\overline{G}$. Now suppose that u and v are non-adjacent in $\overline{G}$. Then, similarly, they have $n-2k+\lambda$ common neighbors in $\overline{G}$. □

Note that the above theorem yields further restrictions on the parameters: by $\bar{\lambda} \geq 0$ and $\bar{\mu} \geq 0$ we obtain $\mu \geq 2k-n+2$ and $\lambda \geq 2k-n$. Next we show that disconnected strongly regular graphs are not too interesting.

Theorem 2.1.9. *Suppose that G is a disconnected strongly regular graph. Then it is the union of some complete graphs of the same size.*

Proof. Let G be an $\mathrm{SRG}(n, k, \lambda, \mu)$ that is disconnected. Take two vertices from two distinct components. Then they cannot have a common neighbor, thus $\mu = 0$. Consider a connected component. If there were two vertices in it at distance at least two, then there were two vertices at distance exactly two, in contradiction with $\mu = 0$. Hence every component is a complete graph, namely K_{k+1}. We remark that $n = c \cdot (k + 1)$ for some integer $c \geq 2$, and $\lambda = k - 1$. □

Example 2.1.10. *Consider the graph on $2n$ vertices that consists of n independent edges (that is, the graph is the union of n disjoint K_2-s). This is called the* ladder *graph. Its complement (also strongly regular) is called the* cocktail party *graph.*

By Theorem 2.1.9, we see that it is enough to treat connected strongly regular graphs whose complement is also connected.

Exercise 2.1.11. *Consider the two element subsets of $\{1; 2; 3; 4; 5\}$ as vertices, and join two of them if and only if they are disjoint. Do you know this graph? (You do.)*

Example 2.1.12. *The lattice-graph $L(m)$ is defined as follows. Consider an $m \times m$ grid, whose m^2 points are the vertices of $L(m)$, and two vertices are adjacent if and only if they are in the same row or column. Formally, let $V = \{1, 2, \ldots, m\} \times \{1, 2, \ldots, m\}$, and (i, j) is adjacent to (i', j') if and only if $i = i'$ or $j = j'$. $L(m)$ is an $\mathrm{SRG}(m^2, 2(m-1), m-2, 2)$.*

As we have seen in the case of the Petersen-graph, sometimes the parameters of a strongly regular graph uniquely determine the graph, but this is not true in general.

Theorem 2.1.13. *For every $4 \neq m \geq 2$ the lattice graph $L(m)$ is the unique $\mathrm{SRG}(m^2, 2(m-1), m-2, 2)$.*

Proof. We prove the theorem for $m > 4$ only; see Exercise 2.1.15. Let v be a vertex of $L(4)$, and let $G(v)$ be the graph induced by the neighbors of v. It is clear that $G(v)$ has $2m - 2$ vertices and it is $(m-2)$-regular. Let u and z be two nonadjacent vertices of $G(v)$, and denote the number of their common neighbors by c. Clearly $c \leq 1$. The number of vertices not adjacent to u nor z in $G(v)$) is precisely c. Assume first that $c = 1$, then let w be the unique point of $G(v)$ that is not incident to u nor z. Again, w and u, and also w and z may have at most one common neighbor, so there are at least $m - 4 \geq 1$ neighbors of w that are not adjacent to u nor z. However, the only such vertex is w itself, thus $c = 0$ must hold. This means that $\overline{G(v)}$ does not contain a triangle. As $\overline{G(v)}$ is $(m-1)$-regular on $2(m-1)$ vertices, this implies $\overline{G(v)} \approx K_{m-1,m-1}$, so $G(v)$ is the union of two disjoint K_{m-1}s. Thus to every vertex v of $L(m)$ belong two K_m subgraphs, altogether $2m$ distinct copies. Two distinct K_m subgraphs may share at most one vertex, otherwise two of their common points would have more than $m - 2$ common neighbors. It is enough to show that the K_ms can be divided into two classes so that elements of each class partition the vertices of $L(4)$ (so the two classes correspond to the rows and columns). Let H be the graph whose vertices

are the K_m subgraphs, and two of them are adjacent if and only if they intersect in one vertex. Then H is triangle-free (otherwise we could find two adjacent vertices in $L(4)$ with more than $m-2$ common neighbors) and every vertex of H has degree at least m; thus H is isomorphic to $K_{m,m}$, whence the assertion follows. □

The next exercise shows that the above result does not hold if $m = 4$.

Exercise 2.1.14. *Let V be the vertex set of the lattice graph $L(4)$, and let S be the set of the four diagonal vertices. We define a new graph G (the* Shrikhande-graph*) on the set V. Let u, v be two distinct vertices of V. If $u \notin S$ and $v \notin S$, then uv is an edge in G if and only if uv is an edge in $L(4)$. If $u \in S$ and $v \notin S$, then uv is an edge in G if and only if uv is not an edge in $L(4)$. No two vertices of S are adjacent. Show that G is a strongly regular graph with the same parameter set as $L(4)$, but G is not isomorphic to $L(4)$.*

Exercise 2.1.15. *Verify Theorem 2.1.13 for $m = 2, 3$.*

Example 2.1.16. *The triangular graph $T(m)$ is defined as follows. Let the vertex set V of $T(m)$ be the set of two-element subsets of $\{1, 2, \ldots, m\}$, and let two of them be adjacent if their intersection is of size one. Then $T(m)$ is an* $\mathrm{SRG}(\frac{m(m-1)}{2}, 2(m-2), m-2, 4)$.

Note that by Exercise 2.1.11 we have that $T(5)$ is the complement of the Petersen-graph.

Exercise 2.1.17. *Consider the even element subsets of $\{1, 2, 3, 4, 5\}$ (including the empty set) and let two be adjacent if their symmetric difference has four elements. Prove that the arising graph is strongly regular. Do you know this graph?*

Exercise 2.1.18. *Construct an* $\mathrm{SRG}(35, 18, 9, 9)$.

Exercise 2.1.19. *Construct an* $\mathrm{SRG}(120, 56, 32, 28)$.

Exercise 2.1.20. *Let p be a prime such that $p \equiv 1 \pmod 4$. The Paley-graph $P(p)$ is defined in the following way: its vertex set is $\{0, 1, \ldots, p-1\}$, and two distinct vertices u and v are connected if and only if $u - v$ is a quadratic residue modulo p. (A number n is a quadratic residue modulo p if $n \equiv x^2 \pmod p$ for some integer x.) Prove that $P(p)$ is an* $\mathrm{SRG}(p, \frac{p-1}{2}, \frac{p-5}{4}, \frac{p-1}{4})$. *(Hint: use automorphisms; consider also $\overline{P(p)}$.)*

Exercise 2.1.21. *Is it possible to color the edges of K_{10} with three colors so that the edges of each color form a Petersen-graph?*

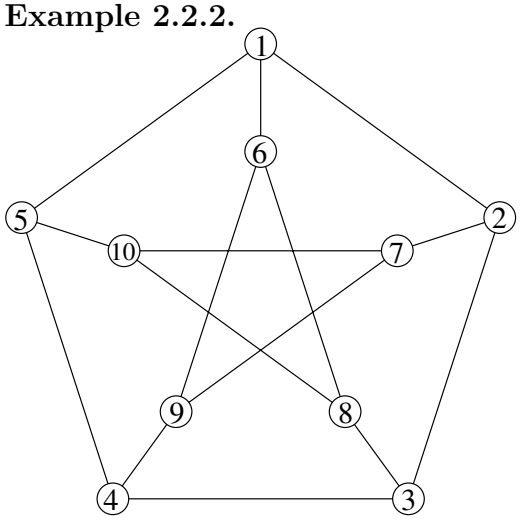
2.2 . . . and eigenvalues

Next we associate a matrix to a graph, which allows us to use linear algebraic techniques and results. Throughout $\mathbf{1}$ denotes the all-one vector (of suitable dimension), I is the identity matrix, J is the all-one matrix.

Definition 2.2.1. Let $G = (V, E)$ be a graph, and suppose that V has some ordering, $V = \{v_1, v_2, \ldots, v_n\}$. The *adjacency matrix* of G is the matrix $A \in \mathbb{R}^{n \times n}$, where $A_{ij} = 1$ if v_i and v_j are adjacent, and zero otherwise.

Note that the adjacency matrix of a graph is symmetric, and it has zeros in the diagonal.

Example 2.2.2.



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

The Petersen-graph and its adjacency matrix.

A graph is completely described by its adjacency matrix, so information on one of them gives information on the other one. We will examine the adjacency matrix of graphs, in particular the eigenvalues and the eigenvectors of it. First we consider some facts from linear algebra. Recall that the trace of a (square) matrix A is the sum of the entries on its diagonal.

Theorem 2.2.3. *Let $A \in \mathbb{R}^{n\times n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$. Then $\prod_{i=1}^{n} \lambda_i = \det(A)$ and $\sum_{i=1}^{n} \lambda_i = \mathrm{trace}(A)$.*

We remark that the trace of the adjacency matrix of a (loopless) graph is zero.

Theorem 2.2.4. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Then there is an orthonormal eigenbasis $v_1, \ldots, v_n$ of $\mathbb{R}^n$ with respect to A; that is,*

- *$v_1, \ldots, v_n$ is a basis of $\mathbb{R}^n$;*
- *$Av_i = \lambda_i v_i$ for some $\lambda_i \in \mathbb{R}$ $(1 \leq i \leq n)$;*
- *$v_i^T v_j = 0$ for all $1 \leq i < j \leq n$;*
- *$v_i^T v_i = 1$ for all $1 \leq i \leq n$.*

Note that the above theorem implies that a real symmetric matrix has real eigenvalues.

Definition 2.2.5. The *spectrum* of a matrix is the multiset of its eigenvalues. The *spectrum of a graph* is that of its adjacency matrix. If the matrix is of dimension $n \times n$, we usually order its eigenvalues as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. We may indicate the multiset of eigenvalues as a set in which the elements have an exponent, which refers to the multiplicity of the eigenvalue.

Exercise 2.2.6. *Show that the spectrum of a graph is the union of the spectra of its connected components.*

The next theorem is a consequence of the more general Frobenius–Perron theorem. We only formulate the results for adjacency matrices of graphs.

Theorem 2.2.7. *Let A be the adjacency matrix of a connected, undirected graph G. Then*

- *the largest eigenvalue λ_1 of A has multiplicity one;*
- *there is an eigenvector of A with eigenvalue λ whose components are positive;*
- *for the smallest eigenvalue λ_n we have $|\lambda_n| \leq \lambda_1$.*

Theorem 2.2.8. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix, and let λ be its largest eigenvalue. Then for all $u \in \mathbb{R}^n$ we have*

$$u^T Au \leq \lambda |u|^2.$$

Equality holds if and only if u is an eigenvector of A with eigenvalue λ.

Proof. Let $v_1, \dots, v_n$ be an orthonormal eigenbasis as in Theorem 2.2.4. Then $u = \sum_{i=1}^n \alpha_i v_i$ for some $\alpha_i \in \mathbb{R}$, and $|u|^2 = u^T u = \sum_{i,j} \alpha_i \alpha_j u_i^T u_j = \sum_{i=1}^n \alpha_i^2$. Thus

$$u^T Au = \left(\sum_{i=1}^n \alpha_i v_i^T\right) A \left(\sum_{j=1}^n \alpha_j v_j\right) = \sum_{i=1}^n \alpha_i v_i^T \sum_{j=1}^n \alpha_j A v_j =$$

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \lambda_j v_i^T v_j = \sum_{i=1}^n \alpha_i^2 \lambda_i \leq \lambda_1 |u|^2.$$

Equality holds if and only if $\lambda_i < \lambda_1$ implies $\alpha_i = 0$, thus u is in the subspace generated by the eigenvectors with eigenvalue λ_1. □

Theorem 2.2.9. *Let G be a graph with average degree $\bar{d}$ and maximum degree Δ. Then $\bar{d} \leq \lambda_1 \leq \Delta$.*

Proof. Let e denote the number of edges in G. Then $\bar{d} = 2e/n$. Suppose that $Av = \lambda v$, $v = (v_1, \dots, v_n) \neq \mathbf{0}$. We may assume that $v_1 \geq v_2 \geq \cdots \geq v_n$ and $v_1 > 0$ (as $-v$ is also an eigenvector). Then $\lambda v_1 = (Av)_1 \leq \Delta v_1$. On the other hand, Theorem 2.2.8 yields $2e = \mathbf{1}^T A \mathbf{1} \leq \lambda_1 n$. □

Theorem 2.2.10. *A graph is regular if and only if $\mathbf{1}$ is an eigenvector of its adjacency matrix. The eigenvalue of $\mathbf{1}$ is the common degree of the vertices.*

Proof. Trivial. □

One may think of an eigenvector and the corresponding eigenvalue of a graph in the following way. Let A be the adjacency matrix of the graph G on n vertices and let v be an eigenvector of A with eigenvalue λ; that is, $Av = \lambda v$. For any $1 \leq i \leq n$ the ith coordinate of the left-hand-side is $(Av)_i = \sum_{v_k \in V:\, v_k \sim v_i} v_k$, while the ith coordinate of the right-hand-side is λv_i. So if we write the entries of the eigenvector v on the corresponding vertices of G, and then replace every entry by the sum of the entries on the neighboring vertices (in the same time), then we get the original value multiplied by λ *on all vertices*. For an illustration, see Figure 2.2.

As an illustration, we give a characterization of bipartite graphs in terms of their spectrum. Recall that $v^T Av \leq \lambda_1 |v|^2$ for all vectors v.

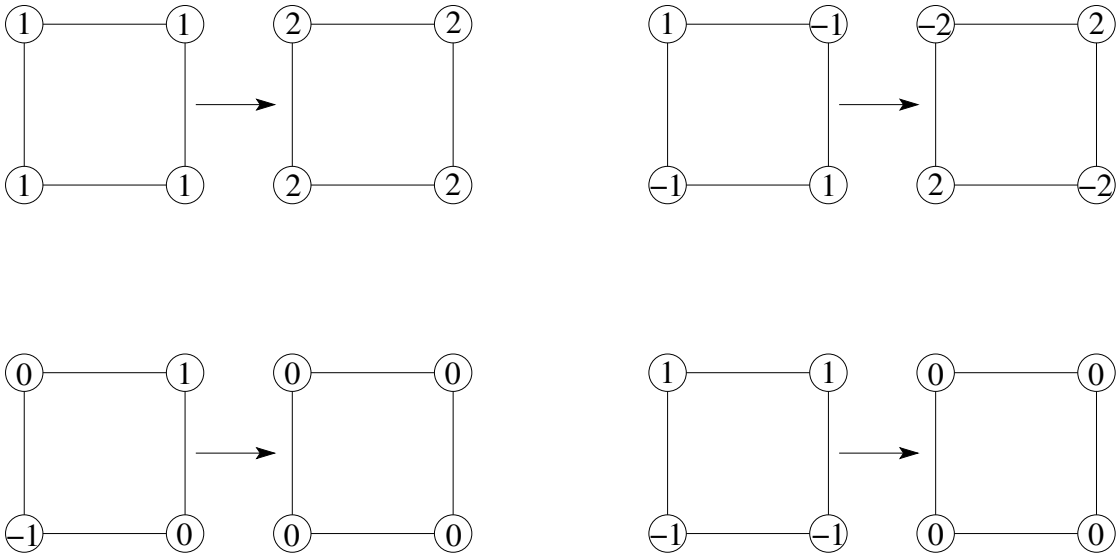


Figure 2.2: The cycle of length four has spectrum $\{1^1, 0^2, -1^1\}$. On the left part we depicted the eigenvector, on the right part we depicted the result after adding up the entries of the neighbors. Ordering the vertices from the top-left corner clockwise, the four eigenvectors are $(1;1;1;1)$, $(1;-1;1;-1)$, $(0;1;0;-1)$, $(1;1;-1;-1)$.

Theorem 2.2.11. *Let G be a graph on n vertices, and let A be its adjacency matrix. Then the following hold.*

- *G is bipartite if and only if the spectrum of A is symmetric (that is, if λ is an eigenvalue of A with multiplicity m, then $-\lambda$ is also an eigenvalue of A with multiplicity m).*
- *G is bipartite if and only if $\lambda_1 = -\lambda_n$.*

Proof. Suppose that G is bipartite on $n + m$ vertices, where the two classes have n and m vertices, respectively, and let A be its adjacency matrix. Let $v = (v_1, \ldots, v_{n+m})$ be an eigenvector of A with eigenvalue λ. By a proper ordering we may assume that the first n coordinates correspond to the vertices of first vertex class. Let $\bar{v} = (-v_1, \ldots, -v_n, v_{n+1}, \ldots, v_{n+m})$. Then $\bar{v}$ is also an eigenvector of A with eigenvalue $-\lambda$, hence the spectrum of A is symmetric.
Now suppose that $\lambda_1 = -\lambda_n$. Then λ_1 and λ_n are eigenvalues of the same component by the Frobenius–Perron theorem, thus we may assume that our graph is connected. Let v be an eigenvector of length $|v| = 1$ with eigenvalue λ_n, and let the vector u be defined by $u_i = |v_i|$ $(1 \le i \le n + m)$. Then also $|u| = 1$. As

$$\lambda_n = v^T A v = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} v_i v_j,$$

we have

$$\lambda_1 = |\lambda_n| = \left| \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} v_i v_j \right| \le \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} |v_i| |v_j| = u^T A u \le \lambda_1.$$

It the second estimate equality holds if and only if u is an eigenvector with eigenvalue λ_1 (Theorem 2.2.8). By the Frobenius–Perron theorem we have that all components of u are positive. As equality holds in the first estimate (triangle-inequality), either $v_iv_j = |v_i||v_j|$ or $v_iv_j = -|v_i||v_j|$ for all pairs i and j such that the corresponding vertices are adjacent. As $\lambda_n < 0$, the second option holds. Thus two vertices may be adjacent only if the corresponding components of v have different sign; that is, the signs of the components of v yield a bipartition. □

If v is an eigenvector of A with eigenvalue λ, then v is also an eigenvector of A^k with eigenvalue λ^k.

Theorem 2.2.12. *$(A^m)_{ij}$ is the number of walks of length m from v_i to v_j.*

Proof. By induction. The cases $m = 0, 1$ are trivial. (Recall that $A^0 = I$ by definition.) We prove the theorem by induction on m. Now

$$(A^m)_{ij} = (A^{m-1}A)_{ij} = \sum_{k=1}^{m} (A^{m-1})_{ik}A_{kj} = \sum_{k:\ v_k \in N(v_j)} (A^{m-1})_{ik},$$

which (by the inductive hypothesis) is the number of walks of length $m-1$ from v_i to some neighbor of v_j, which is just the number of walks of length m from v_i to v_j. □

We give another proof of Theorem 2.2.11.

Theorem 2.2.13. *Let G be a graph on n vertices, and let A be its adjacency matrix. Then the following hold.*

- *G is bipartite if and only if the spectrum of A is symmetric (that is, if λ is an eigenvalue of A with multiplicity m, then $-\lambda$ is also an eigenvalue of A with multiplicity m).*
- *G is bipartite if and only if $\lambda_1 = -\lambda_n$.*

Proof. Let G have n vertices. G is bipartite if and only if the number of closed walks of length m in G is zero for all odd integer m, or equivalently, if $\mathrm{trace}(A^m) = 0$ for all odd integer m. This holds if and only if $s_m := \sum_{i=1}^{n} \lambda_i^m = 0$ for all odd integer m. Suppose, say, $\lambda_1 > -\lambda_n$. Then $\lim_{k\to\infty} s_{2k+1} = \infty$. Thus $\lambda_1 = -\lambda_n$. After that, $\lambda_2 = -\lambda_{n-1}$ also follows etc. □

Now let us examine the spectrum of strongly regular graphs.

Theorem 2.2.14. *Let G be a graph, and let A be its adjacency matrix. Then the following are equivalent:*

1. *G is an $\mathrm{SRG}(n, k, \lambda, \mu)$;*
2. *$A^2 + (\mu - \lambda)A - (k - \mu)I = \mu J$.*

Proof. The entry $(A^2)_{ij}$ is the inner product of the vectors corresponding to v_i and v_j, which is just the number of common neighbors of v_i and v_j. Thus G is an $\mathrm{SRG}(n, k, \lambda, \mu)$ if and only if this quantity is k if $i = j$; λ if v_i and v_j are adjacent; and μ otherwise. In other words, $A^2 = kI + \lambda A + \mu(J - I - A)$, which is equivalent to the formula stated. □

Theorem 2.2.15. *Let G be an* $\mathrm{SRG}(n,k,\lambda,\mu)$*, and let A be its adjacency matrix. Then*

1. *the spectrum of A is $\{k^1, r^f, s^g\}$, where $r > s$ ($r = k$ may occur);*
2. *$rs = \mu - k$ and $r + s = \lambda - \mu$;*
3. *$f, g = \frac{1}{2}\left(n - 1 \pm \frac{(n-1)(\mu-\lambda) - 2k}{\sqrt{(\mu-\lambda)^2 + 4(k-\mu)}}\right)$ are non-negative integers.*

Proof. It is clear that $\mathbf{1}$ is an eigenvector with eigenvalue k. Let v be an eigenvector of A with eigenvalue x and $\mathbf{1}^T v = 0$. Then $Jv = 0$. As $A^2+(\mu-\lambda)A-(k-\mu)I = \mu J$, we obtain

$$x^2 v + (\mu - \lambda)xv - (k - \mu)v = 0,$$

thus $x^2 + (\mu - \lambda)x - (k - \mu) = 0$. Thus

$$x = \frac{\lambda - \mu \pm \sqrt{((\mu-\lambda)^2 + 4(k-\mu))}}{2}.$$

The two roots r and s are different as $(\mu - \lambda)^2 = 4(\mu - k)$ would contradict $\mu \leq k$ and $\lambda \leq k - 1$. Thus the first two points follow. As $f + g = n - 1$ and $\mathrm{trace}(A) = k + fr + gs = 0$, the last assertion also can be obtained easily. □

The third point of the above theorem is a strong restriction on the parameters of strongly regular graphs, and it is called the integrality or rationality condition.

Exercise 2.2.16. *Let G be an* $\mathrm{SRG}(n,k,\lambda,\mu)$ *with three distinct eigenvalues, $k > r > s$. Show that $(k - r)(k - s) = n\mu$.*

Exercise 2.2.17. *Let G be an* $\mathrm{SRG}(n,k,\lambda,\mu)$*. Show that either $(n,k,\lambda,\mu) = (4t + 1, 2t, t - 1, t)$ for some integer t or the eigenvalues of G are integral. (An* $\mathrm{SRG}(4t + 1, 2t, t - 1, t)$ *is called a* conference graph*.)*

Exercise 2.2.18. *Let G be an* $\mathrm{SRG}(n,k,\lambda,\mu)$*, where $n = p$ is a prime. Show that G is a conference-graph.*

Exercise 2.2.19. *We are about to show that the edges of K_{10} cannot be partitioned into three Petersen-graphs in terms of their adjacency matrices: the adjacency matrix of K_{10} is $J - I$, and our aim is to show that it cannot be expressed as $A + B + C$, where A, B and C are adjacency matrices of Petersen-graphs.*

- *Show that the eigenvalue 1 of a Petersen-graph has multiplicity five.*
- *Show that the eigensubspaces belonging to the eigenvalue 1 in two edge-disjoint Petersen-graphs intersect nontrivially. (Hint: there is a 9-dimensional subspace containing both.)*
- *Show that if A and B are the adjacency matrices of two edge-disjoint Petersen-graphs, then -3 is an eigenvalue of C, so C is not the adjacency matrix of a Petersen-graph.*

2.2.1 The Hoffman–Singleton theorem

In the sequel we treat the famous Hoffman–Singleton theorem on strongly regular graphs of girth five, that is, SRGs with $\lambda = 0$ and $\mu = 1$. (The girth of a graph is the length of the shortest cycle in it.) Note that Theorem 2.1.7 yields $n = k^2 + 1$ for this case.

Theorem 2.2.20 (Hoffman–Singleton)**.** *Let G be an* SRG$(n, k, 0, 1)$*. Then $k = 2, 3, 7$ or 57.*

Proof. By the integrality condition we have that

$$\frac{1}{2}\left(n - 1 \pm \frac{(n-1)(\mu-\lambda) - 2k}{\sqrt{(\mu-\lambda)^2 + 4(k-\mu)}}\right) = \frac{1}{2}\left(k^2 \pm \frac{k^2 - 2k}{\sqrt{4k-3}}\right)$$

are non-negative integers. Then either $k^2 - 2k = 0$, thus $k = 2$, or $\sqrt{4k-3}$ is an integer dividing $k(k-2)$. Then $4k-3$ divides $k^2(k-2)^2$, so it also divides $256k^2(k-2)^2 - (64k^3 - 208k^2 + 100k + 75)(4k-3) = 225 = 3^2 \cdot 5^2$. As $4k-3$ is a square, $4k - 3 \in \{9; 25; 225\}$ follows, which proves the assertion. □

For $k = 2$ and 3, the unique SRG$(k^2+1, k, 0, 1)$ graphs are the pentagon and the Petersen-graph. For $k = 7$ we will show a construction of an SRG$(50, 7, 0, 1)$, which is called the Hoffman–Singleton-graph. The existence of an SRG$(3250, 57, 0, 1)$ is still an open question.

Exercise 2.2.21. *Let G be a k-regular graph of girth five. Show that G has at least $k^2 + 1$ vertices, and in case of equality it is strongly regular.*

2.2.2 The Hoffman–Singleton-graph

The next construction is due to Robertson. Let P_m be a pentagon, and let Q_x be a pentagram as seen in Figure 2.3, $0 \le m \le 4$, $0 \le x \le 4$. Let the vertex labeled b of P_m be denoted by the pair $[m, b]$, and let the vertex labeled y of Q_x be denoted by (x, y). Besides the edges of the pentagons and the pentagrams, add an edge between (x, y) and $[m, b]$ if and only if $y \equiv mx + b \pmod 5$. It is clear that there is precisely one edge between any pentagon and pentagram, so the resulting graph is 7-regular. It is also clear that the graph does not contain any triangle. Suppose that we have a quadrangle. Then its four vertices are of form $[m_1, b_1]$, $[m_2, b_2]$, $m_1 \neq m_2$, (x_1, y_1), (x_2, y_2), $x_1 \neq x_2$, where

$$\begin{aligned} y_1 &\equiv m_1x_1 + b_1 \pmod 5 && (2.1)\\ y_2 &\equiv m_1x_2 + b_1 \pmod 5 && (2.2)\\ y_1 &\equiv m_2x_1 + b_2 \pmod 5 && (2.3)\\ y_2 &\equiv m_2x_2 + b_2 \pmod 5. && (2.4) \end{aligned}$$

Then $(2.1) - (2.2) - (2.3) + (2.4) \equiv 0 \pmod 5$, thus

$$(m_1 - m_2)(x_1 - x_2) \equiv 0 \pmod 5,$$

a contradiction.

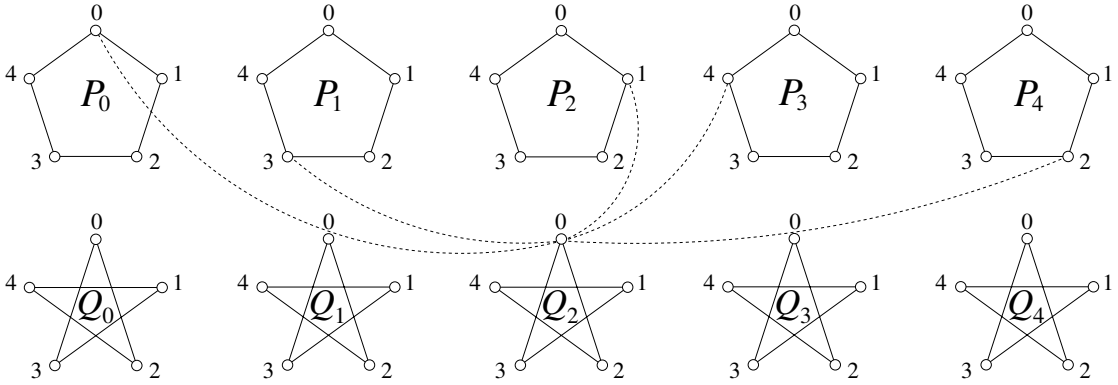


Figure 2.3: The pentagons and the pentagrams in Robertson's construction for the Hoffman–Singleton-graph.

Exercise 2.2.22. *Let P be* any *subgraph of the Hofmann–Singleton-graph isomorphic to the Petersen-graph. Show that each vertex not in P has exactly one neighbor in P.*

Exercise 2.2.23. *Let F be a subset of the vertices of the Hoffman–Singleton-graph that span an empty graph. Show that $|F| \leq 15$ and if $|F| = 15$ then each vertex not in F has precisely three neighbors in F.*

2.3 References

[1] A. E. Brouwer, A. M. Cohen, A. Neumaier: *Distance-Regular Graphs*. Springer-Verlag, New York, 1989, 485 pp., ISBN 3-540-50619-5

[2] A. E. Brouwer, W. H. Haemers: *Spectra of graphs*. Springer-Verlag, New York, 2012, ISBN 978-1-4614-1938-9.

[3] C. Godsil: *Algebraic combinatorics*. Chapman and Hall Mathematics Series. New York: Chapman and Hall, 1993. pp. xvi+362. ISBN 0-412-04131-6.

[4] C. Godsil, G. F. Royle: *Algebraic graph theory*. Springer-Verlag, New York, 2001, 462 pp., ISBN 0-387-95220-9

Chapter 3

István Faragó: Numerical methods for initial value problems

The first part of the textbook is concerned with initial value problems for scalar and systems of ordinary differential equations.

Since we have no hope of solving the vast majority of differential equations in explicit, analytic form, the design of suitable numerical algorithms for accurately approximating solutions is essential. The ubiquity of differential equations throughout mathematics and its applications has driven the tremendous research effort devoted to numerical solution schemes, some dating back to the beginnings of the calculus. Nowadays, one has the luxury of choosing from a wide range of excellent software packages that provide reliable and accurate results for a broad range of systems, at least for solutions over moderately long time periods. However, all of these packages, and the underlying methods, have their limitations, and it is essential that one be able to to recognize when the software is working as advertised, and when it produces spurious results!

3.1 Basic of the of the theory of initial-value problems

For simplicity, in general we will investigate the numerical methods for the scalar case, where $d = 1$. Then the formulation of the problem is as follows.

Let $Q_T := [0, T] \times \mathbb{R} \subset \mathbb{R}^2$, $f : Q_T \to \mathbb{R}$. The problem of the form

$$\frac{du}{dt} = f(t, u), \quad u(0) = u_0 \tag{3.1.1}$$

will be called *initial value problem*,or, alternatively, *Cauchy problem* (We do not emphasize the scalar property.) We always assume that for the given function $f \in C(Q_T)$ the Lipschitz condition

$$|f(t, u_1) - f(t, u_2)| \leq L\,|u_1 - u_2|\,, \quad \forall (t, u_1), (t, u_2) \in Q_T, \tag{3.1.2}$$

is satisfied. Moreover, $u_0 \in \mathbb{R}$ is a given number. Hence, our task is to find a sufficiently smooth function $u : [0, T] \to \mathbb{R}$ such that the relations

$$\frac{du(t)}{dt} = f(t, u(t)), \quad \forall t \in [0, T], \quad u(0) = u_0 \tag{3.1.3}$$

hold.

3.2 Introduction into the one-step numerical methods

Our aim is the numerical solution of the problem

$$\frac{du}{dt} = f(t,u), \quad t \in [0,T], \tag{3.2.1}$$

$$u(0) = u_0 \tag{3.2.2}$$

where $T > 0$ is such that the initial value problem (3.2.1)–(3.2.2) has a unique solution on the interval $[0,T]$. This means that we want to approximate the solution of this problem at a finite number of points of the interval $[0,T]$, denoted by $\{t_0 < t_1 < \cdots < t_N\}$. [1] In the sequel we consider those methods where the value of the approximation at a given time-point t_n is defined only by the approximation at the time-point t_{n-1}. Such methods are called *one-step methods*.

3.2.1 The Taylor method

This is one of the oldest methods. By definition, the solution $u(t)$ of the Cauchy problem satisfies the equation (3.2.1), which results in the equality

$$u'(t) = f(t,u(t)), \quad t \in [0,T]. \tag{3.2.3}$$

We assume that f is an analytical function, therefore it has partial derivatives of any order on the set Q_T. Hence, by use of the chain rule, by differentiation of the identity (3.2.3), at some point $t^\star \in [0,T]$ we get the relation

$$\begin{aligned} u'(t^\star) &= f(t^\star, u(t^\star)), \\ u''(t^\star) &= \partial_1 f(t^\star, u(t^\star)) + \partial_2 f(t^\star, u(t^\star))\ u'(t^\star), \\ u'''(t^\star) &= \partial_{11} f(t^\star, u(t^\star)) + 2\partial_{12} f(t^\star, u(t^\star))\ u'(t^\star) + \partial_{22} f(t^\star, u(t^\star))\ (u'(t^\star))^2 + \\ &\quad + \partial_2 f(t^\star, u(t^\star))\ u''(t^\star). \end{aligned} \tag{3.2.4}$$

Let us notice that knowing the value $u(t^\star)$ all derivatives can be computed exactly.

Hence the following numerical methods can be defined.

a) Taylor method

Let us select $t^\star = 0$, where the initial condition is given.[2]

Then the value $u(t^\star) = u(0)$ is known from the initial condition, and, based on the formula (3.2.4), the derivatives can be computed *exactly* at this point. Hence, using the usual Taylor approximation„ we have

$$u(t) \simeq \sum_{k=0}^{p} \frac{u^{(k)}(0)}{k!} t^k, \tag{3.2.5}$$

where, based on (3.2.4), the values $u^{(k)}(0)$ can be computed.

[1] We mention that, based on these approximate values, using some interpolation method we can define some approximation at any point of the interval $[0,T]$.

[2] According to Section 3.1, the derivatives do exist at the point $t = 0$.

b) Local Taylor method

We consider the following algorithm.

1. On the interval $[0, T]$ we define the points $t_0, t_1, \dots t_N$, which define the *mesh* $\omega_h := \{0 = t_0 < t_1 < \dots < t_{N-1} < t_N = T\}$. The distances between two neighbouring mesh-points, i.e., the values $h_i = t_{i+1} - t_i$, (where $i = 0, 1, \dots N-1$,) are called *step-size*, while $h = \max_i h_i$ denotes the measure of the mesh. (In the sequel, we define the approximation at the mesh-points, and the approximations to the exact values $u(t_i)$ will be denoted by y_i, while the approximations to the k-th derivatives $u^{(k)}(t_i)$ will be denoted by $y_i^{(k)}$, where $k = 0, 1, \dots, p$.[3]
2. The values $y_0^{(k)}$ for $k = 0, 1, \dots, p$ can be defined *exactly* from the formula (3.2.4), by substituting $t^\star = 0$.
3. Then, according to the formula

$$y_1 = \sum_{k=0}^{p} \frac{y_0^{(k)}}{k!} h_0^k, \tag{3.2.6}$$

we define the approximation to $u(t_1)$.

4. For $i = 1, 2, \dots, N-1$, using the values y_i, by (3.2.4) we define *approximately* $y_i^{(k)}$ (for $k = 0, 1, \dots, p$), by the substitution $t^\star = t_i$ and $u(t^\star) = u(t_i) \approx y_i$.
5. Using the formula

$$y_{i+1} = \sum_{k=0}^{p} \frac{y_i^{(k)}}{k!} h_i^k, \tag{3.2.7}$$

we define the approximation to $u(t_{i+1})$.

Using (3.2.7), let us define the algorithm of the local Taylor method for the special cases $p = 0, 1, 2$!

- For $p = 0$, $y_i = y_0$ for each values of i. Therefore this case is not interesting, and we will not investigate it.
- For $p = 1$ we have

$$y_{i+1} = y_i + y_i' h_i = y_i + h_i f(t_i, y_i), \quad i = 0, 1, \dots N-1, \tag{3.2.8}$$

where $y_0 = u_0$ is given.

- For $p = 2$ we have

$$y_{i+1} = y_i + h_i y_i' + \frac{h_i^2}{2} y_i'' = y_i + h_i f(t_i, y_i) + \frac{h_i^2}{2} \left(\partial_1 f(t_i, y_i) + \partial_2 f(t_i, y_i) f(t_i, y_i)\right), \tag{3.2.9}$$

where $i = 0, 1, \dots N-1$, and y_0 is given.

[3] As usual, the zero-th derivative ($k = 0$) denotes the function.

Example 3.2.1. *We consider the Cauchy problem*

$$\begin{aligned} u' &= -u + t + 1, \quad t \in [0,1], \\ u(0) &= 1. \end{aligned} \tag{3.2.10}$$

The exact solution is $u(t) = \exp(-t) + t$.

In this problem $f(t,u) = -u + t + 1$, therefore

$$\begin{aligned} u'(t) &= -u(t) + t + 1, \\ u''(t) &= -u'(t) + 1 = u(t) - t, \\ u'''(t) &= -u(t) + t, \end{aligned} \tag{3.2.11}$$

i.e., $u(0) = 1$, $u'(0) = 0$, $u''(0) = 1$, $u'''(0) = -1$. The global Taylor method results in the following approximation polynomials:

$$\begin{aligned} T_{1,u}(t) &= 1, \\ T_{2,u}(t) &= 1 + t^2/2, \\ T_{3,u}(t) &= 1 + t^2/2 - t^3/6. \end{aligned} \tag{3.2.12}$$

Hence, at the point $t = 1$ we have $T_{1,u}(1) = 1$, $T_{2,u}(1) = 1.5$, $T_{3,u}(1) = 1.333$). (We can also easily define the values $T_{4,u}(1) = 1.375$ and $T_{5,u}(1) = 1.3666$.) As we can see, these values approximate the value of the exact solution $u(1) = 1.367879$ only for larger values of n.

Let us apply now the local Taylor method taking into account the derivatives under (3.2.11). The algorithm of the first order method is

$$y_{i+1} = y_i + h_i(-y_i + t_i + 1), \quad i = 0, 1, \ldots, N-1, \tag{3.2.13}$$

while the algorithm of the second order method is

$$y_{i+1} = y_i + h_i(-y_i + t_i + 1) + \frac{h_i^2}{2}(y_i - t_i), \quad i = 0, 1, \ldots, N-1,$$

where $h_1 + h_2 + \ldots + h_N = T$. In our computations we have used the step-size $h_i = h = 0.1$. In Table 3.2.1 we compared the results of the global and local Taylor methods at the mesh-point of the interval $[0,1]$. (LT1 and LT2 mean the first and second order local Taylor method, while T1, T2 and T3 are the first, second and third order Taylor methods, respectively.)

Using some numerical method, we can define a numerical solution at the mesh-points of the grid. Comparing the numerical solution with the exact solution, we define the error function, which is a grid function on the mesh on which the numerical method is applied. This error function (which is a vector) can be characterized by the maximum norm. In Table 3.2.2 we give the magnitude of the maximum norm of the error function on the meshes for decreasing step-sizes. We can observe that by decreasing h the maximum norm is strictly decreasing for the local Taylor method, while for the global Taylor method the norm does not change. (This is

t_i	the exact solution	LT1	LT2	T1	T2	T3
0.1	1.0048	1.0000	1.0050	1.0000	1.0050	1.0048
0.2	1.0187	1.0100	1.0190	1.0000	1.0200	1.0187
0.3	1.0408	1.0290	1.0412	1.0000	1.0450	1.0405
0.4	1.0703	1.0561	1.0708	1.0000	1.0800	1.0693
0.5	1.1065	1.0905	1.1071	1.0000	1.1250	1.1042
0.6	1.1488	1.1314	1.1494	1.0000	1.1800	1.1440
0.7	1.1966	1.1783	1.1972	1.0000	1.2450	1.1878
0.8	1.2493	1.2305	1.2500	1.0000	1.3200	1.2347
0.9	1.3066	1.2874	1.3072	1.0000	1.4050	1.2835
1.0	1.3679	1.3487	1.3685	1.0000	1.5000	1.3333

Table 3.2.1: Comparison of the local and global Taylor methods on the mesh with mesh-size $h = 0.1$.

mesh-size	LT1	LT2	T1	T2	T3
0.1	$1.92e-02$	$6.62e-04$	0.3679	0.1321	0.0345
0.01	$1.80e-03$	$6.12e-06$	0.3679	0.1321	0.0345
0.001	$1.85e-04$	$6.14e-08$	0.3679	0.1321	0.0345
0.0001	$1.84e-05$	$6.13e-10$	0.3679	0.1321	0.0345

Table 3.2.2: Maximum norm errors for the local and global Taylor methods for decreasing mesh-size h.

a direct consequence to the fact that the global Taylor method is independent of the mesh-size.)

The local Taylor method is a so-called *one-step method* (or, alternatively, *two-level method*). This means that the approximation at the time level $t = t_{i+1}$ is defined with the approximation obtained at the time level $t = t_i$ only. The error analysis is rather complicated. As the above example shows, the difference between the exact solution $u(t_{i+1})$ and the numerical solution y_{i+1} is caused by several reasons.

- The first reason is the *local truncation error*, which is due to the replacement of the Taylor series by the Taylor polynomial, assuming that we know the exact value at the point $t = t_i$. The order of the difference on the interval $[t_i, t_i + h_i]$, i.e., the order of magnitude of the expression $u(t) - T_{n,u}(t)$ defines the *order of the local error*. When this expression has the order $\mathcal{O}(h_i^{p+1})$, then the method is called p-th order.

- When we solve our problem on the interval $[0, t^\star]$, then we consider the difference between the exact solution and the numerical solution at the point $t^\star$. We analyze the error which arises due to the first two sources, and it is called *global error*. Intuitively, we say that some method is convergent at some fixed point $t = t^\star$ when by approaching zero with the maximum step-size of the mesh the global error at this point tends to zero. The order of the convergence of this limit to zero is called *order of convergence* of the method. This order is independent of the round-off error. In the numerical computations, to define the approximation at the point $t = t^\star$, we have to

execute approximately n steps, where $nh = t^\star$. Therefore, in case of local truncation error of the order $\mathcal{O}(h^{p+1})$, the expected magnitude of the global error is $\mathcal{O}(h^p)$. In Table 3.2.2 the results for the methods LT1 and LT2 confirm this conjecture: method LT1 is convergent in the first order, while method LT2 in the second order at the point $t^\star = 1$.

3.2.2 Some simple one-step methods

In the previous section we saw that the local Taylor method, especially for $p = 1$ is beneficial: for the computation by the formula (3.2.8) the knowledge of the partial derivatives of the function f is not necessary, and by decreasing the step-size of the mesh the unknown exact solution is well approximated at the mesh-points. Our aim is to define further one-step methods having similar good properties.

The LT1 method was obtained by the approximation of the solution on the subinterval $[t_i, t_{i+1}]$ by its first order Taylor polynomial.[4] Then the error (the local truncation error) is

$$|u(t_{i+1}) - T_{1,u}(t_{i+1})| = \mathcal{O}(h_i^2), \quad i = 0, 1, \ldots, N-1, \tag{3.2.14}$$

which means that the approximation is exact in the second order. Let us define instead of $T_{1,u}(t)$ some other, first order polynomial $P_1(t)$, for which the estimate (3.2.14) remains true, i.e., the estimation

$$|u(t_{i+1}) - P_1(t_{i+1})| = \mathcal{O}(h_i^2) \tag{3.2.15}$$

holds.
The polynomial $T_{1,u}(t)$ is the tangent line at the point $(t_i, u(t_i))$ to the exact solution. Therefore, we seek such a first order polynomial $P_1(t)$, which passes through this point, but whose direction is defined by the tangent lines to the solution $u(t)$ at the points t_i and t_{i+1}. To this aim, let $P_1(t)$ have the form $P_1(t) := u(t_i) + \alpha(t - t_i)$ $(t \in [t_i, t_{i+1}])$, where $\alpha = \alpha(u'(t_i), u'(t_{i+1}))$. (E.g., by the choice $\alpha = u'(t_i)$ we get $P_1(t) = T_{1,u}(t)$, and then the estimation (3.2.15) holds.)
Is any other suitable choice of α possible? Since

$$u(t_{i+1}) = u(t_i) + u'(t_i)h_i + \mathcal{O}(h_i^2), \tag{3.2.16}$$

therefore

$$u(t_{i+1}) - P_1(t_{i+1}) = h_i(u'(t_i) - \alpha) + \mathcal{O}(h_i^2),$$

i.e., the relation (3.2.15) is satisfied if and only if the estimation

$$\alpha - u'(t_i) = \mathcal{O}(h_i) \tag{3.2.17}$$

holds.

Theorem 3.2.2. *For any $\theta \in \mathbb{R}$, under the choice of α by*

$$\alpha = (1 - \theta)u'(t_i) + \theta u'(t_{i+1}) \tag{3.2.18}$$

the estimation (3.2.17) *is true.*

[4] In each subinterval $[t_i, t_{i+1}]$ we define a different Taylor polynomial of the first order, but the dependence of the polynomial on the index i will not be denoted.

Proof. Let us apply (3.2.16) to the function $u'(t)$. Then we have

$$u'(t_{i+1}) = u'(t_i) + u''(t_i)h_i + \mathcal{O}(h_i^2). \tag{3.2.19}$$

Substituting the relation (3.2.19) into the formula (3.2.18), we get

$$\alpha - u'(t_i) = \theta u''(t_i)h_i + \mathcal{O}(h_i^2), \tag{3.2.20}$$

which proves the statement. ■

Corollary 3.2.3. *The above polynomial $P_1(t)$ defines the one-step numerical method of the form*

$$y_{i+1} = y_i + \alpha h_i, \tag{3.2.21}$$

where, based on the relations (3.2.18) *and* (3.2.1)*, we have*

$$\alpha = (1-\theta)f(t_i, y_i) + \theta f(t_{i+1}, y_{i+1}). \tag{3.2.22}$$

Definition 3.2.4. *The numerical method defined by* (3.2.21)-(3.2.22) *is called θ-method.*

3.2.3 Explicit Euler method

Let us consider the θ-method with the choice $\theta = 0$. Then the formulas (3.2.21) and (3.2.22) result in the following method:

$$y_{i+1} = y_i + h_i f(t_i, y_i), \quad i = 0, 1, \ldots, N-1. \tag{3.2.23}$$

Since y_i is the approximation of the unknown solution $u(t)$ at the point $t = t_i$, therefore

$$y_0 = u(0) = u_0, \tag{3.2.24}$$

i.e., in the iteration (3.2.23) the starting value y_0, corresponding to $i = 0$, is given.

Definition 3.2.5. *The one-step method* (3.2.23)–(3.2.24) *is called explicit Euler method.*

We can characterize explicit Euler method The method (3.2.23)–(3.2.24) on the following example, which gives good inside of the method.

Example 3.2.6. *The simplest initial value problem is*

$$u' = u, \quad u(0) = 1, \tag{3.2.25}$$

whose solution is, of course, the exponential function $u(t) = e^t$.

Since for this problem $f(t, u) = u$, the explicit Euler method with a fixed step size $h > 0$ takes the form

$$y_{i+1} = y_i + hy_i = (1+h)y_i.$$

This is a linear iterative equation, and hence easy to get

$$y_i = (1+h)^i u_0 = (1+h)^i.$$

Then this is the proposed approximation to the solution $u(t_i) = e^{t_i}$ at the mesh point $t_i = ih$. Therefore, the Euler scheme to solve the differential equation, we are effectively approximating the exponential by a power function

$$e^{t_i} = e^{ih} \approx (1+h)^i$$

When we use simply $t^\star$ to indicate the fixed mesh-point $t_i = ih$, we recover, in the limit, a well-known calculus formula:

$$e^{t^\star} = \lim_{h\to 0}(1+h)^{t^\star/h} = \lim_{i\to\infty}(1+t^\star/h)^i$$

A reader familiar with the computation of compound interest will recognize this particular approximation. As the time interval of compounding, h, gets smaller and smaller, the amount in the savings account approaches an exponential.

3.2.4 Implicit Euler method

Let us consider the θ-method by the choice $\theta = 1$. For this case the formulas (3.2.21) and (3.2.22) together generate the following numerical method:

$$y_{i+1} = y_i + h_i f(t_{i+1}, y_{i+1}), \quad i = 0, 1, \ldots, N-1, \tag{3.2.26}$$

where again we put $y_0 = u_0$.

Definition 3.2.7. *The one-step numerical method defined in* (3.2.26)–(3.2.24) *is called* implicit Euler method.

Remark 3.2.8. *The Euler method of the form* (3.2.26) *is called implicit because* y_{i+1}*, the value of the approximation on the new time level* t_{i+1}*, can be obtained by solving a usually non-linear equation.*

- The convergence on the interval $[0, t^\star]$ yields the relation

  $$\lim_{h\to 0} \max_{i=1,2,\ldots,n} |e_i| = 0.$$

  As one can easily see, the relation $|e_n| \le C \cdot h$ holds for both methods (explicit Euler methodand implicit Euler method). Therefore the local truncation error $|e_n|$ can be bounded at any point uniformly on the interval $[0, t^\star]$, which means the convergence in the first order on the interval.

- Since the implicit Euler method is implicit, in each step we must solve a – usually non-linear – equation, namely, find the root of the equation $g(s) := s - hf(t_n, s) - y_n = 0$. This can be done by using some iterative method, such as Newton's method.

t_i	exact solution	EE	IE	TR
0.1	1.0048	1.0000	1.0091	1.0048
0.2	1.0187	1.0100	1.0264	1.0186
0.3	1.0408	1.0290	1.0513	1.0406
0.4	1.0703	1.0561	1.0830	1.0701
0.5	1.1065	1.0905	1.1209	1.1063
0.6	1.1488	1.1314	1.1645	1.1485
0.7	1.1966	1.1783	1.2132	1.1963
0.8	1.2493	1.2305	1.2665	1.2490
0.9	1.3066	1.2874	1.3241	1.3063
1.0	1.3679	1.3487	1.3855	1.3676

Table 3.2.3: Comparison in the maximum norm for the explicit Euler method (EE), the implicit Euler method (IE), and the trapezoidal method (TR) on the mesh with mesh-size $h = 0.1$.

3.2.5 Trapezoidal method

Let us consider the θ-method by the choice $\theta = 0.5$. For this case the formulas (3.2.21) and (3.2.22) generate the numerical method of the form

$$y_{i+1} - y_i = \frac{h_i}{2} \left[f(t_i, y_i) + f(t_{i+1}, y_{i+1}) \right], \quad i = 0, 1, \ldots, N-1, \tag{3.2.27}$$

where $y_0 = u_0$.

Definition 3.2.9. *The one-step method* (3.2.27) *is called* trapezoidal method.

3.2.6 Numerical test for the theta-method

We consider the test equation given in Exercise 3.2.10. In Table 3.2.3 we give the numerical solution by the above listed three numerical methods (explicit Euler method, implicit Euler method, trapezoidal method). In Table (3.2.4) we compare the numerical results obtained on refining meshes in the maximum norm. These results show that on some fixed mesh the explicit Euler method and the implicit Euler method give approximately the same accuracy, while the trapezoidal method is more accurate. On the refining meshes we can observe that the error function of the trapezoidal method has the order $\mathcal{O}(h^2)$, while for the explicit Euler method and thebimplicit Euler method the error function is of $\mathcal{O}(h)$. (These results completely correspond to the theory.)

3.3 Runge-Kutta method

We have seen, the maximum accuracy of the investigated one-step methods (explicit Euler method,implicit Euler method,trapezoidal method) for the Cauchy problem (3.2.1)-(3.2.2) is two. However, from practical point of view, this accuracy isn't enough: typically we require the construction of numerical methods with higher order accuracy. The accuracy of Taylor method is higher, but, in this case the realization of this method requires rather complicated preliminary analysis.

step-size (h)	EE	IE	TR
0.1	$1.92e-02$	$1.92e-02$	$3.06e-04$
0.01	$1.84e-03$	$1.84e-03$	$3.06e-06$
0.001	$1.84e-04$	$1.84e-04$	$3.06e-08$
0.0001	$1.84e-05$	$1.84e-05$	$3.06e-10$
0.0001	$1.84e-06$	$1.84e-06$	$5.54e-12$

Table 3.2.4: The error in the maximum norm for the explicit Euler method (EE), the implicit Euler method (IE), and the trapezoidal method (TR) on the mesh with mesh-size h.

3.3.1 Second order Runge-Kutta methods

Let us consider again the Cauchy problem (3.2.1)-(3.2.2). In order to introduce the Runge-Kutta methods, first of all we define a one-step method of second order accuracy, which is different from the trapezoidal method.
Let us define the first members of the Taylor series of the function $u(t)$ at the point $t = t^\star + h$. Then

$$u(t^\star + h) = u(t^\star) + hu'(t^\star) + \frac{h^2}{2!}u''(t^\star) + \mathcal{O}(h^3). \tag{3.3.1}$$

Using the derivatives (3.2.4), and introducing the notations

$$f = f(t^\star, u(t^\star)), \quad \partial_i f = \partial_i f(t^\star, u(t^\star)), \quad \partial_{ij} f = \partial_{ij} f(t^\star, u(t^\star)), \quad \text{etc.},$$

the equation (3.3.1) can be rewritten as

$$\begin{aligned} u(t^\star + h) =& u(t^\star) + hf + \frac{h^2}{2!}(\partial_1 f + f\partial_2 f) + \mathcal{O}(h^3) \\ =& u(t^\star) + \frac{h}{2}f + \frac{h}{2}[f + h\partial_1 f + hf\partial_2 f] + \mathcal{O}(h^3). \end{aligned} \tag{3.3.2}$$

Since [5]

$$f(t^\star + h, u(t^\star) + hf(t^\star, u(t^\star)) = f + h\partial_1 f + hf\partial_2 f + \mathcal{O}(h^2), \tag{3.3.3}$$

therefore (3.3.2) can be written in the form

$$u(t^\star + h) = u(t^\star) + \frac{h}{2}f + \frac{h}{2}(f(t^\star + h, u(t^\star) + hf(t^\star, u(t^\star))) + \mathcal{O}(h^3). \tag{3.3.4}$$

Therefore applying the formula (3.3.4) at some arbitrary mesh-point $t_i = t^\star$ of ω_h, we can define the following one step, explicit numerical method:

$$y_{i+1} = y_i + \frac{h}{2}f(t_i, y_i) + \frac{h}{2}f(t_{i+1}, y_i + hf(t_i, y_i)). \tag{3.3.5}$$

Let us introduce the notations

$$k_1 = f(t_i, y_i); \quad k_2 = f(t_{i+1}, y_i + hf(t_i, y_i)) = f(t_i + h, y_i + hk_1). \tag{3.3.6}$$

[5] We recall that the first order Taylor polynomial of the function $f : Q_T \to \mathbb{R}$ around the point (t, u) , for arbitrary constants $c_1, c_2 \in \mathbb{R}$ can be written as $f(t + c_1 h, u + c_2 h) = f(t, u) + c_1 h\partial_1 f(t, u) + c_2 h\partial_2 f(t, u) + \mathcal{O}(h^2)$.

Then the method (3.3.5) can be written in the form

$$y_{i+1} = y_i + \frac{h}{2}(k_1 + k_2). \tag{3.3.7}$$

Definition 3.3.1. *The one-step, explicit numerical method* (3.3.6)-(3.3.7) *is called* Heun method

Remark 3.3.2. *Based on* (3.3.4)*, we have*

$$u(t^\star + h) - u(t^\star) - \frac{h}{2}f - \frac{h}{2}(f(t^\star + h, u(t^\star) + hf(t^\star, u(t^\star))) = \mathcal{O}(h^3). \tag{3.3.8}$$

This means that the exact solution of the Cauchy problem (3.2.1)-(3.2.2) *satisfies the formula of the Heun method* (3.3.5) *with the accuracy* $\mathcal{O}(h^3)$*, which means that the Heun method is of second order.*

3.3.2 Higher order Runge-Kutta methods

The following generalization seems to be natural:

$$\begin{aligned} k_1 &= f(t_i, y_i), \\ k_2 &= f(t_i + a_2 h, y_i + h b_{21} k_1), \\ k_3 &= f(t_i + a_3 h, y_i + h b_{31} k_1 + h b_{32} k_2), \end{aligned} \tag{3.3.9}$$

and the approximation on the new mesh-point id defined as

$$y_{i+1} = y_i + h(\sigma_1 k_1 + \sigma_2 k_2 + \sigma_3 k_3). \tag{3.3.10}$$

The parameters of this method, according to the table, can be written as follows

$$\begin{array}{c|ccc} 0 & & & \\ a_2 & b_{21} & & \\ a_3 & b_{31} & b_{32} & \\ \hline & \sigma_1 & \sigma_2 & \sigma_3 \end{array} \tag{3.3.11}$$

Our aim is to define the parameters in (3.3.9) in the way that the corresponding numerical method was of third order accurate. To get this condition, we have to define again the local approximation error, as it was done before. After some long (but not difficult) calculation we obtain the following result.

Theorem 3.3.3. *The numerical method* (3.3.9) *has third order accuracy, if and only if the conditions*

$$\begin{aligned} a_2 = b_{21}, \quad & a_3 = b_{31} + b_{32}, \\ a_3(a_3 - a_2) - b_{32}a_2(2 - 3a_2) = 0, \quad & \sigma_3 b_{32} a_2 = 1/6, \\ \sigma_2 a_2 + \sigma_3 a_3 = 1/2, \quad & \sigma_1 + \sigma_2 + \sigma_3 = 1 \end{aligned} \tag{3.3.12}$$

are satisfied.

Clearly, (3.3.12) yields six equations (conditions) for eight unknown values. From the possible solution we give two cases.

- The method with the parameters

$$\begin{array}{c|ccc} 0 & & & \\ 1/3 & 1/3 & & \\ 2/3 & 0 & 2/3 & \\ \hline & 1/4 & 0 & 3/4 \end{array} \tag{3.3.13}$$

  is very popular in the different applications.

- The following third order method

$$\begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 1 & -1 & 2 & \\ \hline & 1/6 & 2/3 & 1/6 \end{array} \tag{3.3.14}$$

  is also often used for the applied problems. We note that this method has accuracy $\mathcal{O}(h^5)$ for the problems with $f(t,u) = f(t)$ with the Simpson formula. Therefore this method recommended also fort he problems, when the partial derivative $\partial_2 f$ is close to the zero.

When we want the methods higher than three order method ($p > 3$), then we need some more generalization. This will be the following form.

Let $m \geq 1$ some given integer. We define the following, so called *m-stage explicit Runge-Kutta method*:

$$\begin{aligned} k_1 &= f(t_i, y_i), \\ k_2 &= f(t_i + a_2 h, y_i + h b_{21} k_1), \\ k_3 &= f(t_i + a_3 h, y_i + h b_{31} k_1 + h b_{32} k_2), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ k_m &= f(t_i + a_m h, y_i + h b_{m1} k_1 + h b_{m2} k_2 + \ldots + h b_{m,m-1} k_{m-1}) \end{aligned} \tag{3.3.15}$$

$$y_{i+1} = y_i + h(\sigma_1 k_1 + \sigma_2 k_2 + \ldots + \sigma_m k_m). \tag{3.3.16}$$

by giving the values of the parameters in this formula, we define the concrete numerical method. As before, we can write the parameters in a table:.

$$\begin{array}{c|cccc} 0 & & & & \\ a_2 & b_{21} & & & \\ a_3 & b_{31} & b_{32} & & \\ \ldots & \ldots & \ldots & \ldots & \\ a_m & b_{m1} & b_{m2} & \ldots & b_{m,m-1} \\ \hline & \sigma_1 & \sigma_2 & \ldots & \sigma_m \end{array} \tag{3.3.17}$$

For writing the method in compact form, we introduce some notations. Let the vectors $\sigma, \mathbf{a} \in \mathbb{R}^m$ denote the row-vectors with the coordinates σ_i and a_i, respectively, (where we always assume that $a_1 = 0$). Let $\mathbf{B} \in R^{m \times m}$ denote the matrix with the elements b_{ij}, which is a strictly lower triangular matrix, i.e.,

$$\mathbf{B}_{ij} = \begin{cases} b_{ij}, & \text{for } i > j, \\ 0, & \text{for } i \leq j. \end{cases}$$

Definition 3.3.4. *For some explicit Runge-Kutta method the corresponding table of the parameters in the form*

$$\begin{array}{c|c} \boldsymbol{a}^\top & \boldsymbol{B} \\ \hline & \boldsymbol{\sigma} \end{array} \tag{3.3.18}$$

is called Butcher tableau.

In the sequel, when we specify some explicit Runge-Kutta method, then we list the lower triangular part of the matrix $\mathbf{B}$, only.

For some fixed explicit Runge-Kutta method the order of the consistency can be defined by some simple, however usually cumbersome computation:.This can be realized by the following steps.

- First on the right side of the formulas (3.3.15)-(3.3.15) we replace y_i by the values $u(t_i)$.
- Then we replace on the left side of this formula y_{i+1} by $u(t_i + h)$.
- We compute the difference between the two sides, and we define its order by h.

Executing these steps for the explicit Runge-Kutta method, we get the condition of the p-th order of some explicit Runge-Kutta method.

3.4 Multistep numerical methods

In the previous sections we have considered the one-step methods, i.e., the numerical methods such that the value of the approximation to the exact solution is defined by the approximation, already defined at the previous mesh-point. In the sequel we generalize this approach in that way, that the new value of the approximation is defined by not only one but several previous approximations. Such methods are called multistep method.

Further m ($m \geq 1$) denote the number of the mesh-points, at which the approximations are taken into account for the definition at the new mesh-point. Such multistep method is called m-step method. (The one-step methods can be considered as special case of multistep methods with the choice $m = 1$.)

Next we show on two simple examples that by developing into Taylor series the solution of the Cauchy problem (3.2.1)-(3.2.2) around suitably chosen points how can these methods derived.

Example 3.4.1. *Clearly*

$$\begin{aligned} u(t_{i-1}) =& u(t_i) - hu'(t_i) + \frac{h^2}{2}u''(t_i) + \mathcal{O}(h^3), \\ u(t_{i-2}) =& u(t_i) - 2hu'(t_i) + \frac{4h^2}{2}u''(t_i) + \mathcal{O}(h^3). \end{aligned} \tag{3.4.1}$$

Therefore

$$3u(t_i) - 4u(t_{i-1}) + u(t_{i-2}) = 2hu'(t_i) + \mathcal{O}(h^3) = 2hf(t_i, u(t_i)) + \mathcal{O}(h^3).$$

Hence, by using the notation $f_i = f(t_i, y_i)$ we can define the numerical method as follows

$$y_i - \frac{4}{3}y_{i-1} + \frac{1}{3}y_{i-2} = \frac{2}{3}hf_i, \quad i = 2, 3, ... \tag{3.4.2}$$

As we can see, the method (3.4.2) *is two-steps, implicit method, and it has second order consistency.*

Example 3.4.2. *We develop into the Taylor series at the point t_{i-1} the exact solution and its derivative, too. Hence we have the relations*

$$\begin{aligned} u(t_i) =& u(t_{i-1}) + hu'(t_{i-1}) + \frac{h^2}{2}u''(t_{i-1}) + \mathcal{O}(h^3), \\ u'(t_{i-2}) =& u'(t_{i-1}) - hu''(t_{i-1}) + \mathcal{O}(h^2). \end{aligned} \tag{3.4.3}$$

From the second relation we have $hu''(t_{i-1}) = u'(t_{i-1}) - u'(t_{i-2}) + \mathcal{O}(h^2)$. Substituting this expression into the first formula, we obtain

$$u(t_i) = u(t_{i-1}) + \frac{h}{2}[3u'(t_{i-1}) - u'(t_{i-2})] + \mathcal{O}(h^3).$$

Based on this relation, we define the numerical method

$$y_i - y_{i-1} = h[\frac{3}{2}f_{i-1} - \frac{1}{2}f_{i-2}], \quad i = 2, 3, ... \tag{3.4.4}$$

which is two-steps, explicit method, having second order consistency.

3.5 Consistency of the general linear multistep methods

Following the above examples, we can define the linear multistep methodin a general form, too.

Definition 3.5.1. *Let $a_0, a_1, \ldots a_m$ and $b_0, b_1, \ldots b_m$ given numbers. The iteration of the form*

$$a_0y_i + a_1y_{i-1} + \cdots + a_my_{i-m} = h[b_0f_i + b_1f_{i-1} + \ldots + b_mf_{i-m}], \quad i = m, m+1, \ldots, \tag{3.5.1}$$

is called linear, m-step methods.

In the sequel we always assume that $a_0 \neq 0$, otherwise we are not able to define from the known values $y_{i-m}, y_{i-m+1}, \ldots, y_{i-1}$ the unknown approximation y_i. According the notation $f_i = f(t_i, y_i)$, the method (3.5.1)) is *explicit*, when $b_0 = 0$, and it is *implicit*, when $b_0 \neq 0$. We fix some linear multistep method by

giving the parameters a_k and b_k $(k = 0, 1, \ldots, m)$. For instance, the numerical method (3.4.2) is fixed by giving the values $m = 2$, $a_0 = 1$, $a_1 = -4/3$, $a_2 = 1/3$, and $b_0 = 2/3$, $b_1 = 0$, $b_2 = 0.$, respectively. When there are given two m-steps linear multistep methods with the parameters (a_k, b_k) and $(a_k^\star, b_k^\star)$, and there exists a parameter $\beta \neq 0$ such that $a_k = \beta a_k^\star$ and $b_k = \beta b_k^\star$ for any $k = 0, 1, \ldots, m$, then from the given initial values both methods generate the same result. These methods will be considered as the same method. Hence, the $2m + 2$ parameters $((a_k, b_k),\ \ k = 0, 1, \ldots m)$ which define some linear multistep method, one should be a'priori fixed. This is the reason why we always assume that

$$a_0 = 1. \tag{3.5.2}$$

Further we define the consistency and the order of consistency of the linear multistep methodin the form (3.5.1). The local approximation error has the form

$$g_i(h) = \sum_{k=0}^{m} \Big[a_k u(t_{i-k}) - h b_k f(t_{i-k}, u(t_{i-k}))\Big]. \tag{3.5.3}$$

Based on the equation (3.2.1) we have $u'(t_{i-k}) = f(t_{i-k}, u(t_{i-k}))$,. Therefore

$$g_i(h) = \sum_{k=0}^{m} \Big[a_k u(t_{i-k}) - h b_k u'(t_{i-k})\Big]. \tag{3.5.4}$$

We develop into the Taylor series the functions on the right side (3.5.4) at the point $t = t_i$ up to order p and $p - 1$, respectively. Then

$$\begin{aligned} u(t_{i-k}) =& u(t_i - kh) = u(t_i) - khu'(t_i) + \frac{1}{2!}k^2h^2u''(t_i) \\ &+ \ldots + (-1)^p \frac{1}{p!}k^p h^p u^{(p)}(t_i) + \mathcal{O}(h^{p+1}), \end{aligned} \tag{3.5.5}$$

$$\begin{aligned} u'(t_{i-k}) =& u'(t_i - kh) = u'(t_i) - khu''(t_i) \\ &+ \ldots + (-1)^{p-1} \frac{1}{(p-1)!}k^{p-1}h^{p-1}u^{(p)}(t_i) + \mathcal{O}(h^p). \end{aligned} \tag{3.5.6}$$

Substituting these expressions into (3.5.4), for the local approximation error we obtain

$$g_i(h) = d_0 u(t_i) + h d_1 u'(t_i) + h^2 d_2 u''(t_i) + \ldots + h^p d_p u^p(t_i) + \mathcal{O}(h^{p+1}), \tag{3.5.7}$$

where

$$\begin{aligned} d_0 &= \sum_{k=0}^{m} a_k, \\ d_1 &= -\sum_{k=0}^{m} (k a_k + b_k) \\ d_2 &= \sum_{k=0}^{m} (\frac{1}{2}k^2 a_k + k b_k), \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ d_p &= (-1)^p \sum_{k=0}^{m} (\frac{1}{p!}k^p a_k + \frac{1}{(p-1)!}k^{p-1} b_k). \end{aligned} \tag{3.5.8}$$

Some linear multistep method is of consistency order p, when $g_i(h) = \mathcal{O}(h^{p+1})$, which results in the condition $d_0 = d_1 = \ldots = d_p = 0$. Using (3.5.8), we obtained the following statement.

Theorem 3.5.2. *The linear multistep method of the form* (3.5.1) *has p-th order consistency, when for the parameters of the method the following conditions are satisfied:*

$$\begin{aligned} a_0 = 1, \quad \sum_{k=0}^{m} a_k = 0 \\ \frac{1}{j}\sum_{k=0}^{m} k^j a_k + \sum_{k=0}^{m} k^{j-1} b_k = 0, \quad j = 1, 2, \ldots, p. \end{aligned} \tag{3.5.9}$$

Using the above statement, the condition of the consistency of some linear multistep method can be also formulated.

Corollary 3.5.3. *The linear multistep method of the form* (3.5.1) *is consistent if and only if when for the parameters of the method the conditions*

$$\begin{aligned} a_0 = 1, \quad \sum_{k=0}^{m} a_k = 0 \\ \sum_{k=0}^{m} k a_k + \sum_{k=0}^{m} b_k = 0 \end{aligned} \tag{3.5.10}$$

are satisfied.

3.6 Numerical solution of initial value problems with Matlab

The investigated numerical methods can be realized on computer with help of different program packages. In the following we consider the Matlab, which has several built-in routines for the the different methods, however, to prepare own code is also possible, and, since this is not difficult, it is highly recommended for the Readers.

Let consider the explicit Euler method and we prepare the program (so called *m-file* for this method. Then we can use very simple this program as a function, by giving its parameters.

First we describe those steps which are required for the realization.

- In the first step we start the work of Matlab by its running. Then, into the Editor we type the following:

```
function[t,y] = expeuler(diffegy, t0, y0, h, N)
t=zeros(N+1,1);
y = zeros(N+1,1);
t(1) = t0;
y(1) = y0;
for i=1:N
t(i+1) = t(i) + h;
y(i+1) = y(i) + h * diffegy(t(i),y(i));
end
```

- Let us describe the program in more details.

  In the first line of this program we gave how can be called the program. (We gave the name expeuler.) This means, that on the left side of equality symbol = we identify the task by listing the input data, and the required input parameters of the explicit Euler method. On the left side of equality symbol = in the bracket $[\cdot]$ we list the output parameters. (Typically they are such values which are computed within the program and later are used for some purposes.) In our case we have five input and two output parameters. The output parameters (results) are two vectors: the first vector (t) is the vector containing the discrete time points (the mesh-points), and the second vector (y) contains the numerical solution at these points. The first input parameter $(diffegy)$ identifies the differential equation by giving the function, standing on the right side of the differential equation (which was denoted by f). The second parameter $(t0)$ denotes the point, where the initial condition is given. The third parameter $(y0)$ is the initial value at this point.The next parameter (h) which denotes the step-size of the mesh, where the numerical solution is defined. Finally, N denotes the number of the steps on this mesh. (I.e., the numerical solution is defined at the equidistant mesh-point of the interval $[t0, Nh]$.)

  In the second and third lines we give zero value for the vectors t and y, where the numerical approximations will be computed. In the next two lines we define the starting values for these vectors.

  In fact, these steps were the preparation work for the method.

  From the next line we start to give the algorithm of the method. Within a cycle first we give the values of t_i, then we compute the slope of the approximation, and then we compute y_i according to the explicit Euler method.

- With this program we cannot compute directly the numerical solution of the problem, because we have to identify the function "diffegy", too. (We remind that this function describes the right side of the equation f.)

  We will consider the example, given as

  $$u'(t) = -u(t) + t + 1,$$

  where $t \in [0, 1]$ and $u(0) = 1$.[6] To prepare the function "diffegy", we open a new m-file, and we write the following:

```
function dydt = diffegy(t,y)
dydt = -y + t + 1;
```

- When both routines are ready, we can run the program. We will use the parameters $h = 0.1$, $h = 0.01$ and $h = 0.001$ on the interval $[0, 1]$. In the windows Command we type the following:

  `[T1,Ye] = expeuler(@diffegy, 0, 1, 0.1, 10).`

  After pushing Enter we obtain the vectors T1 and Ye, which contains the place and the values of the approximation. If we want to draw the solution, then by giving the command

[6] We remind that this example was already considered previously. In this section section we will solve this problem by Runge-Kutta methods, and linear multistep method, too.

`plot(T1,Ye)`

we can do it, and then in the separate windows we will see the graphic of the numerical solution. (For $h = 0.01$ and $h = 0.001$ we should type

`[T1,Ye] = expeuler(@diffegy, 0, 1, 0.01, 100)`

and

`[T1,Ye] = expeuler(@diffegy, 0, 1, 0.001, 1000)`,

respectively.)

In order to get the global error of the explicit Euler method we have to compare the numerical solution with the exact solution. The solution of our problem is the function

$$u(t) = e^{-t} + t.$$

On the Figures 3.6.1-3.6.3 we can see the accuracy of the method for the step-sizes $h = 0.1$, $h = 0.01$ and $h = 0.001$, respectively. In accordance with the theory, by decreasing h the graph of the numerical solution approaches to the graph of the exact solution.

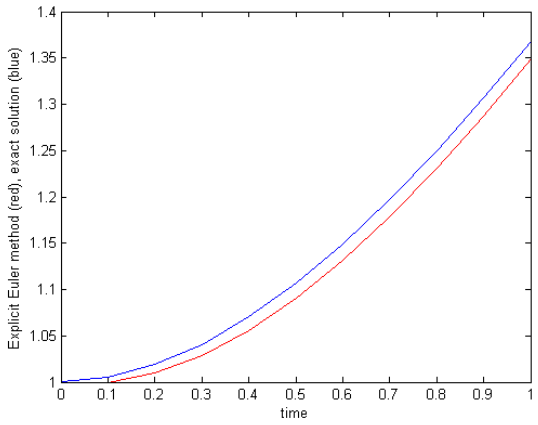


Figure 3.6.1: The explicit Euler method with step-size $h = 0.1$.

3.7 In-built Matlab programs

The Matlab has possibility to use in-built programs, which, in fact realize the most useful methods. These methods allow us to get the numerical solution of the initial value problems efficiently with high accuracy.

One of the methods is called ODE45, which is based on the embedded Dormand-Prince method. We recall that these methods are bases on two Runge-Kutta methods with different order, but in their Butcher tableau $\mathbf{a}^\top$ and $\mathbf{B}$ are the same. (However, the weighting vectors $\boldsymbol{\sigma}$ are different, and therefore their orders are also different. The suitable combination of these methods give us possibility for the right choice of the varying step-size of the combined method.) This method based on the combination of some fourth and fifth order Runge-Kutta methods, and the

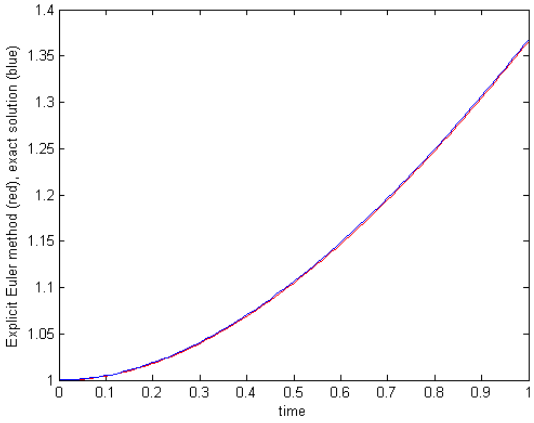


Figure 3.6.2: The explicit Euler method with step-size $h = 0.01$.

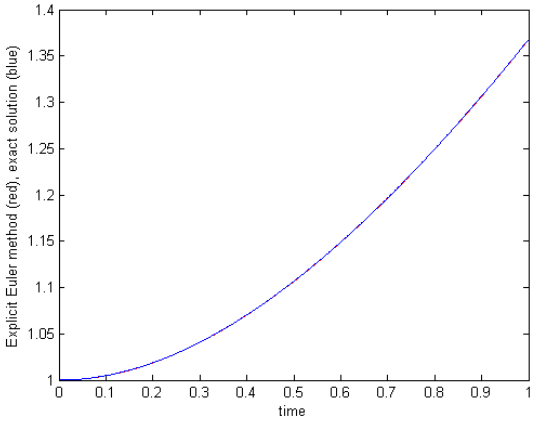


Figure 3.6.3: The explicit Euler method with step-size $h = 0.001$.

step-size is chosen that the error of the combined method is the error of the fourth order method. In the Table 3.7.1 we give the Butcher tableau of the method. (The first row for σ belongs tot eh fourth order method, while the second row does to the fifth order method.) The routine ODE45 can be called in the same way as the routines written by us in the previous section. Namely, the form is `[T1, Y45] = ode45(@diffegy, T1, 1)`. We have two output vectors again, which give the time points and the values of the approximations. The input parameters are the following: "diffegy" is the function, which describes the right side of the equation; "T1" is the vector, which give those points where we aim the get the numerical solution, and finally we give the initial value for the problem.[7]

The accuracy of the method ODE45 we have checked on the previous Exercise

[7]We note that it is also possible a fourth (optional) parameter for numerical integrating. However, typically its default value is enough for the practical purposes. If one wants to change the default value, it is possible by using the program ODE45, and the details can be found in the help list of Matlab.

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$\frac{-56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$\frac{-25360}{2187}$	$\frac{64448}{6561}$	$\frac{-212}{729}$			
1	$\frac{9017}{3168}$	$\frac{-355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$\frac{-5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$\frac{-2187}{6784}$	$\frac{11}{84}$	
	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$\frac{-92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$
	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$\frac{-2187}{6784}$	$\frac{11}{84}$	0

Table 3.7.1: The parameters of the embedded Dormand-Prince RK in ODE45 routine

3.2.10, with the step-sizes $h = 0.1$ and $h = 0.01$, respectively. The corresponding results are included into the Tables 3.7.2 and 3.7.3. We can see that the accuracy of the method doesn't decrease with decreasing the step-size. The reason is that, due to the adaptivity of the choice of the mesh-size in the method, the obtainable accuracy is reached already for the first choice $h = 0.1$.

Another in-built and widely used routine is the embedded Runge-Kutta method ODE23, which is also called as Bogacki-Shampine method . This program can

t_i	exact solution	numerical solution	error
0	1.0000	1.0000	0
0.1000	1.0048	1.0048	$2.9737e-010$
0.2000	1.0187	1.0187	$5.3815e-010$
0.3000	1.0408	1.0408	$7.3041e-010$
0.4000	1.0703	1.0703	$8.8120e-010$
0.5000	1.1065	1.1065	$9.9668e-010$
0.6000	1.1488	1.1488	$1.0822e-009$
0.7000	1.1966	1.1966	$1.1424e-009$
0.8000	1.2493	1.2493	$1.1814e-009$
0.9000	1.3066	1.3066	$1.2026e-009$
1.0000	1.3679	1.3679	$1.2090e-009$

Table 3.7.2: Results for the ODE45 method with step-size $h = 0.1$

t_i	exact solutuion	numerical solution	error
0	1.0000	1.0000	0
0.0100	1.0000	1.0000	$2.2204e-016$
0.0200	1.0002	1.0002	$8.8940e-011$
⋮	⋮	⋮	⋮
0.1000	1.0048	1.0048	$5.4080e-009$
⋮	⋮	⋮	⋮
0.5000	1.1065	1.1065	$2.8278e-009$
⋮	⋮	⋮	⋮
0.8000	1.2493	1.2493	$1.6519e-009$
⋮	⋮	⋮	⋮
0.9000	1.3066	1.3066	$1.3610e-009$
⋮	⋮	⋮	⋮
0.9900	1.3616	1.3616	$1.0920e-009$

Table 3.7.3: Results for ODE45 method with step-size $h = 0.01$.

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{3}{4}$	0	$\frac{3}{4}$		
1	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	
	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	0
	$\frac{7}{24}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{8}$

Table 3.7.4: The parameters of the embedded Bogacki-Shampine RK in the routine ODE23

be called as `[T1, Y23] = ode23(@diffegy, T1, 1)`, , similarly to the ODE45 method. The Table 3.7.4 contains the Butcher tableau of the method.

We emphasize that Bogacki-Shampine method is explicit, $(2, 3)$-type Runge-Kutta method. This method is especially useful when we want to get numerical solution quickly and with cheap but low accuracy. (Usually the methods ODE45 and ODE23 are used only for non-stiff problems.)

Testing the ODE23 routine on the Exercise 3.2.10 for the mesh-sizes $h = 0.1$ and $h = 0.01$, we obtain the Tables 3.7.5 and 3.7.6, respectively.

In Matlab we can also find in-built routines for the linear multistep methods. Such routine is the ODE113 method , which order can change from 1 until 13,

t_i	exact solution	numerical solution	error
0	1.0000	1.0000	0
0.1000	1.0048	1.0048	$4.0847e-006$
0.2000	1.0187	1.0187	$7.3920e-006$
0.3000	1.0408	1.0408	$1.0033e-005$
0.4000	1.0703	1.0703	$1.2104e-005$
0.5000	1.1065	1.1065	$1.3690e-005$
0.6000	1.1488	1.1488	$1.4865e-005$
0.7000	1.1966	1.1966	$1.5692e-005$
0.8000	1.2493	1.2493	$1.6227e-005$
0.9000	1.3066	1.3066	$1.6518e-005$
1.0000	1.3679	1.3679	$1.6607e-005$

Table 3.7.5: Results of the routine ODE23 for $h = 0.1$.

t_i	exact solution	numerical solution	error
0	1.0000	1.0000	0
0.0100	1.0000	1.0000	$4.1583e-010$
0.0200	1.0002	1.0002	$3.3825e-008$
⋮	⋮	⋮	⋮
0.1000	1.0048	1.0048	$1.8521e-006$
⋮	⋮	⋮	⋮
0.5000	1.1065	1.1065	$1.2194e-005$
⋮	⋮	⋮	⋮
0.9000	1.3066	1.3066	$1.5515e-005$
⋮	⋮	⋮	⋮
0.9900	1.3616	1.3616	$1.5233e-005$
1.0000	1.3679	1.3679	$1.5087e-005$

Table 3.7.6: Results of the routine ODE23 for $h = 0.01$.

t_i	exact solution	numerical solution	error
0	1.0000	1.0000	0
0.1000	1.0048	1.0048	$6.2300e-006$
0.2000	1.0187	1.0187	$1.8714e-005$
0.3000	1.0408	1.0408	$2.7885e-005$
0.4000	1.0703	1.0703	$2.1933e-005$
0.5000	1.1065	1.1065	$1.8889e-005$
0.6000	1.1488	1.1488	$1.7254e-005$
0.7000	1.1966	1.1966	$1.5668e-005$
0.8000	1.2493	1.2493	$1.4228e-005$
0.9000	1.3066	1.3066	$1.2872e-005$
1.0000	1.3679	1.3679	$1.1643e-005$

Table 3.7.7: Results of the routine ODE113 for $h = 0.1$.

and based on the Adams-Bashforth-Moulton method. Comparing with the ODE45 method, we can conclude that the ODE113 method is less accurate but cheaper. This method is especially recommended when the evaluation of the function f is expensive.

The syntax of the routine is the usual `[T1, Y113] = ode113(@diffegy, T1, 1)` with the same output and input parameters, and the method is applicable for the non-stiff problems. The accuracy of the method, for the step-sizes $h = 0.1$ and $h = 0.01$, on the previous test problem can be seen on the Tables 3.7.7 and 3.7.8, respectively.

Below we summarize the accuracy of the different methods, investigated in this part, for the same problem (3.2.10). We compare the errors at the time-point $t^\star = 1$ on the meshes with mesh-size $h = 0.1$ and $h = 0.01$, respectively.

method	e_{n^*}	
	$h_1 = 0.1$	$h_2 = 0.01$
explicit Euler	1.9201e-002	1.8471e-003
improved Euler	6.6154e-004	6.1775e-006
implicit Euler (more accurate)	1.7664e-002	1.8318e-003
implicit Euler	2.1537e-002	1.8687e-003
ODE45	1.2090e-009	1.0903e-009
ODE23	1.6607e-005	1.5087e-005
ODE113	1.1643e-005	1.6360e-008

Finally we note that in Matlab there are in-built routines for the stiff problems, too. Such method is ODE23S, which can be called by the usual command [T1, Y23S] = ODE23S(@DIFFEGY, T1, 1)). The routines ODE123T and ODE123TB are recommended for the numerical solution of stiff problems with mild stiff number. These methods have only moderate accuracy. The method ODE15S based on the backward differentiation formulas, which are also called Gear methods. The

t_i	exact solution	numerical solution	error
0	1.0000	1.0000	0
0.0100	1.0000	1.0001	$1.6625e-007$
0.0200	1.0002	1.0002	$1.4800e-007$
⋮	⋮	⋮	⋮
0.1000	1.0048	1.0048	$1.4004e-007$
⋮	⋮	⋮	⋮
0.5000	1.1065	1.1065	$1.7178e-008$
⋮	⋮	⋮	⋮
0.8000	1.2493	1.2493	$2.0090e-008$
⋮	⋮	⋮	⋮
0.9000	1.3066	1.3066	$1.8052e-008$
⋮	⋮	⋮	⋮
0.9800	1.3553	1.3553	$1.6692e-008$
0.9900	1.3616	1.3616	$1.6525e-008$
1.0000	1.3679	1.3679	$1.6360e-008$

Table 3.7.8: Results of the routine ODE113 for $h = 0.01$.

algorithm uses the varying step-size formula. This method is especially applicable when the method ODE45 is very slow or it is failed. The method ODE23S is a second order, modified one-step Rosenbrock- method. Because this is a one-step method, usually it is more economic than the ODE15S method.

Chapter 4

Ferenc Izsák: Computer aided simulation of time dependent phenomena

4.1 Introduction

In the natural sciences, many time dependent phenomena can be simulated with partial differential equations (PDE's). For the simulations, we have to solve the correspondig equations. This is almost impossible with analytic methods, therefore, we have to use efficient numerical methods. From the theoretical point of view, we have to prove convergence results, i.e., we have to verify that the corresponding numerical approximations converge to the analytic solutions and a convergence speed have to be given.

We deal with linear partial differential equations, or in details, with initial-boundary value problems for these equations [2]. In concrete terms, the following phenomena will be simulated:

- Diffusion - transport governed by concentration differences.
- Advection - transport governed by an external force.
- Wave propagation - transport of deformation/dilatation governed by an elastic medium.

4.2 The basic principles in the numerical solution

We apply finite difference approach such that both sides of the equations are approximated with finite differences and solve the corresponding system of equations. To clarify the details, we introduce the following notations, where for the simplicity, the spatial dimension is two.

- Unknowns in the numerical approximation are $u^n_{j,k} \approx u(t_n, x_j, y_k)$, so the solution is approximated at distinct points at distinct times, where
  - (x_j, y_k) denote the vertices in a given rectangular grid Ω_h of the original domain Ω.

- $t_1, \dots, t_n$ denote distinct times, $t_n = n \cdot \delta,\ t_N = T$.
- Unknowns at time $n\delta$: $\mathbf{u}^n = \{u^n_{j,k} : (x_j, y_k) \in \Omega_h\}$, all unknowns: $\mathbf{u}_{\text{disc}} = \{u^n_{j,k} : n = 1, 2, \dots, N\}$.
- In any case, the initial conditions are given as $\mathbf{u}^0$ and we use appropriate boundary values for the computations.

4.3 Motivation: a model problem and the chief questions

To demonstrate the main principles and the chief questions, we discuss a simple model problem for a one-dimensional diffusion problem

$$\begin{cases} \partial_t u(t,x) = \sigma_D \partial_{xx} u(t,x) \quad t \in \mathbb{R}^+,\ x \in (0,\pi) \\ u(t,0) = u(t,\pi) = 0 \quad t \in \mathbb{R}^+ \\ u(0,x) = \sin x \quad x \in (0,\pi). \end{cases} \tag{4.3.1}$$

We approximate then both sides of (4.3.1) and give the resulting system of linear equations.

4.3.1 First step: approximate the differential operators

For the time derivative

$$\partial_t u(t,x) \approx \frac{1}{\delta}(u(t+\delta,x) - u(t,x)), \tag{4.3.2}$$

such that we have the error term

$$\partial_t u(t,x) = \frac{1}{\delta}(u(t+\delta,x) - u(t,x)) + \mathcal{O}(\delta).$$

For the spatial derivative we use

$$\sigma_D \partial_{xx} u(t,x) \approx \frac{1}{h}(u(t,x+h) - 2u(t,x) + u(t,x-h)), \tag{4.3.3}$$

4.3.2 Second step: construct a linear system to solve

Using (4.3.2) and (4.3.3), for any (t,x) with $x - h \in [0,\pi]$ and $x + h \in [0,\pi]$ we have

$$\frac{1}{\delta}(u(t+\delta,x) - u(t,x)) \approx \sigma_D \frac{1}{h^2}(u(t,x+h) - 2u(t,x) + u(t,x-h)), \tag{4.3.4}$$

which componentwise results in the following equalities for the numerical approximations:

$$u_k^{n+1} = u_k^n + \sigma_D \frac{\delta}{h^2}(u_{k-1}^n - 2u_k^n + u_{k+1}^n) = u_k^n + r(u_{k-1}^n - 2u_k^n + u_{k+1}^n) \quad k = 1, 2, \dots, N$$

with $r = \frac{\sigma_D \delta}{h^2}$, where N is the number of internal gridpoints in $[0, \pi]$. According to the boundary conditions, we also have $u_0^{n+1} = u_{N+1}^{n+1} = 0$.
Summarized, we arrive at the *numerical scheme*

$$\begin{cases} u_k^0 = \sin(kh), \ k = 0, 1, \dots, N, N+1 \\ u_k^{n+1} = u_k^n + r(u_{k-1}^n - 2u_k^n + u_{k+1}^n), \ k = 1, 2, \dots, N \\ u_0^{n+1} = u_{N+1}^{n+1} = 0, \end{cases} \tag{4.3.5}$$

where the *time step* can be recasted into the matrix form

$$\begin{pmatrix} u_1^{n+1} \\ u_2^{n+1} \\ \vdots \\ u_{N-1}^{n+1} \\ u_N^{n+1} \end{pmatrix} = \begin{pmatrix} 1-2r & r & 0 & \dots & 0 & 0 & 0 \\ r & 1-2r & r & \dots & 0 & 0 & 0 \\ \vdots & \vdots & & \ddots & & & \vdots \\ 0 & 0 & 0 & \dots & r & 1-2r & r \\ 0 & 0 & 0 & \dots & 0 & r & 1-2r \end{pmatrix} \begin{pmatrix} u_1^n \\ u_2^n \\ \vdots \\ u_{N-1}^n \\ u_N^n \end{pmatrix}.$$

In this way, the approximations $\mathbf{u}^1, \mathbf{u}^2, \dots$ in the consecutive time steps can be computed.

4.3.3 Third step: some numerical experiments

In all cases, we simulated on the time interval $(0, 1)$ with $\delta = 0.01$. The relative error can be seen for various parameters at $t = 1$ on the computational grid using 100 time steps.

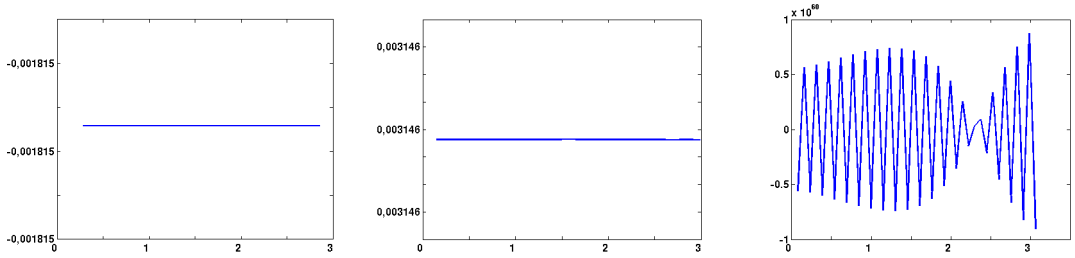


Figure 4.3.1: Linear interpolation of the relative error (proportion of the error and the analytic solution) at the grid points in the case of 10 (left) 20 (middle) and 40 (right) internal grid points.

It is really strange that incresing the spatial accuracy (taking more grid points) the aprroximation results can be worse. This clearly demonstrates that some mathematical analysis is necessary to have a somewhat deeper insight into these numerical methods. In the remaining part of lecture we want to track this problem and develop numerical methods to avoid this phenomenon.

4.3.4 Theoretical basis

We investigate the general problem

$$\partial_t u = Lu,$$

where L is a differential operator which consists the boundary conditions, as well. We use the notation u_{disc} for the restriction of u to the set $(0, \delta, \dots, T) \times \Omega_h$.

It is natural to require that the approximations of the both sides are accurate:

$$\partial_t u(n\delta, x_i, y_j) \approx [D_t u_{\mathrm{disc}}]^n_{j,k}$$

and

$$Lu(n\delta, x_i, y_j) \approx [D_L u_{\mathrm{disc}}]^n_{j,k}.$$

We should also demand that the consecutive time steps should not increase too much the (approximation) error. Anyway, in the solution process

$$\mathbf{u}^0 \xrightarrow{Q_{\mathbf{h},\delta,1}} \mathbf{u}^1 \xrightarrow{Q_{\mathbf{h},\delta,2}} \mathbf{u}^2 \dots \xrightarrow{Q_{\mathbf{h},\delta,N}} \mathbf{u}^N \tag{4.3.6}$$

the error will increase as we perform time/space refinement.

Here the given operators $Q_{\mathbf{h},\delta,1}, Q_{\mathbf{h},\delta,2}, \dots, Q_{\mathbf{h},\delta,N}$ are called the *time step operators*.

For the precise setting we observe that for the analytic solution u the left and the right hand side coincide:

$$\partial_t u - Lu = 0.$$

Definition 4.3.1. *We call the approximation given with D_t and D_L consistent in order $(\alpha_0, \alpha_1, \dots, \alpha_d)$ with respect to the $\|\cdot\|_*$ norm if for the analytic solution u*

$$\|D_t u_{disc} - D_L u_{disc}\|_* = \mathcal{O}(\delta^{\alpha_0}) + \mathcal{O}(h_1^{\alpha_1}) + \dots + \mathcal{O}(h_d^{\alpha_d}). \tag{4.3.7}$$

Definition 4.3.2. *If there is a constant C_{stab} such that for each vector $\mathbf{v}$*

$$Q_{\mathbf{h},\delta,1} Q_{\mathbf{h},\delta,2} \cdots Q_{\mathbf{h},\delta,N} \|\mathbf{v}\|_* \le C_{stab} \|\mathbf{v}\|_*, \tag{4.3.8}$$

independently of $N = T/\delta$ and h then the method is called unconditionally stable.

If the inequality in (4.3.8) holds only under a certain condition (mostly depending on the discretization parameters) then the corresponding method is called conditionally stable under this condition.

The fundamental theorem in the analysis can be written shortly as

$$\text{stability} + \text{consistency} \Leftrightarrow \text{convergence},$$

which was published by Lax and Richtmyer in 1956. In a bit more details:

Theorem 4.3.3. *If the method is consistent of order $(\alpha_0, \alpha_1, \dots, \alpha_d)$ with respect to the $\|\cdot\|_*$ norm and stable with respect to the $\|\cdot\|_*$ norm then the method is convergent of order $(\alpha_0, \alpha_1, \dots, \alpha_d)$ in the $\|\cdot\|_*$ norm,*

where the notion of the convergence has still to be clarified as the approximations are vectors with finite entries while, the analytic solution is a continuous function. Indeed, the converse of the theorem is also true.

In the remaining part, we discuss how the conditions of the above theorem can be verified for different numerical methods.

A detailed exposition of the theoretical basis can be found in [6] and [7].

4.4 Tools for the analysis

4.4.1 Consistency analysis: finite differences

The consistency order of the approximations are calculated usually using Taylor expansions. This is only justified if the analytic solution of the problem under investigation is sufficiently smooth (differentiable). For diffusion equations, it is satisfied automatically, and for wave equations and advection equations it depends on the boundary condition. This requirement is a weakness of the finite difference methods.

Instead of the norm consistency, we usually verify that the approximation order in (4.3.7) has pointwise a sufficient order.

Example 4.4.1. *Central difference approximation of the first order derivative:*

$$\partial_x u(t,x) \approx \frac{1}{2h}(u(t,x+h) - u(t,x-h)).$$

Here we obtain using Taylor expansion that

$$u(t,x+h) = u(t,x) + h\partial_x u(t,x) + \frac{h^2}{2}\partial_{xx}u(t,x) + \frac{h^3}{6}\partial_{xxx}u(t,x) + \mathcal{O}(h^4)$$

and similarly

$$u(t,x-h) = u(t,x) - h\partial_x u(t,x) + \frac{h^2}{2}\partial_{xx}u(t,x) - \frac{h^3}{6}\partial_{xxx}u(t,x) + \mathcal{O}(h^4),$$

which give then

$$\frac{1}{2h}(u(t,x+h)-u(t,x-h)) = \partial_x u(t,x)+\frac{h^2}{12}\partial_{xxx}u(t,x)+\mathcal{O}(h^4) = \partial_x u(t,x)+\mathcal{O}(h^2)$$

provided that $\partial_{xxx}u(t,x)$ is continuous in a neighborhood of (t,x).

4.4.2 Stability analysis: discrete time Fourier transform

We investigate l_2 stability for approximation vectors of form $\mathbf{v} = \ldots, v_{-1}, v_0, v_1, v_2, \ldots$, i.e. for the case if $\Omega = \mathbb{R}$.

Definition 4.4.2. *(discrete time Fourier transform) We define the mapping $\mathcal{F} : l_2 \to L_2[-\pi,\pi]$ as follows*

$$\mathcal{F}(\mathbf{u})(s) = \frac{1}{\sqrt{2\pi}}\sum_{k=-\infty}^{\infty} e^{-iks}u_k,$$

where l_2 is the linear space of series $\ldots, v_{-1}, v_0, v_1, v_2, \ldots$ for which $\sum_{j=-\infty}^{\infty} v_j^2 < \infty$.

One can easily verify the following.

- $\mathcal{F}$ is an isometry, i.e.
$$\|\mathcal{F}(\mathbf{u})\|_2 = \|\mathbf{u}\|_2,$$
we we have used the symbol $\|\cdot\|_2$ for two different norms.
- The inverse $\mathcal{F}^{-1} : L_2[-\pi,\pi] \to l_2$ of $\mathcal{F}$ can be given as
$$(\mathcal{F}^{-1}g)_k = \frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi} e^{iks}g(s)\,\mathrm{d}s.$$

Some important discrete time Fourier transform

If $_+$ and $_-$ denotes the right and the left-shift operator on l_2, we have

- $\mathcal{F}(\mathbf{v}_+ - \mathbf{v}) = (e^{is} - 1)\mathcal{F}(\mathbf{v})$,
- $\mathcal{F}(\mathbf{v} - \mathbf{v}_-) = (1 - e^{-is})\mathcal{F}(\mathbf{v})$,
- $\mathcal{F}(\mathbf{v}_+ - \mathbf{v}_-) = 2i \sin s \cdot \mathcal{F}(\mathbf{v})$.

We will investigate the coefficient $\rho : [-\pi, \pi] \to \mathbb{R}$ with

$$\rho(s) = \frac{\mathcal{F}(\mathbf{u}^n)(s)}{\mathcal{F}(\mathbf{u}^{n-1})(s)},$$

which is called the *amplification factor*, which indeed, depends on δ and h. Then we also have

$$\mathcal{F}(\mathbf{u}^n)(s) = \rho^n(s)\mathcal{F}(\mathbf{u}^{n-1})(s)$$

and therefore, we also obtain the following.

Theorem 4.4.3. *If $|\rho(s)| \leq 1$ for each $s \in [-\pi, \pi]$ then the corresponding scheme is stable.*

This approach has two drawbacks: first the theory can be applied only for constant coefficient problems and only on infinite domain.

4.4.3 Matrix analysis

To investigate real-life problems, we have to take a bounded Ω such that the operators in (4.3.6) become matrices and for the case $Q_{\mathbf{h},\delta,j} = Q$ stability is satisfied if and only if

$$\|\mathbf{u}^N\|_* \leq C\|Q^N\mathbf{u}^0\|_*$$

is valid for some mesh-independent constant C. A trivial sufficient condition is given in the following:

Theorem 4.4.4. *The numerical scheme with the constant one-step operator Q is stable in the $\|\cdot\|_2$ norm if*

- $\|Q\|_2 \leq 1$
- $s(Q) = \max\{|\lambda| : \lambda \textit{is an eigenvalue of } Q\} \leq 1$, *whenever Q is symmeric.*

Further details, see [6] and [7].

4.5 Numerical methods for diffusion problems

We again investigate

$$\begin{cases} \partial_t u(t,x) = \sigma_D \partial_{xx} u(t,x), & x \in \mathbb{R},\ t \in \mathbb{R}^+ \\ u(0,x) = u_0(x), & x \in \mathbb{R} \end{cases}$$

now on an infinite domain. The corresponding scheme is

$$\begin{cases} u_k^0 = u_0(kh), & k \in \mathbb{Z} \\ u_k^{n+1} = u_k^n + \frac{\sigma_D \delta}{h^2}(u_{k-1}^n - 2u_k^n + u_{k+1}^n), & k \in \mathbb{Z},\ n = 0, 1, 2, \dots \end{cases} \tag{4.5.1}$$

Proposition 4.5.1. *The scheme in* (4.5) *is conditionally stable: it is stable with respect to the l_2-norm if and only if $r \leq \frac{1}{2}$.*

A tiny modification in the scheme can make it undonditionally convergent:

Proposition 4.5.2. *The scheme*

$$\begin{cases} u_k^0 = u_0(kh), \quad k \in \mathbb{Z} \\ u_k^{n+1} = u_k^n + \frac{\sigma_D \delta}{h^2}(u_{k-1}^{n+1} - 2u_k^{n+1} + u_{k+1}^{n+1}), \quad k \in \mathbb{Z},\ n = 0, 1, 2, \dots \end{cases}$$

is unconditionally stable.

This method is *implicit* since the time steps can not be given with a simple explicit formula but one have to solve a linear system in each time step. One can prove similar statements also in the case of a bounded bounded domain. This requires the stability results on matrices in Section 4.4.3.

4.6 Numerical methods for advection problems

The numerical methods will be demonstrated for the simple advection problem

$$\begin{cases} \partial_t u(t,x) + a\partial_x u(t,x) = 0 \quad t \in \mathbb{R}^+,\ x \in I = (b_l, b_r) \\ u(0,x) = u_0(x) \quad x \in I \\ u(t, b_l) = u_{\text{left}}(t) \quad \text{if } \ b_l \neq -\infty \end{cases} \tag{4.6.1}$$

where the initial function u_0 and the left-hand side values u_{left} (if applicable) are given and the advection speed a is positive. We use the parameter $R = a\frac{\delta}{h}$. The following statements can be verified.

Proposition 4.6.1. *The scheme*

$$\begin{cases} u_k^0 = u_0(kh), \quad k \in \mathbb{Z} \\ u_k^{n+1} = u_k^n - \frac{R}{2}(u_{k+1}^n - u_{k-1}^{n+1}), \quad k \in \mathbb{Z},\ n = 0, 1, 2, \dots \end{cases}$$

is consistent with (4.6.1) *for $I = \mathbb{R}$ and the order of consistency is* $(2, 2)$ *but it can never be stable.*

Proposition 4.6.2. *The schemes*

$$\begin{cases} u_k^0 = u_0(kh), \quad k \in \mathbb{Z} \\ u_k^{n+1} = u_k^n - R(u_k^n - u_{k-1}^{n+1}), \quad k \in \mathbb{Z},\ n = 0, 1, 2, \dots \end{cases}$$

and

$$\begin{cases} u_k^0 = u_0(kh), \quad 0 < kh < 1 \\ u_k^{n+1} = u_k^n - R(u_k^n - u_{k-1}^{n+1}), \quad 0 < kh < 1\ n = 0, 1, 2, \dots \\ u_0^n = u_{left}(n\delta), \quad n = 0, 1, 2, \dots \end{cases}$$

are both conditionally stable under the condition $R \leq 1$. The first corresponds to the case of the infinite domain and the second one to the case $I = (0, 1)$.

Since the condition $a > 0$ is crucial in the preceding statements, we also mention a scheme where this condition can be relaxed.

Proposition 4.6.3. *The scheme*

$$\begin{cases} u_k^0 = u_0(kh), \quad k \in \mathbb{Z} \\ u_k^{n+1} = \left(\frac{R^2}{2} + \frac{R}{2}\right) u_{k-1}^n + (1 - R^2) u_k^n + \left(\frac{R^2}{2} - \frac{R}{2}\right) u_{k+1}^n, \quad k \in \mathbb{Z},\ n = 0, 1, 2, \dots \end{cases}$$

is consistent with (4.6.1) *for* $I = \mathbb{R}$ *and the order of consistency is* $(2, 2)$. *It is stable if and only if* $|R| \leq 1$. *This is valid also for the modification of* (4.6.1) *with* $a < 0$.

For further schemes and key questions, see [1], [7].

4.7 Numerical methods for the one-dimensional wave-equations

A numerical methods will be given for the simple wave propagation problem

$$\begin{cases} \partial_{tt} u(t, x) = \partial_{xx} u(t, x) \quad t \in \mathbb{R}^+,\ x \in I = (b_l, b_r) \\ u(0, x) = u_0(x) \quad x \in I \\ \partial_t u(0, x) = g(x) \quad x \in I \\ u(t, b_l) = u_{\text{left}}(t),\ u(t, b_r) = u_{\text{right}}(t) \quad \text{if } b_l \neq -\infty \text{ and } b_l \neq -\infty \end{cases} \tag{4.7.1}$$

where the initial function u_0, the initial derivative g and the boundary values $u_{\text{left}}, u_{\text{right}}$ (if applicable) are given. We use the parameter $R = \frac{\delta}{h}$.

Here we have to apply a *two-step* method to obtain a consistent approximation of the second order spatial derivative.

The following statement can be verified.

Proposition 4.7.1. *For* $I = \mathbb{R}$ *the scheme*

$$\begin{cases} u_k^0 = u_0(kh), \quad k = \dots, -1, 0, 1, \dots \\ \frac{1}{\delta^2}(u_k^{n+1} - 2u_k^n + u_k^{n-1}) = \frac{1}{h^2}(u_{k+1}^n - 2u_k^n + u_{k-1}^n), \quad k \in \mathbb{Z},\ n = 1, 2, \dots \end{cases} \tag{4.7.2}$$

is consistent with the first equation in (4.7.1) *and the order of consistency is* $(2, 2)$. *Furthermore, it is stable if and only if* $|R| \leq 1$.

In the practice, we have to deal with problems on finite intervals and one should also give the approximation $\mathbf{u}^1$ to compute with the consecutive time steps in (4.7.2).

Proposition 4.7.2. *If we use the initialization*

$$u_k^1 = u_k^0 + \frac{\delta}{2} g(kh) + \frac{u_{k-1}^0 - 2u_k^0 + u_{k+1}^0}{2h^2} \tag{4.7.3}$$

then the order of consistency with (4.7.1) *remains* $(2, 2)$.

For the detailed analysis, we refer to [7].

4.8 References

[1] U.U.M. Ascher. *Numerical methods for evolutionary differential equations*. Computational science and engineering. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2008.

[2] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[3] E. Hairer, S. P. N**o**rsett, and G. Wanner. *Solving ordinary differential equations. I*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 1993. Nonstiff problems.

[4] David Kincaid and Ward Cheney. *Numerical analysis. Mathematics of scientific computing. Reprint of the 3rd ed. 2002 published by Brooks/Cole.* Providence, RI: American Mathematical Society (AMS), reprint of the 3rd ed. 2002 published by brooks/cole edition, 2009.

[5] J. D. Lambert. *Numerical Methods for Ordinary Differential Systems: the Initial Value Problem*. John Wiley & Sons, Inc., New York, NY, USA, 2000.

[6] J.C. Strikwerda. *Finite Difference Schemes And Partial Differential Equations*. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2004.

[7] J. W. Thomas. *Numerical partial differential equations: finite difference methods*, volume 22 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 1995.

Chapter 5

András Frank: The graph orientation problem

5.1 Introduction

By orienting an undirected edge $e = uv$, we mean the operation that replaces e by one of the two directed edges uv and vu. An orientation of an undirected graph G arises from G by orienting each edge of G. In the basic form of the orientation problem, we are interested in finding an orientation of G meeting some specified properties. A more general form arises when some edges are already oriented and only the undirected edges are requested (and allowed) to be oriented. In other words, in this case we want to find an orientation of a mixed graph.

The goal of this mini-course is to outline some of the basic results and techniques concerning orientations as well as to exhibit several applications. The interested reader can get deeper inside from the book Connections in Combinatorial Optimization [5].

One may be interested, for example, in a root-connected orientation which means that every node is reachable from a specified root-node r_0 along a one-way path. The reader will easily find a proof for the following observation.

Proposition 5.1.1. A graph G has a root-connected orientation if and only if G is connected.

Somewhat trickier, but still rather easy, is to prove the following pretty result of Robbins.

THEOREM 5.1.2 (Robbins)**.** An undirected graph G has a strongly connected *(or just, strong)* orientation if and only if G is 2-edge-connected.

Proof. Let s be a specified node and compute a Depth-first-search tree F of root s. Define an arborescence $\vec{F}$ by orienting the edges of F away from s. By a basic property of DFS trees, the unique path connecting the end-nodes of every non-tree edge determines a directed path in $\vec{F}$. Orient all the non-tree edges so as to form a directed circuit with this path (that is, toward s). We claim that each arc of the digraph D obtained in this way belongs to a di-circuit, and hence D is strongly connected. As a result of this construction, each non-tree edge belongs to a di-circuit. Let $f = uv$ be an arc of $\vec{F}$ and let X denote the subset of nodes reachable in $\vec{F}$ from v. Since G is 2-edge-connected, there is an edge e in G leaving X. Since there is no cross-edge to F, f belongs to the di-circuit defined by e. •

Problem 5.1.1. Find a necessary and suffcient condition for a mixed graph to have a **(a)** root-connected **(b)** strong orientation.

We call a digraph **smooth** if the in-degree and the out-degree of every node differ by at most 1.

Proposition 5.1.3. Every graph has a smooth orientation.

Proof. We may assume that G has at least one edge. If G includes a circuit K, then inductively we find first a smooth orientation of the subgraph $G - K$, and orient then K so as to obtain a one-way circuit. Suppose now that G is a forest. Then it has a node v of degree 1. Let $e = uv$ be the single edge incident to v. Inductively there is a smooth orientation of $G-e$. We may assume that $\varrho(v) \leq \delta(v)$ since the reorientation of a smooth digraph is also smooth. By orienting e toward v we obtain a smooth orientation of G. •

When the initial graph is Eulerian (that is, every degree is even) we obtain the following observation.

Proposition 5.1.4. An Eulerian graph has a di-Eulerian orientation.

After these easy orientation results, let us mention a significantly more difficult one.

THEOREM 5.1.5 (Nash-Williams, [14])**.** An undirected graph $G = (V, E)$ has a k-edge-connected orientation if and only if G is $2k$-edge-connected.

In this theorem the necessity of the given condition is straightforward since in a k-edge-connected digraph there are at least k edges entering X and at least k edges leaving X for every non-empty $X \subset V$. The proper difficulty lies in proving sufficiency.

In some cases, however, it is not straightforward even to figure out the right condition. For example, when does a mixed graph have a di-Eulerian orientation? Or, when does a graph have an orientation in which there are k edge-disjoint paths from a specified root-node to every other node?

It turns out that orientation theorems help answer questions arising in apparently remote areas of graph theory. The goal of this mini-course is to outline the basic orientation results as well as their applications.

5.2 Degree-constrained orientations

We pointed out in Proposition 5.1.3 that every undirected graph has a smooth orientation. The question naturally arises: When does a graph have an orientation in which the in-degree of each node belongs to a prescribed interval? A special case is when each interval has exactly one element, that is, the in-degree of each node is specified. Note that one could consider out-degrees as well, however the in-degree plus the out-degree of a node is its undirected degree, which is independent of the orientation, and therefore a constraint for the out-degree can easily be transformed to a constraint for the in-degree.

5.2.1 In-degree specification

The earliest result of this type is by Landau [11], who solved it for complete graphs. By the **in-degree sequence** (or **vector**) of a directed graph with n nodes, we mean a sequence consisting of the in-degrees of the nodes of the digraph.

THEOREM 5.2.1 (Landau)**.** A sequence $m_1 \geq m_2 \geq \cdots \geq m_n$ of non-negative integers is the in-degree sequence of a tournament if and only if

$$\sum_{i=1}^{n} m_i = \binom{n}{2} \tag{5.2.1}$$

and

$$\sum_{i=1}^{k} m_i \leq k(k-1)/2 + k(n-k) \quad (k = 1, \ldots, n) \tag{5.2.2}$$

hold, which are equivalent to requiring *(5.2.1)* and

$$\sum_{i=n-h+1}^{n} m_i \geq \binom{h}{2} \quad (h = 1, \ldots, n). \tag{5.2.3}$$

We prove the theorem in a more general form.

THEOREM 5.2.2 (Orientation lemma, Hakimi, [8])**.** For an undirected graph $G = (V, E)$ and a function $m : V \to \mathbf{Z}$ satifying $\widetilde{m}(V) = |E|$, the following are equivalent.

(A) G has an orientation so that $\dot{\varrho} = m$, that is,

$$\varrho(v) = m(v) \text{ for every node } v, \tag{5.2.4}$$

(B) $e_G \geq m$, that is,

$$e_G(X) \geq \widetilde{m}(X) \text{ for every subset } X \subseteq V \tag{5.2.5}$$

where $e_G(X)$ denotes the number of edges having at least one end-node in X.
(C) $i_G \leq m$, that is,

$$i_G(Y) \leq \widetilde{m}(Y) \text{ for every subset } Y \subseteq V \tag{5.2.6}$$

where $i_G(Y)$ denotes the number of edges induced by Y.

Proof. Since $e_G(X) + i_G(V - X) = |E| = \widetilde{m}(V) = \widetilde{m}(X) + \widetilde{m}(V - X)$, the equivalence of (5.2.5) and (5.2.6) is evident.

Suppose now that there exists a requested orientation. Then $e_G(X) = \sum[\varrho(v) : v \in X] + \delta(X) \geq \sum[m(v) : v \in X] = \widetilde{m}(X)$ holds for any subset $X \subseteq V$, and hence (A) implies (B).

Finally, suppose that (5.2.5) is met. The function e_G can easily be seen to be submodular, that is, $e_G(X) + e_G(Y) \geq e_G(X \cap Y) + e_G(X \cup Y)$. Call a subset X **tight** if $\widetilde{m}(X) = e_G(X)$. The empty set is tight and so is V by the hypothesis $\widetilde{m}(V) = |E|$.

Proposition 5.2.3. The intersection and the union of two tight sets X and Y are also tight.

Proof. $\widetilde{m}(X)+\widetilde{m}(Y) = e_G(X)+e_G(Y) \geq e_G(X\cap Y)+e_G(X\cup Y) \geq \widetilde{m}(X\cap Y)+\widetilde{m}(X\cup Y) = \widetilde{m}(X)+\widetilde{m}(Y)$ and the proposition follows. •

We proceed by induction on $\widetilde{m}(V)$. The statement is straightforward when $\widetilde{m}(V) = |E| = 0$, so we can assume that there is a node s for which $m(s) > 0$. The proposition implies that there is a unique largest tight set Z not containing s. There exists an edge $f = us$ for which $u \notin Z$, for otherwise $e_G(Z+s) = e_G(Z) = \widetilde{m}(Z) = \widetilde{m}(Z+s) - m(s) < \widetilde{m}(Z+s)$, and hence $Z+s$ would violate condition (5.2.5). Delete f and reduce the value of $m(s)$ by one. We claim that condition (5.2.5) also holds for the resulting graph G' and for the revised in-degree specification m'. Indeed, if a subset X would violate (5.2.5), then X would originally be a tight $u\bar{s}$-set. From the maximal choice of Z we would have $X \subseteq Z$, contrary to the assumption $u \notin Z$.

By induction, G' has an orientation of in-degree vector m', from which we obtain an orientation of G with in-degree vector m by adding the directed edge us. • •

Theorem 5.2.1 follows immediately from the Orientation lemma, since in a complete graph the number $i_G(X)$ of edges induced by a subset $X \subseteq V$ is the same for each h-element subset, namely, $h(h-1)/2$. Therefore, it suffices to require the condition $i_G(X) \leq \widetilde{m}(X)$ only for the h smallest values m_i.

Problem 5.2.1. At a chess tournament, the winner of a game gets one point, the loser no points, while both players get half point for a draw. In order to avoid fractions, multiply everything by two. Then the winner, for example, gets 2 points. Under this assumption, when can a sequence $m_1 \geq m_2 \geq \cdots \geq m_n$ be the final score of a chess tournament?

Research problem 5.2.2 (A. Iványi)**.** Decide if a sequence of n integers can be the final score of a football tournament of n teams. The winner of a game gets 3 points, the loser no point, while both teams get 1 point for a draw.

Euler orientations of mixed graphs

As mentioned above, an undirected Euler graph always has an Euler orientation. The Orientation lemma allows one to extend this observation to mixed graphs.

THEOREM 5.2.4 (Ford and Fulkerson)**.** Let $M = (V, A+E)$ be a mixed graph consisting of an undirected graph $G = (V, E)$ and a digraph $D = (V, A)$. It is possible to orient the edges of E in such a way that the resulting directed graph is Eulerian if and only if every node of M is incident to an even number of *(directed or undirected)* edges, that is,

$$\delta_D(v) + \varrho_D(v) + d_G(v) \text{ is even} \tag{5.2.7}$$

and

$$d_G(X) \geq \varrho_D(X) - \delta_D(X) \text{ holds for every subset } X \subseteq V. \tag{5.2.8}$$

Proof. Let $\varrho_{\vec{G}}$ and $\delta_{\vec{G}}$ denote, respectively, the in-degree and the out-degree functions of an orientation $\vec{G} = (V, \vec{E})$ of G. The digraph $D+\vec{G}$ is Eulerian if and only if $\varrho_D(v)+\varrho_{\vec{G}}(v) = \delta_D(v)+\delta_{\vec{G}}(v)$ for every node v. This equality is equivalent,

via $\varrho_{\vec{G}}(v) + \delta_{\vec{G}}(v) = d_G(v)$, to $\varrho_{\vec{G}}(v) = (\delta_D(v) - \varrho_D(v) + d_G(v))/2$. Denote the right-hand side by $m(v)$. By (5.2.7), m is integer-valued. Apply theorem 5.2.2 and observe that the requirements (5.2.8) and (5.2.5) are equivalent for the specified m. •

5.2.2 Upper and lower bounds

The Orientation lemma follows immediately from a more general result of Hakimi [8], in which lower bounds, rather than exact values, are given for the in-degrees. This problem is equivalent to that of imposing upper bounds on the out-degree of the nodes. We now combine these two and solve the orientation problem when both lower and upper bounds are prescribed for the in-degree of the nodes. To this end, let $f : V \to \mathbf{Z}_+ \cup \{-\infty\}$ and $g : V \to \mathbf{Z}_+ \cup \{\infty\}$ be two functions for which $f \leq g$. (The lower bound $-\infty$ on a node means that there is no actual lower bound. We could have replaced $-\infty$ by zero, but $-\infty$ shows more directly that these nodes do not play any role. Analogous is the situation with the upper bound ∞.)

THEOREM 5.2.5. An undirected graph $G = (V, E)$ has an orientation for which

(A) $\varrho(v) \geq f(v)$ for every node v if and only if

$$e_G \geq \widetilde{f} \quad \text{(that is, } e_G(X) \geq \widetilde{f}(X) \text{ for every subset } X \subseteq V\text{)}, \tag{5.2.9}$$

(B) $\varrho(v) \leq g(v)$ for every node v if and only if

$$i_G \leq \widetilde{g}, \tag{5.2.10}$$

(C) $f(v) \leq \varrho(v) \leq g(v)$ for every node v if and only if both $e_G \geq \widetilde{f}$ and $i_G \leq \widetilde{g}$ hold.

Proof. If a requested orientation exists, then $\widetilde{f}(X) \leq \sum[\varrho(v) : v \in X] \leq e_G(X)$, and the necessity of (5.2.9) follows.

For proving sufficiency, assume (5.2.9). In an orientation of G we define a node s **deficient** if $\varrho(s) < f(s)$. Let us choose an orientation of G in which the total deficiency defined by the sum $\sum[f(v) - \varrho(v) : v \text{ deficient}]$ is minimum. If this sum is positive, then there is a deficient node s. Let X denote the set of nodes reachable from s in the given orientation. No directed edge leaves X, hence $\sum[\varrho(v) : v \in X] = e_G(X)$. Now X must contain a node t, for which $\varrho(t) > f(t)$. Otherwise $\widetilde{f}(X) > \sum[\varrho(v) : v \in X] = e_G(X)$, contradicting (5.2.9). By reorienting a directed path from s to t, we obtain another orientation of G in which the total deficiency is smaller, contradicting the choice of the original orientation. Therefore, the total deficiency must be 0, that is, there is no deficient node, and we are done.

Part (B) can be proved analogously (with the difference that a node t is deficient now if in the current orientation $\varrho(t) > g(t)$ and X denotes the set of nodes from which t is reachable). Alternatively, the second part is formally equivalent to that version of the first one when the out-degree of a node v is at least $f(v) := d_G(v) - g(v)$.

Finally, to see sufficiency in Part (C), let us start with an orientation of G for which $(*)$ $\varrho(v) \leq g(v)$ holds for every node v. Apply the algorithm of Part (A)

and observe that the in-degree of a node s can increase only if $\varrho(s) < f(s) \leq g(s)$, and hence $(*)$ remains automatically valid. •

Note that the Part (C) involves the first two parts.

Corollary 5.2.6. Let $G = (V, E)$ be an undirected graph with a specified subset $U \subseteq V$ of nodes and let $m : U \to \mathbf{Z}_+$ be an in-degree specification on U. There exists an orientation of G such that $\varrho(v) = m(v)$ for every $v \in U$ if and only if $i_G(X) \leq \widetilde{m}(X) \leq e_G(X)$ holds for every subset $X \subseteq U$.

Proof. Let $f(v) := g(v) := m(v)$ if $v \in U$, and $f(v) := -\infty$ and $g(v) := \infty$ if $v \in V - U$. Apply Theorem 5.2.5. •

It is worth emphasizing the following interesting consequence.

Corollary 5.2.7. Let f and g be integer-valued functions on E for which $f \leq g$. If the graph $G = (V, E)$ has an orientation for which $\varrho_1(v) \geq f(v)$ for every node v, and G has an orientation for which $\varrho_2(v) \leq g(v)$ for every node v, then there is an orientation of G meeting both requirements. •

This property is called the **linking** property. One of its earliest occurrences appeared in a paper of Mendelsohn and Dulmage [12]. It was formulated by Ford and Fulkerson [3] (Page 49) in a related theorem on the existence of integral matrices for which the row-sums and the column-sums lie between specified bounds. The concept was investigated in detail in the book of Mirsky [13]. The linking property shows up under much more complicated circumstances, too.

Note that the Orientation lemma follows immediately from Part (A) of Theorem 5.2.5.

Alternative proof of (the non-trivial part of) the Orientation lemma. Observe that for $f := m$, (5.2.5) and (5.2.9) are the same. Therefore, Theorem 5.2.5 implies the existence of an orientation of G for which $\varrho(v) \geq m(v)$ for every node v. Since $|E| = \sum[\varrho(v) : v \in V] \geq \sum[m(v) : v \in V] = \widetilde{m}(V) = |E|$, we must have equality for every node v, that is, $\varrho(v) = m(v)$. •

Exercise 5.2.3. Prove that if D_1 and D_2 are two orientations of the same undirected graph such that $\varrho_1(v) = \varrho_2(v)$ holds for each node v, then it is possible to get from D_1 to D_2 by a sequence of reorienting directed circuits.

Exercise 5.2.4. Show that if ϱ and ϱ' denote the in-degree functions of two orientations of G for which $\varrho(v) = m(v) = \varrho'(v)$ for every node v, then $\varrho(X) = \varrho'(X)$ for every subset $X \subseteq V$.

Problem 5.2.5. Develop a necessary and sufficient condition for the existence of an orientation where lower and uppers bounds are given for the in-degrees as well as for the out-degrees.

Alternative algorithmic proof: a push-relabel approach

Consider again Part (A) of Theorem 5.2.5, where the goal was to find an orientation for which the in-degrees obey lower bounds on the nodes. Suppose for a moment that we do not know yet the path reorientation technique used above and that we want to solve the orientation problem from scratch. A naive

approach would be to start with an arbitrary orientation, reorient any edge leaving an arbitrary node with $\varrho(v) < f(v)$, and repeat these edge reorientations as long as there are deficient nodes. Not surprisingly, such a primitive procedure need not terminate: For example, if e is an uv-edge and both u and v are deficient, then the algorithm may choose to reorient e every time, in which case there is no progress. However, if we introduce a clever control parameter (to be called level) on the nodes to select the proper deficient node and the proper leaving edge, then this approach does work. Precisely this approach is the idea of the push-relabel algorithm of Goldberg and Tarjan [7] that was developed for computing a maximum flow (and beat the alternating path method). The algorithm below is merely an adaptation of the algorithm of Goldberg and Tarjan for orientations. It may help the reader to capture the very essence of this technique in the present, particularly simple setting. In Part III, we shall see that the push-relabel approach extends well beyond network flows.

The procedure starts with an arbitrary orientation and works throughout with a non-negative integer-valued level function $\Theta : V \to \mathbf{Z}_+$ on the nodes. The following two level properties will be maintained:

(L1) Every oversaturated node v (that is, one with $\varrho(v) > f(v)$) is on level 0.
(L2) $\Theta(v) \geq \Theta(u) - 1$ holds for every directed edge uv.

The algorithm terminates when one of the following stopping rules holds.

(A) There are no more in-deficient nodes where v is in-deficient if $\varrho(v) < f(v)$.
(B) There exists an in-deficient node z and an empty level j under the level of z, that is, $j < \Theta(z)$ and $\{u \in V : \Theta(u) = j\} = \emptyset$.

Stopping rule (A) means that the current orientation satisfies the requirement $\varrho(v) \geq f(v)$ for every node v. We claim that Stopping rule (B) implies that $e_G(X) < \widetilde{f}(X)$ for subset $X := \{u : \Theta(u) \geq j\}$, and hence X violates (5.2.9). Indeed, on the one hand, X contains no oversaturated node by Property (L1), whereas it contains the in-deficient z, from which $\sum[\varrho(u) : u \in X] < \sum[f(u) : u \in X] = \widetilde{f}(X)$ follows. On the other hand, no edge leaves X by Property (L2), from which $e_G(X) = \sum[\varrho(u) : u \in X] < \widetilde{f}(X)$.

The algorithm runs as follows. At the beginning $\Theta \equiv 0$. As long as there are in-deficient nodes, select arbitrarily one, to be denoted by z. If there is an arc $e = zv$ with $\Theta(v) = \Theta(z) - 1$, then reorient e. When this operation makes Stopping rule (A) valid, the algorithm terminates by returning a requested orientation.

If there is no such an arc, then increase $\Theta(z)$ by one. When this operation makes Stopping rule (B) valid, the algorithm terminates by returning a subset X violating (5.2.9). Note that both operations maintain properties (L1) and (L2).

The level of a node can be increased at most n times, since once a node reaches level $n = |V|$ there must be an empty level under it, in which case (B) certainly holds. Therefore, the total number of level increases is at most n^2.

Since an edge is reoriented only if its head has a lower level than its tail, the sum $\Theta(u) + \Theta(v)$ increases by at least 2 between two consecutive reorientations of uv. Since the level of every node is less than n, the sum $\Theta(u) + \Theta(v)$ is at most $2n$, and hence every edge can be reoriented at most $n = 2n/2$ times. Therefore, the total number of reorientations is at most mn, implying that the overall complexity of the algorithm is $O(nm)$. •

5.3 Applications

5.3.1 Paths and matchings

Menger's theorem

There are several versions of Menger's theorem. Here we derive the directed edge-version.

THEOREM 5.3.1 (Menger)**.** Let $D = (V, A)$ be a digraph with a source-node s and a sink-node t so that no edge enters s and no edge leaves t. There are k edge-disjoint st-paths if and only if

$$\delta(X) \geq k \text{ holds for every } s\bar{t}\text{-set } X. \tag{5.3.1}$$

Proof. A necessity is straightforward, we prove only sufficiency.

Observe first that it suffices to construct a subgraph $D' = (V, A')$ of D in which

$$\delta_{D'}(s) = k, \varrho_{D'}(s) = 0 \text{ and } \varrho_{D'}(v) = \delta'(v) \text{ holds for every } v \in V - \{s, t\} \tag{5.3.2}$$

since in such a D' one can find in a greedy way the k edge-disjoint st-walks.

We assumed that $\varrho_D(s) = 0 = \delta_D(t)$. Let G denote the underlying undirected graph of D and define $m : V \to \mathbf{Z}$ by

$$m(v) := \begin{cases} \varrho_D(v) & \text{if } v \in V - \{s, t\} \\ k & \text{if } v = s \\ \varrho_D(t) - k & \text{if } v = t \end{cases} . \tag{5.3.3}$$

Claim 5.3.2. $\widetilde{m}(X) \leq e_G(X)$ for every $X \subseteq V$, with equality for $X = V$.

Proof. First, $\widetilde{m}(V) = \sum[m(v) : v \in V] = \sum[\varrho_D(v) : v \in V] + k - k = e_D(V) = e_G(V)$. Second, observe that (5.3.1) is equivalent to requiring that $\delta_D(X) \geq k(|X \cap \{s\}| - |X \cap \{t\}|)$ for *every* subset X of V. We have

$$\widetilde{m}(X) = \sum[\varrho_D(v) : v \in X] + k(|X \cap \{s\}| - |X \cap \{t\}|) = e_G(X) - \delta_D(X) + k(|X \cap \{s\}| - |X \cap \{t\}|)$$

from which $\widetilde{m}(X) \leq e_G(X)$ follows. •

By the Orientation lemma (Theorem 5.2.2), there is an orientation $\vec{G}$ of G with in-degree specification m. Let D' denote the subgraph of D consisting of those edges which have been reversed in $\vec{G}$. Because of the definition of m, D' satisfies the properties given by (5.3.2). • •

By adapting the path-reversing proof technique described in the proof of Theorem 5.2.5, one obtains the following simple algorithm for finding the requested degree-specified reorientation of D. Start with D. As long as there is a directed st-path in the current reorientation of D, find one and reorient its edges. Since this operation reduces the out-degree of each $s\bar{t}$-set by one, after k reorientations we shall have arrived at the one satisfying the requested in-degree specification m.

Factors of bipartite graphs

THEOREM 5.3.3 (Hall)**.** A bipartite graph $G = (S, T; E)$ has a perfect matching if and only of $|S| = |T|$ and

$$|\Gamma(Y_T)| \geq |Y_T| \text{ for every } Y_T \subseteq T \tag{5.3.4}$$

where $\Gamma(Y_T)$ denotes the subset of nodes in S having a neighbour in Y_T.

Proof. Necessity is being straightforward, we prove only sufficiency.

Observe that by orienting the edges of a perfect matching M toward S while all the other edges toward T we obtain an orientation of G in which the in-degree of every node $s \in S$ is $m(s) := 1$ and the in-degree of every node $t \in T$ is $m(t) := d_G(t) - 1$, and conversely, in an orientation of G with this in-degree specification m the set of edges entering S is a perfect matching.

By the Orientation lemma, there is an orientation (that is, there is a perfect matching) if and only if $\widetilde{m}(V) = |E|$, where $V = S \cup T$, and $\widetilde{m}(Y) \geq i(Y)$ for every $Y \subseteq V$. Clearly $\widetilde{m}(V) = \widetilde{m}(T) + \widetilde{m}(S) = \sum[d(t) - 1 : t \in T] + \sum[1 : t \in S] = |E| - |T| + |S| = |E|$. For a subset $Y \subseteq V$, let $Y_S := Y \cap S$ and $Y_T := Y \cap T$. By (5.3.4) we have $|Y_S| + d(Y, S) \geq |\Gamma(Y_T)| \geq |Y_T|$ from which $\widetilde{m}(Y) = |Y_S| + [d(Y_T) - |Y_T|] = |Y_S| + i(Y) + d(Y, S) - |Y_T| \geq i(Y)$. •

Problem 5.3.1. Let $G = (S, T; E)$ be a bipartite graph and let $b : S \cup T \to \mathbf{Z}_+$ be a function. There exists a subset $F \subseteq E$ of edges so that $d_F(v) = b(v)$ for every $v \in S \cup T$ if and only if $\widetilde{b}(S) = \widetilde{b}(T)$ and $\widetilde{b}(\Gamma(Y)) \geq \widetilde{b}(Y)$ holds for every $Y \subseteq T$.

Finding degree-constrained subgraphs

Let $G = (V, E)$ be an undirected graph. At every node v of G, we are given a set $F(v) \subseteq \{0, 1, \ldots, d_G(v)\}$ of forbidden degrees, where $d_G(v)$ denotes the degree of v. A subgraph $G' = (V, E')$ of G is called **F-avoiding** if $d_{G'}(v) \notin F(v)$ for every $v \in V$. The problem of deciding if there is an F-avoiding subgraph is **NP**-complete in general. Indeed, the hypergraph perfect matching problem (that seeks to find a partition of the node set consisting of hyperedges), which is known to be **NP**-complete even for 3-uniform hypergraphs, can easily be formulated this way. To this end, consider the bipartite graph $G = (V, U; E)$ associated with hypergraph $H = (V, \mathcal{F})$. Note that the degree of every node $u \in U$ is 3. At every node $v \in V$, let $F(v) := \{0, 2, 3, \ldots, d_G(v)\}$. This means that for $v \in V$ the set $F(v)$ the complementary set of $\{1\}$. At every node $u \in U$, let $F(u) = \{1, 2\}$. Clearly, there is a one-to-one correspondence between the perfect matchings of H and the F-avoiding subgraphs of G.

Intuition suggests that there must be an F-avoiding subgraph if each forbidden set $F(v)$ is sufficiently small and this natural feeling is formulated in the next result of Shirazi and Verstraëte [17].

THEOREM 5.3.4. If

$$|F(v)| \leq \lfloor d_G(v)/2 \rfloor \text{ for every } v \in V, \tag{5.3.5}$$

then G admits an F-avoiding subgraph.

Interestingly, the original proof of this disarmingly simple-sounding statement made use of a fundamental combinatorial result concerning polynomials, a result by Alon [1]. The simple combinatorial proof below neatly exemplifies the paradigm that it can be highly rewarding to figure out a proper extension of the statement in order to get a short and simple proof.

We have seen in Theorem 5.1.3 that every graph G has a smooth orientation $D = (V, \vec{E})$. In such an orientation, $\varrho_D(v) \geq \lfloor d_G(v)/2 \rfloor$ for every node v. Therefore, the following result of Frank, Lao, and Szabó [6] implies Theorem 5.3.4.

THEOREM 5.3.5. If a graph G has an orientation $D = (V, \vec{E})$ in which $\varrho_D(v) \geq |F(v)|$ for every node v, then G has an F-avoiding subgraph.

Proof. We proceed by induction on the number of edges. For an edge $e \in E$ of G, let $\vec{e}$ denote the corresponding directed edge of D. If 0 does not occur at any node as a forbidden degree, then the graph $(V, \emptyset)$ is an F-avoiding subgraph of G. Suppose now that $0 \in F(t)$ for some node t. Then $\varrho_D(t) \geq |F(t)| \geq 1$, and therefore there is an edge $e = st$ of G for which $\vec{e}$ is directed toward t.

Let $G^- := G - e$ and $D^- := D - \vec{e}$. Define F^- as follows.

$$F^-(v) := \begin{cases} \{i - 1 : i \in F(t) - \{0\}\} & \text{if } v \in \{s, t\}, \\ F(v) & \text{if } v \in V - \{s, t\}. \end{cases} \tag{5.3.6}$$

Since $|F^-(t)| = |F(t)| - 1$, we have $\varrho_{D^-}(v) \geq |F^-(v)|$ for every node v. By induction, G^- has an F^--avoiding subgraph G''. It follows from the definition of F^- that the subgraph $G' := G'' + e$ of G is F-avoiding. •

This approach shows that the hypothesis (5.3.5) in Theorem 5.3.4 can be made more flexible. By combining Theorem 5.3.5 with Part (A) of Theorem 5.2.5, one obtains the following.

THEOREM 5.3.6. Suppose that $e_G(X) \geq \sum[|F(v)| : v \in X]$ holds for every subset $X \subseteq V$ of nodes of an undirected graph G. Then G has an F-avoiding subgraph. •

5.3.2 Sparsity and list-colouring

Recognizing k-sparse graphs

A interesting feature of graph orientation problems is that, in several applications, our primary interest is not the existence of a certain orientation. Instead, we want to check, for example, the validity of criterion (5.2.9) or (5.2.10). We call a subset $Z \subseteq V$ of nodes of an undirected graph $G = (V, E)$ k**-sparse** if

$$i_G(X) \leq k(|X| - 1) \tag{5.3.7}$$

holds for every non-empty subset $X \subseteq Z$. The graph G itself is also said to be k**-sparse** if V is k-sparse. A classic theorem of Nash-Williams [15] states that the edge-set of G can be partitioned into k forests if and only if G is k-sparse. Here we show how it is possible to test a graph for k-sparsity with the help of degree-constrained orientations.

It follows from Part (B) of Theorem 5.2.5 that there is an orientation of G so that the in-degree of a specified node s is 0 while all other in-degrees are at most

k if and only if $i_G(X) \leq k(|X|-1)$ holds for every subset $X \subseteq V$ containing s and $i_G(X) \leq k|X|$ holds for every subset $X \subseteq V$ not containing s. Therefore all we need to do is to check if there is such an orientation for every possible choice of $s \in V$. If the answer is yes for every s, then the graph is k-sparse, while if the answer is no for at least one $s \in V$, then the algorithm returns a subset X violating (5.3.7), showing that the graph is not k-sparse.

List-colouring of bipartite graphs

The chromatic number of a graph G is the smallest integer k for which the nodes of G can be coloured with k colours so that no colour class induces any edge. In the list-colouring problem, each node of G admits a list of colours and we are interested in a colouration of nodes so that the colour of each node is taken from its list. G is said to be k list-colourable if there is a list-colouration whenever the list of each node has at least k colours. The smallest such k is the **list-chromatic** number of G. Clearly, the size of a maximum clique is a lower bound for the chromatic number and hence it is a lower bound for the list-chromatic number.

The list-chromatic number can be arbitrarily larger than the chromatic number even for bipartite graphs. However an upper bound depending on the edge density can be obtained.

We say that a stable subset K of a digraph $D = (V, A)$ is a **kernel** if there is an edge $uv \in A$, $u \in K$ for every node $v \in V - K$. A directed circuit of odd length shows that not every digraph has a kernel.

THEOREM 5.3.7 (Fleiner)**.** The union of two transitive and acyclic digraphs has a kernel. Equivalently, given two posets P_1 and P_2 on a common groundset V, there is an antichain A in common such that, for every element $x \in V - A$, there is a $p \in A$ which is larger than x in at least one of the two posets.

Proof. Let A_1 be the set of maximal elements in P_1. This is clearly an antichain in P_1. If A_1 happens to be an antichain of $P2$, too, then A_1 will do.

Suppose now that there are elements $p, y \in A_1$ for which p is larger than y in P_2. By induction, after deleting y there is an antichain A' in common with the required property. If $p \in A'$, then A' will do for the original posets. If $p \notin A'$, then there is an element $p' \in A'$ that is larger than p in one of the posets. Due the the definition of A_1, p' is larger than p in P_2. Now the transitivity implies that p' is larger than y in P_2. •

The following is a special case.

THEOREM 5.3.8. A directed bipartite graph has a kernel. •

THEOREM 5.3.9. If a bipartite graph G has an orientation in which every degree is at most k *(that is, by the Orientation lemma, every subset Z of nodes induces at most $k|Z|$ edges)*, then the list-clouring number of G is at most $k+1$.

Proof. We prove the following stronger statement.

Let D be an orientation of a bipartite graph $G = (V, E)$. For each $v \in V$ we are given a list $L(v)$ of colours of size at least $\varrho_D(v) + 1$. Then G admits a list-colouring.

To see this, let s be a colour appearing in a list and consider the set V_s of nodes whose list contains s. By Theorem 5.3.9 the subdigraph D_s induced by V_s includes a kernel K_s. Assign colour s to each node in K_s and delete s from each list. In the subgraph $D' = D - K_s$, for the reduced lists we claim that $|L'(v)| \geq \varrho_{D'}(v)+1$. This clearly holds for a node v with $|L'(v)| = |L(v)|$ so assume that $|L'(v)| = |L(v)|-1$. Then $v \in V_s - K_s$. Since K_s is a kernel in D_s, there is a directed edge from K_s to v. Therefore $\varrho_{D'}(v) < \varrho_D(v)$ and hence $|L'(v)| = |L(v)| - 1 \geq \varrho_D(v) \geq \varrho_{D'}(v) + 1$, as required.

By induction, there is a list-colouring of D' with respect to the lists $L'(v)$. By adding K_s as an additional class coloured by s, we obtain a list-colouring of G. •

For planar bipartite graphs we have the following corollary.

THEOREM 5.3.10 (Alon and Tarsi, [2])**.** The list-colouring number of a planar bipartite graph is at most 3.

Proof. We may assume that G is simple. Let n and m denote the number of nodes and edges, respectively. In a simple planar bipartite graph each region s bounded by at least 4 edges. Therefore the number t of regions is at most $m/2$.

By Euler's formula, $t + n = m + 2$ from which $m/2 + n \geq m + 2$ and hence $m \leq 2n - 4$. Since every subgraph of G is also planar and bipartite, it follows that each subset Z of nodes of G induces at most $2|Z| - 4$ edges. By the Orientation lemma, G admits an orientation of D in which every in-degree is at most 2. By Theorem 5.3.9, G is 3-list-colourable. •

Problems

5.3.2. Let ℓ and k be non-negative integers where $\ell \leq k$. Extend the approach above to test, for a graph G, whether $i_G(X) \leq k|X| - \ell$ holds for every non-empty subset $X \subseteq V$.

5.3.3 (Kampen, [10])**.** Let G be a simple planar graph. Relying on the Euler's formula, prove that G has an orientation in which every in-degree is at most 3.

5.3.4. Let $G = (V, E)$ be a simple maximal *(that is, triangulated)* planar graph on at least five nodes. Prove that E can be partitioned into claws, where a claw is the complete bipartite graph $K_{1,3}$.

5.3.5. Adapt the graph-orientation idea above to test in a graph if $i_G(X) \leq \widetilde{m}(X) - \ell$ holds for every non-empty subset X of V where $m : V \to \mathbf{Z}_+$ is a non-negative integer-valued function and ℓ is a non-negative integer for which $\ell \leq m(v)$ for every $v \in V$.

5.3.6. Develop a *(slightly more sophisticated)* approach to test in a graph if $i_G(X) \leq 2|X| - 3$ holds for every subset $X \subseteq V$ with at least 2 elements.

5.4 References

[1] N. Alon, *Combinatorial Nullstellensatz*, in: Recent trends in combinatorics (Mátraháza, 1995). Combin. Probab. Comput. 8. no.1-2 (1999), 7-29.

[2] N. Alon and M. Tarsi, *Colorings and orientations of graphs*, Combinatorica 12 (1992) 125-134.

[3] L.R. Ford and D.R. Fulkerson, Flows in Networks, Princeton Univ. Press, Princeton NJ., 1962.

[4] A. Frank and A. Gyárfás, *How to orient the edges of a graph,* in: Combinatorics, (Keszthely 1976), Coll. Math. Soc. J. Bolyai 18, 353-364, North-Holland.

[5] A. Frank, Connections in Combinatorial Optimization, Oxford University Press, 2011 (ISBN 978-0-19-920527-1). Oxford Lecture Series in Mathematics and its Applications, 38.

[6] A. Frank, L.C. Lao, and J. Szabó, *A note on degree-constrained subgraphs*, Discrete Mathematics, (2008) Vol 308, 12, 2647-2648.

[7] A.V. Goldberg and R.E. Tarjan, *A new approach to the maximum-flow problem*, J. Association for Computing Machinery, 35 (1988) 921-940.

[8] S.L. Hakimi, *On the degrees of the vertices of a directed graph*, J. Franklin Inst. 279 (4) (1965) 290-308.

[9] P. Hall, *On representatives of subsets*, J. London Math. Soc., Vol. 10 (1935) 26-30.

[10] G.R. Kampen,*Orienting planar graphs*, Discrete Mathematics, Volume 14, Issue 4, (1976) 337-341.

[11] H.G. Landau, *On dominance relations and the structure of animal societies III. The condition for the core structure*, Bull. Math. Biophys. 15 (1953) 143-148.

[12] N.S. Mendelsohn and A.L. Dulmage, *Some generalizations of the problem of distinct representatives*, Canadian Journal of Mathematics, 10 (1958) 230-241.

[13] L. Mirsky, Transversal Theory, Academic Press, 1971.

[14] C.St.J.A. Nash-Williams, *On orientations, connectivity and odd vertex pairings in finite graphs*, Canad. J. Math. 12 (1960) 555-567.

[15] C.St.J.A. Nash-Williams, *Decomposition of finite graphs into forests,* J. London Math. Soc. 39 (1964) 12.

[16] H.E. Robbins, *A theorem on graphs with an application to a problem of traffic control*, American Math. Monthly **46** (1939) 281–283.

[17] H. Shirazi and J. Verstraëte, *A note on polynomials and f-factors of graphs*, Electronic Journal of Combinatorics, 15 (2008) No. 22.

Chapter 6

Péter E. Frenkel: Algebraic inequalities and sums of squares

Many of the most important inequalities in mathematics are, or can be reformulated as, algebraic inequalities. An algebraic inequality is one which asserts that some given polynomial is nonnegative everywhere (or is nonnegative on some specified set).

6.1 Inequalities between means

As examples, consider the inequalities listed below. In each of these, the variables x_i, X_i, Y_i are meant to be *nonnegative* reals.

- The inequality
$$\sqrt[q]{\frac{1}{n}\sum_{i=1}^{n} x_i^q} \leq \sqrt[p]{\frac{1}{n}\sum_{i=1}^{n} x_i^p} \tag{6.1.1}$$
between power means. This holds for any real exponents $p \geq q > 0$, and can be rewritten as an algebraic inequality when p and q are integers.

- The more general inequality of Lyapunov:
$$\left(\sum_{i=1}^{n} x_i^q\right)^{p-r} \leq \left(\sum_{i=1}^{n} x_i^p\right)^{q-r} \left(\sum_{i=1}^{n} x_i^r\right)^{p-q}. \tag{6.1.2}$$
This holds for any real exponents $p \geq q \geq r \geq 0$, and is an algebraic inequality when p, q and r are integers. Note that the special case $r = 0$ is the preceding power mean inequality.

- Maclaurin's inequality
$$\sqrt[q]{\binom{n}{q}^{-1} \sum_{1\leq i_1<\cdots<i_q\leq n} x_{i_1}\cdots x_{i_q}} \geq \sqrt[p]{\binom{n}{p}^{-1} \sum_{1\leq i_1<\cdots<i_p\leq n} x_{i_1}\cdots x_{i_p}} \tag{6.1.3}$$
between elementary symmetric means. This holds for integers $n \geq p \geq q \geq 1$, and can be rewritten as an algebraic inequality. Note that the special case $q = 1$, $p = n$ is the inequality between the arithmetic and the geometric mean.

- A Lyapunov type generalization of Maclaurin's inequality:

$$\left(\frac{\sum_{i_1<\cdots<i_q} x_{i_1}\cdots x_{i_q}}{\binom{n}{q}}\right)^{p-r} \geq$$
$$\geq \left(\frac{\sum_{i_1<\cdots<i_p} x_{i_1}\cdots x_{i_p}}{\binom{n}{p}}\right)^{q-r} \left(\frac{\sum_{i_1<\cdots<i_r} x_{i_1}\cdots x_{i_r}}{\binom{n}{r}}\right)^{p-q}$$

  for the elementary symmetric means. This holds for integers $n \geq p \geq q \geq r \geq 0$, and is an algebraic inequality. Note that the special case $r = 0$ is Maclaurin's inequality, and the special case $r = q-1$, $p = q+1$ is Newton's well-known inequality. This latter special case is easily seen to imply the general case.

- Minkowski's inequality (superadditivity of the geometric mean):

$$\sqrt[n]{\prod_{i=1}^{n} X_i} + \sqrt[n]{\prod_{i=1}^{n} Y_i} \leq \sqrt[n]{\prod_{i=1}^{n}(X_i + Y_i)}.$$

One way of proving an algebraic inequality $f \geq 0$ is to rewrite f as a sum of squares. A polynomial $f \in \mathbb{R}[x]$ which is nonnegative on the real line is always a sum of two squares of polynomials. This is an easy consequence of the fundamental theorem of algebra. Minkowski conjectured that this fails for multivariate polynomials, i.e., a polynomial $f \in \mathbb{R}[x_1, \ldots, x_n]$ that is nonnegative everywhere on $\mathbb{R}^n$ is not necessarily a sum of squares (of any number of polynomials). Minkowski's conjecture was proved by Hilbert. The simplest known example showing that Minkowski was right was given by Motzkin. This example is given below, after Exercise 5.

However, if the inequality $f \geq 0$ is 'classical' and 'famous' enough, then f usually turns out to be representable as a sum of squares, although such a representation is not always easy to find. For example, the most standard proof of the Cauchy–Schwarz inequality is *not* the one that rewrites the difference of the two sides as a sum of squares, but such a rewriting is possible (and almost as well known). More interestingly, the inequality between the arithmetic and the geometric mean also has such a proof, as was demonstrated by Hurwitz [6] in 1891. The paper of Fujisawa [4] gives numerous further examples of this phenomenon.

Such a purely algebraic proof of an algebraic inequality, even if it is not the simplest proof, gives some extra understanding of why the inequality 'must be true'.

In many cases *binomial squares*, i.e., squares of binomials $ax^\alpha + bx^\beta$ suffice. Here x^α is a shorthand for $x_1^{\alpha_1}\cdots x_n^{\alpha_n}$, and the coefficients a and b are real numbers. For example, for $n = 2$,

$$\frac{x^2+y^2}{2} - \left(\frac{x+y}{2}\right)^2 = \left(\frac{x-y}{2}\right)^2$$

is a binomial square.

In the course, I will explain — following [5] — how positive polynomials arising from some of the inequalities listed above can be represented as sums of binomial squares. As a warm-up, the reader is encouraged to try solving the following exercises before reading the solutions.

1. (a) Can
$$f(x,y,z) = \frac{x^2+y^2+z^2}{3} - \left(\frac{x+y+z}{3}\right)^2$$
be written as the square of a polynomial?

   (b) Can it be written as a sum of squares of polynomials?

   Solution. (a) No. Since the given polynomial f is homogeneous of degree 2, the square root would have to be homogeneous of degree 1, i.e., of the form $g(x,y,z) = ax + by + cz$. Comparing coefficients in f and g^2 shows that $f = g^2$ is impossible.

   (b) Yes, e.g.
$$f(x,y,z) = \left(\frac{x-y}{3}\right)^2 + \left(\frac{y-z}{3}\right)^2 + \left(\frac{z-x}{3}\right)^2.$$

2. (a) More generally, can
$$\frac{x_1^2+\cdots+x_n^2}{n} - \left(\frac{x_1+\cdots+x_n}{n}\right)^2$$
be written as a sum of squares of polynomials?

   (b) Same question for
$$\frac{x^3+y^3}{2} - \left(\frac{x+y}{2}\right)^3.$$

   Solution. (a) Yes, e.g.
$$\sum_{i<j}\left(\frac{x_i-x_j}{n}\right)^2.$$

   (b) Of course not! It is negative for $x = y = -1$.

Classical inequalities often involve *nonnegative* real variables as opposed to real variables. In the setting of nonnegative variables, the suitable analog of the semiring of sums of squares is the semiring

$$S = S_n = \left\{ \sum_{\varepsilon_1=0}^{1} \cdots \sum_{\varepsilon_n=0}^{1} r_{\underline{\varepsilon}} \prod_{j=1}^{n} x_j^{\varepsilon_j} \mid r_{\underline{\varepsilon}} \text{ is a sum of squares in } \mathbb{R}[x_1,\ldots,x_n] \right\}.$$

It is immediately seen that S is indeed a semiring, i.e., it is closed under addition and multiplication. In fact, S is the semiring generated by the variables x_1, ..., x_n and by the squares of all polynomials.

Note that $p \in S$ implies that p is nonnegative for $x_1,\ldots,x_n \geq 0$, but not conversely. Clearly, p is nonnegative for $x_1,\ldots,x_n \geq 0$ if and only if $p(x_1^2,\ldots,x_n^2)$ is nonnegative everywhere. The relevance of the semiring S is explained by the following Lemma.

Lemma 6.1.1 ([5]). *Let $p \in \mathbb{R}[x_1,\ldots,x_n]$. Then $p \in S$ if and only if $p(x_1^2,\ldots,x_n^2)$ is a sum of squares in $\mathbb{R}[x_1,\ldots,x_n]$.*

Proof. The 'only if' part is trivial. For the 'if' part, we consider the linear operator

$$\mathcal{R} : \mathbb{R}[x_1, \ldots, x_n] \to \mathbb{R}[x_1^2, \ldots, x_n^2]$$

that maps the monomial $\prod x_j^{k_j}$ to itself if all k_j are even and maps it to zero otherwise. We assume that $p(x_1^2, \ldots, x_n^2)$ is a sum of squares, i.e., there exist polynomials r_i such that

$$q(x_1, \ldots, x_n) := p(x_1^2, \ldots, x_n^2) = \sum_{i=1}^{k} r_i^2(x_1, \ldots, x_n).$$

In r_i, we group terms according to the parity of the exponents of x_1, …, x_n. We define the polynomials $r_{i,\underline{\varepsilon}}$ for each $\underline{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \in \{0,1\}^n$ so that

$$r_i(x_1, \ldots, x_n) = \sum_{\varepsilon_1=0}^{1} \ldots \sum_{\varepsilon_n=0}^{1} r_{i,\underline{\varepsilon}}(x_1^2, \ldots, x_n^2) \prod_{j=1}^{n} x_j^{\varepsilon_j}.$$

Apply $\mathcal{R}$ to r_i^2, then

$$(\mathcal{R}r_i^2)(x_1, \ldots, x_n) = \sum_{\varepsilon_1=0}^{1} \ldots \sum_{\varepsilon_n=0}^{1} r_{i,\underline{\varepsilon}}^2(x_1^2, \ldots, x_n^2) \prod_{j=1}^{n} x_j^{2\varepsilon_j}.$$

Hence,

$$p(x_1^2, \ldots, x_n^2) = q(x_1, \ldots, x_n) = (\mathcal{R}q)(x_1, \ldots, x_n) =$$

$$= \sum_{i=1}^{k} (\mathcal{R}r_i^2)(x_1, \ldots, x_n) = \sum_{i=1}^{k} \sum_{\varepsilon_1=0}^{1} \ldots \sum_{\varepsilon_n=0}^{1} r_{i,\underline{\varepsilon}}^2(x_1^2, \ldots, x_n^2) \prod_{j=1}^{n} x_j^{2\varepsilon_j}.$$

Therefore,

$$p(x_1, \ldots, x_n) = \sum_{i=1}^{k} \sum_{\varepsilon_1=0}^{1} \ldots \sum_{\varepsilon_n=0}^{1} r_{i,\underline{\varepsilon}}^2(x_1, \ldots, x_n) \prod_{j=1}^{n} x_j^{\varepsilon_j},$$

whence $p \in S$. □

Let us return to our exercises.

3. (a) Is the polynomial of Exercise 2(b) in S ? I.e., can it be obtained by addition and multiplication (but no subtraction) from x, y and squares of suitable polynomials?

   (b) More generally, is

   $$\frac{x_1^p + \cdots + x_n^p}{n} - \left(\frac{x_1 + \cdots + x_n}{n}\right)^p$$

   in S ? (Note that p is a natural number.)

   Solution. (a) Yes, it is $(3/8)(x+y)(x-y)^2$.

   (b) Yes. Hint: use induction on p.

4. Is the polynomial
$$\left(\frac{x+y+z}{3}\right)^3 - xyz$$
in S ?

   Answer. Yes.

5. Suppose that the polynomial $f(x_1, \ldots, x_n)$ is non-negative for all non-negative x_1, x_2, $\ldots$, x_n (as in all examples above). Does it follow that $f \in S$?

   Solution. No. Let $n = 2$ and
$$f(x, y) = 1 - 3xy + xy^2 + x^2y.$$
   It follows from the inequality between the arithmetic and the geometric mean that $f \geq 0$ whenever x and y are nonnegative. However, $f \notin S$. We leave the proof to the reader but, as a hint, we define the main tool. The *Newton polygon* of a polynomial $g(x, y) = \sum a_{ij}x^iy^j$ is the convex hull af the lattice points $(i, j) \in \boldsymbol{Z}^2_{\geq 0}$ such that $a_{ij} \neq 0$. For example, the Newton polygon of f is the triangle with vertices $(0, 0)$, $(1, 2)$ and $(2, 1)$.

By Lemma 6.1.1, this means that the polynomial
$$f(x^2, y^2) = 1 - 3x^2y^2 + x^2y^4 + x^4y^2$$
is not a sum of squares, although it is nonnegative everywhere on $\mathbb{R}^2$. This is Motzkin's example.

6.2 Positive semi-definite matrices

There is another type of algebraic inequality that I will also discuss in the minicourse: determinantal and permanental inequalities for positive semidefinite matrices.

Recall that the determinant and the permanent of an $m \times m$ matrix $C = (c_{i,j})$ are defined by
$$\det C = \sum_{\pi \in \mathfrak{S}_m} (-1)^\pi \prod_{i=1}^m c_{i,\pi(i)}, \qquad \operatorname{per} C = \sum_{\pi \in \mathfrak{S}_m} \prod_{i=1}^m c_{i,\pi(i)},$$
where $\mathfrak{S}_m$ is the symmetric group on m elements. Throughout this section, we assume that C is a positive semi-definite Hermitian $m \times m$ matrix (we write $C \geq 0$). For such C, Hadamard proved that
$$\det C \leq \prod_{i=1}^m c_{i,i}, \tag{6.2.1}$$
with equality if and only if C has a zero row or is a diagonal matrix. Fischer generalized this to
$$\det C \leq \det B' \cdot \det B'' \tag{6.2.2}$$
for
$$C = \begin{pmatrix} B' & A \\ A^* & B'' \end{pmatrix} \geq 0, \tag{6.2.3}$$

with equality if and only if $\det B' \cdot \det B'' \cdot A = 0$.

Concerning the permanent of a positive semi-definite matrix, Marcus [8, 9] proved that

$$\operatorname{per} C \geq \prod_{i=1}^{m} c_{i,i}, \tag{6.2.4}$$

with equality if and only if C has a zero row or is a diagonal matrix. Lieb [7] generalized this to

$$\operatorname{per} C \geq \operatorname{per} B' \cdot \operatorname{per} B'' \tag{6.2.5}$$

for C as in (6.2.3), with equality if and only if C has a zero row or $A = 0$. Moreover, he proved that in the polynomial $P(\lambda)$ of degree n (=size of B') defined by

$$P(\lambda) = \operatorname{per} \begin{pmatrix} B' & A \\ \lambda A^* & B'' \end{pmatrix} = \sum_{t=0}^{n} c_t \lambda^t,$$

all coefficients c_t are real and non-negative. This is indeed a stronger theorem since it implies

$$\operatorname{per} C = P(1) = \sum_{t=0}^{n} c_t \geq c_0 = \operatorname{per} B' \cdot \operatorname{per} B''.$$

If $m = 2n$, then the inequalities $c_t \geq 0$ even imply

$$\operatorname{per} C \geq \operatorname{per} B' \cdot \operatorname{per} B'' + |\operatorname{per} A|^2, \tag{6.2.6}$$

since the right hand side is $c_0 + c_n$. Inequality (6.2.6) is case $p = 2$ of the following conjecture of Marcus: If C is a positive semi-definite Hermitian $pn \times pn$ matrix partitioned into $p \times p$ blocks $A_{i,j}$, each of size $n \times n$, then

$$\operatorname{per} C \geq \operatorname{per}((\operatorname{per} A_{i,j})_{i,j}).$$

This is itself a special case of the so-called permanental dominance conjecture, which we do not state here.

Đoković [3,11] gave a simple proof of Lieb's above inequalities, and showed also that if B' and B'' are positive definite then $c_t = 0$ if and only if all sub-permanents of A of order t vanish. Lieb [7] also states an analogous (and analogously provable) theorem for determinants: for C as in (6.2.3), let

$$D(\lambda) = \det \begin{pmatrix} B' & A \\ -\lambda A^* & B'' \end{pmatrix} = \sum_{t=0}^{n} d_t \lambda^t. \tag{6.2.7}$$

If $\det B' \cdot \det B'' = 0$, then $D(\lambda) = 0$. If B' and B'' are positive definite, then d_t is positive for $t \leq \operatorname{rk} A$ and is zero for $t > \operatorname{rk} A$. In contrast to the above deduction of the permanental Lieb inequality (6.2.5) from $c_t \geq 0$, there is no obvious way of deducing the Fischer inequality (6.2.2) from $d_t \geq 0$. Instead of (6.2.2), we get

$$D(1) = \det \begin{pmatrix} B' & A \\ -A^* & B'' \end{pmatrix} \geq \det B' \cdot \det B''. \tag{6.2.8}$$

Remark 6.2.1. In all of Lieb's inequalities mentioned above, the condition that the matrix C is positive semi-definite can be replaced by the weaker condition that the diagonal blocks B' and B'' are positive semi-definite. The proof goes through virtually unchanged. Alternatively, this stronger form of the inequalities can be easily deduced from the seemingly weaker form above.

In the minicourse, I will explain proofs for some of these inequalities. All proofs will be based on representing the relevant positive polynomial as a sum of squares.

6.3 References

[1] R.B. Bapat, Recent developments and open problems in the theory of permanents, Math. Student 76 (2007), 55–69.

[2] A. Caicedo's teaching blog, `http://caicedoteaching.wordpress.com/2008/11/11/275-positive-polynomials/`

[3] D. Ž. Đoković, Simple proof of a theorem on permanents, Glasgow Math. J. 10 (1969), 52–54.

[4] R. Fujisawa, Algebraic means, Proc. Imp. Acad. Volume 1, Number 5 (1918), 159–170.

[5] P.E. Frenkel, P. Horváth, Minkowski's inequality and sums of squares, `http://arxiv.org/abs/1206.5783`

[6] Hurwitz, Über den Vergleich des arithmetischen und des geometrischen Mittels, in: Math. Werke, 505–507, Basel, E. Berkhäuser, 1933.

[7] E. H. Lieb, Proofs of some conjectures on permanents, J. Math. Mech. 16 (1966), 127–134.

[8] M. Marcus, The permanent analogue of the Hadamard determinant theorem, Bull. Amer. Math. Soc. 69 (1963), 494–496.

[9] M. Marcus, The Hadamard theorem for permanents, Proc. Amer. Math. Soc. 15 (1964), 967–973.

[10] M. Marcus, M. Newman, The permanent function as an inner product, Bull. Amer. Math. Soc. 67 (1961), 223–224.

[11] Henryk Minc, Permanents. Cambridge University Press, 1984.

[12] Walter Rudin, Sums of Squares of Polynomials, The American Mathematical Monthly 107/9 (2000), 813–821;

[13] J. Michael Steele, The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities, Cambridge University Press, 2004.

Chapter 7

Tibor Jordán: Location and localization problems in networks

7.1 Introduction

In the network localization problem the locations of some nodes (called anchors) of a network as well as the distances between some pairs of nodes are known, and the goal is to determine the location of all nodes. This is one of the fundamental algorithmic problems in the theory of wireless sensor networks, see for example [1].

A natural additional question is whether a solution to the localization problem is unique. The network, with the given locations and distances, is said to be *uniquely localizable* if there is a unique set of locations consistent with the given data. The unique localizability of a two-dimensional network, whose nodes are 'in generic position', can be characterized by using results from graph rigidity theory. In this case unique localizability depends only on the combinatorial properties of the network and can be tested by efficient algorithms.

The goal of this series of lectures is to explore the combinatorial background of this characterization and the corresponding algorithms. After proving some of the classical results of combinatorial rigidity theory and discussing the necessary algorithmic tools, we shall investigate several versions and extension of the network localization problems and their solutions.

7.1.1 Basic definitions

In what follows we shall summarize the basic concepts and some of the key preliminary results. See the Appendix for more definitions concerning graphs and matroids.

As we shall see, unique localizability (in the 'generic case') is determined completely by the *distance graph* of the network and the set of anchors, or equivalently, by the *grounded graph* of the network and the number of anchors. The vertices of the distance and grounded graph correspond to the nodes of the network. In both graphs two vertices are connected by an edge if the corresponding distance is explicitly known. In the grounded graph we have additional edges: all pairs of vertices corresponding to anchor nodes are adjacent. The grounded graph represents all known distances, since the distance between two anchors can be obtained from their locations. Before stating the basic observation about unique localizability

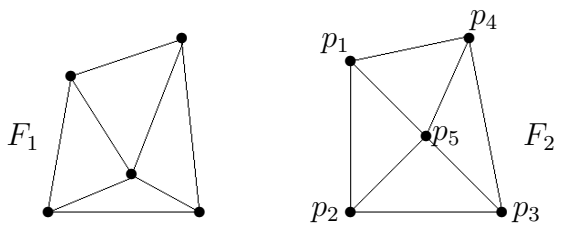


Figure 7.1.1: Two realizations of the same graph G in $\mathbb{R}^2$: F_1 is globally rigid; F_2 is not since we can obtain a realization of G which is equivalent but not congruent to F_2 by reflecting p_2 in the line through p_1, p_5, p_3.

we need some additional terminology. It is convenient to investigate localization problems with distance information by using frameworks, the central objects of rigidity theory.

A d-dimensional *framework* (also called *geometric graph* or *formation*) is a pair (G, p), where $G = (V, E)$ is a graph and p is a map from V to $\mathbb{R}^d$. We consider the framework to be a straight line *realization* of G in $\mathbb{R}^d$. Two frameworks (G, p) and (G, q) are *equivalent* if corresponding edges have the same lengths, that is, if $||p(u) - p(v)|| = ||q(u) - q(v)||$ holds for all pairs u, v with $uv \in E$, where $||.||$ denotes the Euclidean norm in $\mathbb{R}^d$. Frameworks (G, p), (G, q) are *congruent* if $||p(u) - p(v)|| = ||q(u) - q(v)||$ holds for all pairs u, v with $u, v \in V$. This is the same as saying that (G, q) can be obtained from (G, p) by an isometry of $\mathbb{R}^d$. We shall say that (G, p) is *globally rigid*, or that (G, p) is a *unique realization* of G, if every framework which is equivalent to (G, p) is congruent to (G, p), see Figure 7.1.1.

The next observation shows that the theory of globally rigid frameworks is the mathematical background which is needed to investigate the unique localizability of networks.

Theorem 7.1.1. *Let N be a network in $\mathbb{R}^d$ consisting of m anchors located at positions $p_1, ..., p_m$ and $n - m$ ordinary nodes located at $p_{m+1}, ..., p_n$. Suppose that there are at least $d + 1$ anchors in general position. Let G be the grounded graph of N and let $p = (p_1, ..., p_n)$. Then the network is uniquely localizable if and only if (G, p) is globally rigid.*

7.1.2 Generic frameworks

It is a hard problem to decide if a given framework is globally rigid. Indeed Saxe [8] has shown that this problem is NP-hard even for 1-dimensional frameworks. The problem becomes more tractable, however, if we assume that there are no algebraic dependencies between the coordinates of the points of the framework.

A framework (G, p) is said to be *generic* if the set containing the coordinates of all its points is algebraically independent over the rationals. (Recall that a set $\{\alpha_1, \alpha_2, \ldots, \alpha_t\}$ of real numbers is *algebraically independent* over the rationals if, for all non-zero polynomials with rational coefficients $p(x_1, x_2, \ldots, x_t)$, we have

$p(\alpha_1, \alpha_2, \ldots, \alpha_t) \neq 0$.) Restricting to generic frameworks gives us two important 'stability properties'. The first is that, if (G, p) is a globally rigid d-dimensional generic framework then there exists an $\epsilon > 0$ such that all frameworks (G, q) which satisfy $||p(v) - q(v)|| < \epsilon$ for all $v \in V$ are also globally rigid. The second, which follows from a recent result of Gortler at al. [4], is that if some d-dimensional generic realization of a graph G is globally rigid, then all d-dimensional generic realizations of G are globally rigid.

7.2 Rigidity and global rigidity of graphs

Rigidity, which is a weaker property of frameworks than global rigidity, plays an important role in the exploration of the structural results of global rigidity as well as in the corresponding algorithmic problems. Intuitively, we can think of a d-dimensional framework (G, p) as a collection of bars and joints where vertices correspond to joints and each edge to a rigid bar joining its end-points. The framework is rigid if it has no continuous deformations. Equivalently, and more formally, a framework (G, p) is *rigid* if there exists an $\epsilon > 0$ such that, if (G, q) is equivalent to (G, p) and $||p(u) - q(u)|| < \epsilon$ for all $v \in V$, then (G, q) is congruent to (G, p).

Rigidity, like global rigidity, is a generic property of frameworks, that is, the rigidity of a generic realization of a graph G depends only on the graph G and not the particular realization. We say that the graph G is *rigid*, respectively *globally rigid* or *uniquely realizable*, in $\mathbb{R}^d$ if every (or equivalently, if some) generic realization of G in $\mathbb{R}^d$ is rigid, respectively globally rigid.

The problem of characterizing when a graph is rigid in $\mathbb{R}^d$ has been solved for $d = 1, 2$. We refer the reader to [5, 10, 11] for a detailed survey of the rigidity of d-dimensional frameworks. A similar situation holds for global rigidity: the problem of characterizing when a generic framework is globally rigid in $\mathbb{R}^d$ has also been solved for $d = 1, 2$.

We shall state these characterizations and study their algorithmic implications. Here we only mention a general necessary condition, due to Hendrickson, which is valid in all dimensions. We say that G is *redundantly rigid* in $\mathbb{R}^d$ if $G - e$ is rigid in $\mathbb{R}^d$ for all edges e of G.

Theorem 7.2.1. *[6] Let (G, p) be a generic framework in $\mathbb{R}^d$. If (G, p) is globally rigid then either G is a complete graph with at most $d+1$ vertices, or G is $(d+1)$-connected and redundantly rigid in $\mathbb{R}^d$.*

7.3 Rigidity matrices and matroids

A matroid is an abstract structure which extends the notion of linear independence of vectors in a vector space. We will see that many of the rigidity properties of a generic framework (G, p) are determined by an associated matroid defined on the edge set of G. (See the Appendix for the basic definitions and [7, 9] for more information on matroids.)

Let (G, p) be a d-dimensional realization of a graph $G = (V, E)$. The *rigidity matrix* of the framework (G, p) is the matrix $R(G, p)$ of size $|E| \times d|V|$, where, for each edge $e = v_i v_j \in E$, in the row corresponding to e, the entries in the two

columns corresponding to vertices i and j contain the d coordinates of $(p(v_i)-p(v_j))$ and $(p(v_j)-p(v_i))$, respectively, and the remaining entries are zeros. See [5,10] for more details. The rigidity matrix of (G,p) defines the *rigidity matroid* of (G,p) on the ground set E where a set of edges $F \subseteq E$ is *independent* if and only if the rows of the rigidity matrix indexed by F are linearly independent. Any two generic d-dimensional frameworks (G,p) and (G,q) have the same rigidity matroid. We call this the *d-dimensional rigidity matroid* $\mathcal{R}_d(G)$ of the graph G. We denote the rank of $\mathcal{R}_d(G)$ by $r_d(G)$.

As an example, consider a 1-dimensional framework (G,p). In this case, the rows of $R(G,p)$ are just scalar multiples of a directed incidence matrix of G. It is well known that a set of rows in this matrix is independent if and only if the corresponding edges induce a forest in G. Thus $\mathcal{R}_1(G)$ is the cycle matroid of G.

Gluck characterized rigid graphs in terms of their rank.

Theorem 7.3.1. *[3] Let $G=(V,E)$ be a graph. Then G is rigid in $\mathbb{R}^d$ if and only if either $|V| \leq d+1$ and G is complete, or $|V| \geq d+2$ and $r_d(G) = d|V| - \binom{d+1}{2}$.*

This characterization does not give rise to a polynomial algorithm for deciding whether a graph is rigid in $\mathbb{R}^d$. The problem is that to compute $r_d(G)$ we need to determine the rank of the rigidity matrix of a generic realization of G in $\mathbb{R}^d$. There is no known polynomial algorithm for calculating the rank of a matrix in which the entries are linear functions of algebraically independent numbers.

We say that a graph $G=(V,E)$ is *M-independent* in $\mathbb{R}^d$ if E is independent in $\mathcal{R}_d(G)$. Knowing when subgraphs of G are M-independent allows us to determine the rank of G (and hence determine whether G is rigid), since we can construct a base for $\mathcal{R}_d(G)$ by greedily constructing a maximal independent set of $\mathcal{R}_d(G)$. This follows from axiom (M3) which guarantees that an independent set which is maximal with respect to inclusion is also an independent set of maximum cardinality. For example, when $d=1$, we have seen that a subgraph is independent if and only if it is a forest. Thus we can determine the rank of G by greedily growing a maximal forest F in G. By Theorem 7.3.1, G is rigid if and only if F has $|V|-1$ edges, i.e. F is a spanning tree of G.

7.4 Warm up exercises

The following exercises may help warm up for these lectures.

Exercise 7.4.1. Show that a framework (G,p) is rigid in $\mathbb{R}^1$ if and only if G is connected.

Exercise 7.4.2. Characterize the redundantly rigid graphs in $\mathbb{R}^1$ and develop an efficient algorithm for testing whether a given graph has this property.

Exercise 7.4.3. Construct two-dimensional frameworks (G,p) on n vertices for all $n \geq 2$ which are rigid and have $2n-3$ edges. Define a family of graphs which contains a rigid graph in $\mathbb{R}^2$ on n vertices and with $2n-3$ edges for all $n \geq 2$.

Exercise 7.4.4. Construct two-dimensional frameworks (G,p) on n vertices for all $n \geq 4$ which are rigid and have $2n-4$ edges. Can you do that so that the framework is in generic (or general) position?

Exercise 7.4.5. Construct globally rigid graphs in $\mathbb{R}^2$ on n vertices for all $n \geq 2$. Try to do it so that the number of edges is as small as possible.

7.5 Appendix

In what follows we introduce the basic graph (and matroid) theoretical notions. For more details see for example [2].

A *graph* $G = (V, E)$ consists of two sets V and E. The elements of V are called *vertices* (or *nodes*). The elements of E are called *edges*. Each edge $e \in E$ joins two vertices from V, which are called the *endvertices* of e. The notations $V(G)$ and $E(G)$ are also used for the vertex- and edge-sets of a graph G. If vertex v is an endvertex of edge e then v is said to be *incident* with e and e is incident with v. A vertex v is *adjacent* to vertex u if they are joined by an edge. A graph is *simple* if the pairs of endvertices of its edges are pairwise distinct.

The *degree* of a vertex v in a graph G, denoted by $d_G(v)$, is the number of edges incident with v. A graph is *regular* if every vertex is of the same degree. It is *k-regular* if every vertex is of degree k.

A *path* in a graph G from vertex u to vertex v is an alternating sequence of vertices and edges, which starts and ends with u and v (which are its initial and final vertices, respectively), and for which consecutive elements are incident with each other and no internal vertex is repeated. A *cycle* is a path which contains at least one edge and for which the initial vertex is also the final vertex. A graph is *connected* if between every pair of vertices there is a path.

A *subgraph* of a graph G is a graph H with $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$. In a graph G the *induced subgraph* on a set X of vertices, denoted by $G[X]$, has X as its vertex set and it contains every edge of G whose endvertices are in X. A subgraph H is a *spanning subgraph* if $V(H) = V(G)$. A *component* of a graph G is a maximal connected subgraph. A *k-factor* of a graph G is a k-regular spanning subgraph.

The operation of *deleting a vertex set* $X \subseteq V(G)$ from a graph G removes the vertices in X from $V(G)$ and also removes every edge which has an endvertex in X from $E(G)$. The resulting graph is denoted by $G - X$ (or $G - x$, if $X = \{x\}$ is a single vertex). The operation of *deleting an edge set* $F \subseteq E(G)$ from a graph G removes the edges in F from $E(G)$. The resulting graph is denoted by $G - F$ (or $G - f$, if $F = \{f\}$ is a single edge).

A *forest* is a graph without cycles and a *tree* is a connected forest. A *spanning tree* of a graph G is a spanning subgraph which is a tree.

A graph is a *complete graph* if each pair of its vertices is joined by an edge. A complete graph on n vertices is denoted by K_n. A graph is *bipartite* if its vertices can be partitioned into two sets in such a way that no edge joins two vertices in the same set. A *complete bipartite graph* is a bipartite graph in which each vertex in one partite set is adjacent to all vertices in the other partite set. If the two partite sets have cardinalitites m and n, then this graph is denoted by $K_{m,n}$. A graph G on n vertices is a *wheel*, denoted by W_n, if it has an induced subgraph which is a cycle on $n-1$ vertices and the remaining vertex is joined to all vertices of this cycle.

A *k-vertex-cut* in a graph G is a set $X \subseteq V(G)$ of k vertices for which $G - X$ is not connected. A *k-edge-cut* is a set $F \subseteq E(G)$ of k edges for which $G - F$ is not connected. A graph is called *k-vertex-connected* (or *k-connected*) if it has at least $k + 1$ vertices and contains no l-vertex-cut for $l \leq k - 1$. A graph is *k-edge-connected* if it contains no l-edge-cuts for $l \leq k - 1$.

Two paths are called *openly disjoint* if they have no common internal vertex.

They are called *edge disjoint* if they have no common edge. A fundamental theorem of Menger states that if u and v are non-adjacent vertices in graph G then the smallest integer k for which there is a k-vertex-cut X in G such that u and v are in different components of $G - X$ is equal to the maximum number of pairwise openly disjoint paths from u to v. The edge disjoint version of Menger's theorem is as follows. For any pair of vertices u, v in G the smallest integer k for which there is a k-edge-cut F in G such that u and v are in different components of $G - F$ is equal to the maximum number of pairwise edge disjoint paths from u to v.

An *isomorphism* between two graphs G and H is a vertex bijection $\phi : V(G) \to V(H)$ such that $uv \in E(G)$ if and only if $\phi(u)\phi(v) \in E(H)$. A *graph automorphism* is an isomorphism of the graph to itself. The *orbit of a vertex* u of a graph G is the set of all vertices $v \in V(G)$ such that there is an automorphism ϕ such that $\phi(u) = v$. A graph is *vertex-transitive* if all the vertices are in the same orbit.

The *incidence matrix* of a graph $G = (V, E)$ is an $|E| \times |V|$ matrix I where the entry in the row of edge e and vertex v is equal to 1 if e is incident with v, and 0 otherwise. The *directed incidence matrix* of G is obtained from I by replacing exactly one of the two 1's in each row of I by -1.

7.5.1 Matroids

A *matroid* is an ordered pair $\mathcal{M} = (E, \mathcal{I})$ where E is a finite set, and $\mathcal{I}$ is a family of subsets of E, called *independent sets*, which satisfy the following three axioms.

(M1) $\emptyset \in \mathcal{I}$,
(M2) if $I \in \mathcal{I}$ and $D \subseteq I$ then $D \in \mathcal{I}$,
(M3) for all $F \subseteq E$, the maximal independent subsets of F have the same cardinality.

The fundamental example of a matroid is obtained by taking E to be a set of vectors in a vector space and $\mathcal{I}$ to be the family of all linearly independent subsets of E.

Given a matroid $\mathcal{M} = (E, \mathcal{I})$, the cardinality of a maximum independent subset of a set $F \subseteq E$ is defined to be the *rank* of F and denoted by $r(F)$. The rank of E is referred to as the *rank of* $\mathcal{M}$. A *base* of $\mathcal{M}$ is a maximum independent subset of E. A subset of E which is not independent is said to be *dependent*. A *circuit* of $\mathcal{M}$ is a minimal dependent subset of E. The matroid $\mathcal{M}$ is said to be *connected* if every pair of elements of E are contained in a circuit.

Given a graph $G = (V, E)$, we may define a matroid $\mathcal{M} = (E, \mathcal{I})$ by letting $\mathcal{I}$ be the family of all edge sets of forests in G. The rank of a set $F \subseteq E$ is given by $r(F) = |V| - k(F)$, where $k(F)$ denotes the number of connected components in the graph (V, F). A base of $\mathcal{M}$ is the edge set of a forest which has the same number of components as G. A circuit of $\mathcal{M}$ is the edge set of a cycle of G, and $\mathcal{M}$ is connected if and only if G is 2-connected. This matroid is called the *cycle matroid of* G.

7.6 References

[1] J. Aspnes, T. Eren, D.K. Goldenberg, A.S. Morse, W. Whiteley,

Y.R. Yang, B.D.O. Anderson, P.N. Belhumeur, A theory of network localization, *IEEE Trans. on Mobile Computing*, vol. 5, issue 12, pp. 1663-1678, 2006.

[2] J.A. Bondy, U.S.R. Murty, *Graph theory*, Springer, 2008.

[3] H. Gluck, Almost all simply connected closed surfaces are rigid, *Geometric topology (Proc. Conf., Park City, Utah, 1974)*, pp. 225–239. Lecture Notes in Math., Vol. 438, Springer, Berlin, 1975.

[4] S.J. Gortler, A.D. Healy, and D.P. Thurston, Characterizing generic global rigidity, 2007, arXiv:0710.0926v3.

[5] J. Graver, B. Servatius, and H. Servatius, Combinatorial Rigidity, AMS Graduate Studies in Mathematics Vol. 2, 1993.

[6] B. Hendrickson, Conditions for unique graph realizations, *SIAM J. Comput.* **21** (1992), no. 1, 65-84.

[7] A. Recski, *Matroid theory and its applications in electric network theory and in statics*, Akadémiai Kiadó, Budapest, 1989.

[8] J.B. Saxe, Embeddability of weighted graphs in k-space is strongly NP-hard, Tech. Report, Computer Science Department, Carnegie-Mellon University, Pittsburgh, PA, 1979.

[9] A. Schrijver, *Combinatorial Optimization*, Springer, Berlin, 2003.

[10] W. Whiteley, Some matroids from discrete applied geometry. Matroid theory (Seattle, WA, 1995), 171–311, Contemp. Math., 197, Amer. Math. Soc., Providence, RI, 1996.

[11] W. Whiteley, Rigidity and scene analysis, in: Handbook of Discrete and Computational Geometry (J. E. Goodman and J. O'Rourke, eds.), CRC Press, Second Edition, pp. 1327-1354, 2004.

Chapter 8

Márton Naszódi: A Glimpse of Discrete Geometry

On the lectures at the ELTE Summer School in Mathematics 2013 we will discuss various topics in Geometry that are accessible at BSc. level and yet, lead to contemporary research. The main goal in the selection of the problems is to present a diverse set of methods and thus invite you to a field where linear algebra, combinatorics, probability and analysis all come together.

8.1 Borsuk's partitioning problem

The diameter of a convex set K in $\mathbb{R}^d$ is the supremum of distances between points of K. In 1932 Borsuk [4] asked whether any convex body can be partitioned into $d+1$ pieces of smaller diameter – he did not actually conjectured it to be true, but it still became to be known as Borsuk's Conjecture. The answer is affirmative for the Euclidean ball (of any dimension) as well as for any planar convex body. As we will see in section 8.2, it is true for any smooth convex body (that is a convex body with a unique supporting hyperplane at each boundary point).

Theorem 8.1.1. *If K is a convex body that is symmetric about a point then K has a Borsuk partition of cardinality $d+1$.*

We may define the quantity $b(d) := \max\{b(K) : K \subset \mathbb{R}^d \text{ a convex body}\}$, where $b(K)$ is the minimal cardinality of a family of sets of diameter less than that of K that cover K.

In 1993, Kahn a Kalai [9] surprised the mathematical community by showing that certain sets in $\mathbb{R}^d$ require at least $1.2^{\sqrt{d}}$ parts (if d is large), that is $b(d) \geq 1.2^{\sqrt{d}}$. On the other hand, Schramm (see later) showed that, roughly, $b(d) \leq 1.23^d$. In conclusion, we know that $b(d)$ is far greater than $d+1$, but we still do not understand its order of magnitude.

8.2 Covering by translates of a convex body

Say you have a penny that you want to cover with other pennies, but the obvious solution of putting one right on top of the other is not allowed. How many do you need? Next, you have a square that you want to cover with slightly smaller

squares whose sides are parallel to the sides of the bigger square. How many do you need now? How about hexagons, etc.? How many pennies do you need to cover a one dollar bill? These questions yield the following

Definition 8.2.1. *Let K and L be convex set in $\mathbb{R}^d$ with non-empty interior. The* covering number $N(K,L)$ *of K by L is the minimum number of translates of L that cover K.*

Exercise 8.2.2. *Prove that $N(K,M) \leq N(K,L)N(L,M)$ for any three convex bodies K, L and M.*

The main topic of this lecture is estimating $N(K, \mathrm{int}K)$, where $\mathrm{int}K$ denotes the (topological) interior of the set K. This question turns out to be the same as covering K by slightly smaller translates of itself.

Exercise 8.2.3. *Prove that $N(K, \mathrm{int}K) = \min\{N(K, \lambda K) : \lambda < 1\}$ for any convex body K.*

Exercise 8.2.4. *Prove that this quantity is affine-invariant, that is $N(K, \mathrm{int}K) = N(\tilde{K}, \mathrm{int}\tilde{K})$ for any convex body K and its non-degenerate affine image $\tilde{K}$.*

Gohberg and Markus [6] conjectured that 2^d translates suffice.

Conjecture 8.2.5. *For any convex body $K \subset \mathbb{R}^d$ we have $N(K, \mathrm{int}K) \leq 2^d$ and equality holds only if K is a parallelotope.*

Exercise 8.2.6. *Prove the conjecture on the plane, that is for $d = 2$.*

Boltyanski [3] and Hadwiger [7,8] raised the following — rather different looking — problem: We say that a direction $u \in \mathbb{R}^d$(with $|u| = 1$) *illuminates* a boundary point $b \in \mathrm{bd}K$ of K if the ray emanating from b with direction u (that is $\{b + \lambda u : \lambda > 0\}$) intersects the interior of K. The *illumination number* of K is defined as the minimum number of directions that illuminate the whole boundary of K.

Exercise 8.2.7. *We could define the following version of illumination: a point $p \in \mathbb{R}^d$ (think of it as a light source)* illuminates *a boundary point $b \in \mathrm{bd}K$ of K if the ray emanating from b with direction $\overrightarrow{pb}$ (that is $\{b + \lambda(b-p) : \lambda > 0\}$) intersects the interior of K. Prove that the illumination number of K is equal to the minimum number of points that illuminate the whole boundary of K.*

Finally, it is not difficult to show that the two problems are equivalent:

Exercise 8.2.8. *Show that the illumination number of K is equal to $N(K, \mathrm{int}K)$.*

8.2.1 Results

Several special cases of the Illumination Conjecture have been settled. For example, if K is a smooth convex body (that is it has a unique supporting hyperplane at each point of it boundary) than the illumination number is far from maximal, it is $d + 1$. We will prove this using the Gauss-map on the boundary of a convex body. This will explain why the Illumination Conjecture can be thought of as an "integration problem": find a necessary condition for a partition of the unit sphere to be the Gaussian image (ie. the "derivative") of a convex body.

Exercise 8.2.9. *Show that the illumination number of the euclidean ball is $d+1$.*

Exercise 8.2.10. *Show that the illumination number of the cube* $[-1,1]^d$ *is* 2^d.

Definition 8.2.11. *A convex body* $W \subset \mathbb{R}^d$ *is a* set of constant width *one if the distance of any pair of parallel supporting hyperplanes of* W *is one.*

Exercise 8.2.12. *Give an example of a planar set of constant width one. Now, give and example that is not the circle.*

Schramm [15] proved the Illumiation Conjecture for sets of constant width.

We also know that three-dimensional sets that are symmetric about a point can be illuminated by 8 light directions [10] [2]. However, in general the best upper bound (due to Rogers [12]) is over half a century old.

Theorem 8.2.13.

$$N(K, \mathrm{int}K) \leq \frac{\mathrm{vol}(K-K)}{\mathrm{vol}(K)}(d\ln d + d\ln\ln d + 5d) \leq \tag{8.2.1}$$

$$\begin{cases} 2^d(d\ln d + d\ln\ln d + 5d) & \text{if } K = -K, \\ \binom{2d}{d}(d\ln d + d\ln\ln d + 5d) & \text{otherwise.} \end{cases}$$

To prove this result, we will first cover the whole space, $\mathbb{R}^d$ by (infinitely many) translates of $\mathrm{int}K$ in an economical way, and then show that this covering yields a covering of K.

Definition 8.2.14. *For a family* $\mathcal{K}$ *of translates of a convex body* $K \subset \mathbb{R}^d$ *within the cube* $C_s = [-s,s]^d$ *we define*

$$\rho_-(K, C_s) := \frac{1}{(2s)^d} \sum_{K \in \mathcal{K}, K \subset C_s} \mathrm{vol}K.$$

Next, the lower density *of* $\mathcal{K}$ *is defined as*

$$\rho_-(K) := \liminf_{s\to\infty} \rho_-(K, C_s)$$

Finally, the covering density *of* K *is*

$$\theta(K) := \inf_{\mathcal{K}} \rho_-(\mathcal{K})$$

where the infimum is taken over all families $\mathcal{K}$ *of translates of* K.

Theorem 8.2.15 (Rogers, [12])**.** *For any convex body* $K \subset \mathbb{R}^d$*, we have*

$$\theta(K) \leq d\log d + d\log\log d + 5d.$$

We will use this result to prove Theorem 8.2.13.

8.3 Rigidity of Polyhedra

A three dimensional convex polyhedron is the convex hull of finitely many points in $\mathbb{R}^3$ which do not lie in a plane. What parameters determine a convex polyhedron? One of Cauchy's most famous and elegant results is his proof of the fact that knowing each face and how they are connected to one another enables one to reconstruct the polyhedron. To make this statement more precise, we will first visit the basics of the theory of convex polyhedra. This will be beneficial for those too, who lack a strong interest in geometry, since these objects appear in many areas of mathematics.

Definition 8.3.1. *The convex hull of a finite set in $\mathbb{R}^d$ is called a* convex polytope.

Definition 8.3.2. *The intersection of finitely many half spaces in $\mathbb{R}^d$ is a* polyhedral set.

Exercise 8.3.3. *Are these two notions the same? Give an example showing that they are not.*

As a somewhat confusing (but historically firmly agreed) terminology is to call three dimensional convex polytopes *convex polyhedra.*

Let $P \subset \mathbb{R}^d$ be a convex polytope and H be a hyperplane that supports P, that is, H intersects P but not the interior of P. Then we call $F := H \cap P$ a face of P. Since F is a convex subset of H, its dimension is defined in the natural way. Moreover, through each boundary point of P there is a supporting hyperplane of P, and thus, the boundary of P is covered by its faces. It is not difficult to see that the intersection of two faces is a face again and that P has finitely many faces. To simplify some arguments, P itself is also a face of P, as well as the empty set (whose dimension is agreed to be -1).

Theorem 8.3.4. *The set of faces of a convex polytope form a lattice with respect to containment.*

We call this lattice the *face lattice* of P. This lattice is the combinatorial structure of P. Next, we introduce its metric structure.

Two sets in $\mathbb{R}^d$ are *congruent* if there is an isometry $\phi : \mathbb{R}^d \to \mathbb{R}^d$ that maps one set onto the other.

Theorem 8.3.5 (Cauchy [5])**.** *Let P and Q be (three dimensional) convex polyhedra with isomorphic face lattices and assume that the corresponding edges and faces are congruent. Then P and Q are congruent.*

Exercise 8.3.6. *Show that the assumption of convexity cannot be dropped, that is, construct two polyhedra with isomorphic face lattices and with congruent corresponding edges and faces that are not congruent.*

We will discuss Cauchy's ingenious proof that combines a "global" combinatorial and a "local" geometric argument.

One way of phrasing the theorem is to say that convex polyhedra are *globally rigid*. A local version of rigidity can also be introduced: a polyhedron P is *locally rigid*, if there is an $\varepsilon > 0$ such that for any polyhedron Q whose face lattice is isomorphic to that of P and whose corresponding edges and faces are congruent to those of P the following holds: if the vertices of Q are at distance at most ε from the corresponding vertex of P then Q is congruent to P. If time permits, we will discuss a few results on local rigidity, too.

Cauchy's result can be considered as the starting point of a theory that offers a number of open questions. One of them is

Stoker's Conjecture: *The dihedral angles (that is, the angles of faces connected by an edge) and the face lattice of a polytope all of whose faces are triangles determine the face angles.*

Exercise 8.3.7. *Show that the dihedral angles and the face lattice do not determine a polytope up to similarity.*

Exercise 8.3.8. *Show that the face angles and the face lattice determine the dihedral angles of a polytope all of whose vertices are adjacent to three edges.*

8.4 References

[1] Martin Aigner and Günter M. Ziegler. *Proofs from The Book*. Springer-Verlag, Berlin, fourth edition, 2010.

[2] Károly Bezdek. The problem of illumination of the boundary of a convex body by affine subspaces. *Mathematika*, 38(2):362–375 (1992), 1991.

[3] V. G. Boltjanskii. The problem of the illumination the boundary a convex body (in russian). *Izv. Mold. Fil. Akad. Nauk SSSR*, 10(76):79–86, 1960.

[4] Karol Borsuk. Über Schnitte der n-dimensionalen Euklidischen Räume. *Math. Ann.*, 106(1):239–248, 1932.

[5] A. Cauchy. Sur les polygones et les polyèdres, seconde mémoire. *J. École Polytechnique XVIe Cahier, Tome IX, Œuvres Complètes, IIe Série, Vol. 1, Paris 1905*, pages 87–98, 1813.

[6] I. Gohberg and A. Markus. A problem on covering of convex figures by similar figures. *Izv. Mold. Fil. Akad. Nauk. SSSR*, 10:87—-90, 1960.

[7] H. Hadwiger. Ungelösteprobleme nr. 20. *Elem. der Math.*, 12:121, 1957.

[8] H. Hadwiger. Ungelösteprobleme nr. 38. *Elem. der Math.*, 15:130–131, 1960.

[9] Jeff Kahn and Gil Kalai. A counterexample to Borsuk's conjecture. *Bull. Amer. Math. Soc. (N.S.)*, 29(1):60–62, 1993.

[10] Marek Lassak. Solution of Hadwiger's covering problem for centrally symmetric convex bodies in E^3. *J. London Math. Soc. (2)*, 30(3):501–511, 1984.

[11] Jiří Matoušek. *Lectures on discrete geometry*, volume 212 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2002.

[12] C. A. Rogers. A note on coverings. *Mathematika*, 4:1–6, 1957.

[13] C. A. Rogers and G. C. Shephard. The difference body of a convex body. *Arch. Math. (Basel)*, 8:220–233, 1957.

[14] C. A. Rogers and C. Zong. Covering convex bodies by translates of convex bodies. *Mathematika*, 44(1):215–218, 1997.

[15] Oded Schramm. Illuminating sets of constant width. *Mathematika*, 35(2):180–189, 1988.

Chapter 9

Dömötör Pálvölgyi: Algorithmic problems

The minicourse focuses on three interesting results and their corollaries from the field of computer science. They do not build on each other, so missing a class is no problem.

First, I present a simple algorithm that draws a planar graph with n vertices onto a $2n-4 \times n-2$ grid.

Second, I will talk about online competitive algorithms focusing on the k-server problem. Here we have a graph on n vertices, with given edge lengths and some information in each vertex. We can move k servers that can read out and transfer the information. At every step one of the data is requested and our goal is to move the servers such that the sum of the distances they travel is minimized. How much does it help if we know the queries in advance? It turns out that it helps at most a constant factor.

Finally, I would like to state the PCP theorem (probabilistically checkable proof) and a few of its corollaries about approximations. It shows that one can verify a proof's correctness with high probability with reading just a few bits of it, if the proof is in a given format. A corollary is that it is very hard to tell whether a graph with n vertices contains a clique of size $n^{0.99}$ or if each clique is smaller than $n^{0.01}$.

9.1 Embedding planar graphs on a small grid

9.1.1 Planar graphs

A graph is *planar* if it can be embedded in the plane such that its vertices are points and its edges are non-crossing simple curves. We call an already embedded graph a *plane graph*. Planar graphs have several useful properties, maybe the most well-known is Euler's formula: $v(G)-e(G)+f(G) \geq 2$ where equality holds if and only if G is connected. Here $v(G)$ is the number of vertices, $e(G)$ the number of edges and $f(G)$ the number of faces, so we can already conclude that the number of faces does not depend on the embedding, only on the graph. From the formula we can also reduce that if G has at least 3 vertices, then $e(G) \leq 3v(G)-6$ and $f(G) \leq 2v(G)-4$ where equality holds for *triangulated* graphs, i.e. if all faces of the graph have 3 sides. Another simple consequence is that the *chromatic number*

of any planar graph G, $\chi(G)$ is at most 5. The strengthening of this, $\chi(G) \leq 4$, is the famous Four color theorem that was a conjecture for a long time.

Here we will focus on straight-line drawings, i.e. when every edge is embedded as a straight-line segment. The existence of such an embedding was discovered independently by Fáry, Tutte and Wagner. We will give a different proof of this result. In our case not only the edges will be straight-line segments, but even the vertices will have small, integer coordinates. This has applications in computer science to draw a graph on a screen or can be used in theoretical computer science to give a polynomial *witness* of the planarity of a graph.

9.1.2 Canonical ordering

We need the following observation.

Lemma 9.1.1. *Let G be a plane graph, whose exterior face is bounded by a cycle $u_1, u_2, \ldots, u_k$. Then there is a vertex u_i $(i \neq 1, k)$ not adjacent to any u_j other than u_{i-1} and u_{i+1}.*

Proof. If there are no two non-consecutive vertices along the boundary of the exterior face that are adjacent, then there is nothing to prove. Otherwise, pick an edge $u_iu_j \in E(G)$, for which $j > i+1$ and $j-i$ is minimal. Then u_{i+1} cannot be adjacent to any element of $\{u_1, \ldots, u_{i-1}, u_{j+1}, \ldots, u_k\}$ by planarity, nor can it be adjacent to any other vertex of the exterior face different from u_i and u_{i+2}, by minimality of $j-i$. □

Theorem 9.1.2 (Canonical Ordering)**.** *Let G be a triangulation of n vertices, with exterior face uvw. Then there is an ordering of the vertices $v_1 = u, v_2 = v, v_3, \ldots, v_n = w$ satisfying the following conditions for every k $(4 \leq k \leq n)$:*

- *(i) the boundary of the exterior face of the subgraph G_{k-1} of G induced by $\{v_1, v_2, \ldots, v_{k-1}\}$ is a cycle C_{k-1} containing the edge uv;*
- *(ii) v_k is in the exterior face of G_{k-1}, and its neighbors in $V(G_{k-1})$ are some (at least two) consecutive elements along the path obtained from C_{k-1} by removal of the edge uv. (See Figure 9.1.1)*

Proof. The vertices $v_n, v_{n-1}, \ldots, v_3$ will be defined by reverse induction. Set $v_n = w$, and let G_{n-1} be the graph obtained from G by the deletion of v_n. Since G is a triangulation, the neighbors of w form a cycle C_{n-1} containing uv, and this cycle is the boundary of the exterior face of G_{n-1}.

Let $4 \leq k \leq n$ be fixed and assume that $v_n, v_{n-1}, \ldots, v_k$ have already been determined so that the subgraph G_{k-1} induced by $V(G) \setminus \{v_k, v_{k+1}, \ldots, v_n\}$ satisfies condition (i) and (ii). Let C_{k-1} denote the boundary of the exterior face of G_{k-1}. Applying Lemma 9.1.1 to G_{k-1}, we obtain that there is a vertex u' on C_{k-1}, different from u and v, which is adjacent only to two other points of C_{k-1} (i.e., to its immediate neighbors). Letting $v_{k-1} = u'$, the subgraph $G_{k-2} \subseteq G$ induced by $V(G) \setminus \{v_{k-1}, v_k, \ldots, v_n\}$ obviously meets the requirements. □

Using this theorem, we can easily prove the main result of this section.

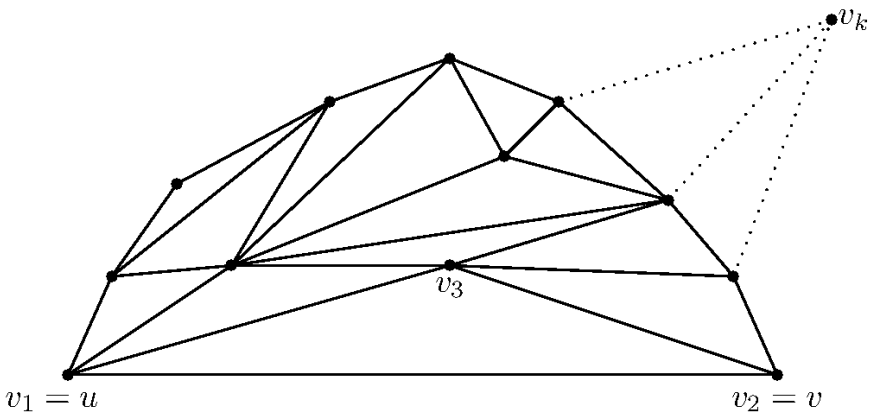


Figure 9.1.1: G_{k-1} and v_k in the exterior

Corollary 9.1.3. *Every planar graph has a straight-line embedding in the plane.*

Proof. It is sufficient to show that the statement is true for triangular planar graphs.

Let G be any triangulation with the canonical ordering $v_1 = u$, $v_2 = v$, $v_3, \ldots, v_n = w$, described above. We will determine the positions $f(v_k) = (x(v_k), y(v_k))$ of the vertices by induction on k.

Set $f(v_1) = (0,0)$, $f(v_2) = (2,0)$, $f(v_3) = (1,1)$. Assume that $f(v_1)$, $f(v_2)$, …, $f(v_{k-1})$ have already been defined for some $k \geq 4$ such that, connecting the images of the adjacent vertex pairs by segments, we obtain a straight-line embedding of G_{k-1}, whose exterior face is bounded by the segments corresponding to the edges of C_{k-1}. Suppose further that

$$\begin{aligned} &x(u_1) < x(u_2) < \ldots < x(u_m), \\ &y(u_i) > 0 \qquad \text{for } 1 < i < m, \end{aligned} \tag{9.1.1}$$

where $u_1 = u$, u_2, u_3, …, $u_m = v$ denote the vertices of C_{k-1} listed in cyclic order. By condition (ii) of Theorem 9.1.2, v_k is connected to u_p, u_{p+1}, …, u_q for some $1 \leq p \leq q \leq m$. Let $x(v_k)$ be any number strictly between $x(u_p)$ and $x(u_q)$. If we choose $y(v_k) > 0$ to be sufficiently large and connect $f(v_k) = (x(v_k), y(v_k))$ to $f(u_p)$, $f(u_{p+1})$, …,$f(u_q)$ by segments, then we obtain a straight-line embedding of G_k meeting all the requirements (including the auxiliary Hypothesis (9.1.1) for the vertices of C_k). □

9.1.3 Embedding on the grid

Now we shall restrict our attention to straight-line drawings, where each point is mapped into a *grid point*, i.e. a point with integer coordinates. Our goal is to minimize the size of the grid needed for the embedding of any planar graph of n vertices. The set of all grid points (x, y) with $0 \leq x \leq m$, $0 \leq y \leq n$ is said to be an $m \times n$ *grid*.

Theorem 9.1.4. *Any planar graph with n vertices has a straight-line embedding on the $2n-4$ by $n-2$ grid.*

Proof. It suffices to prove the theorem for triangulations. Let G be a triangulation with exterior face uvw, and let $v_1 = u$, $v_2 = v$, v_3, $\ldots$, $v_n = w$ be a canonical labelling of the vertices (see Theorem 9.1.2).

We are going to show by induction on k that G_k, the subgraph of G induced by $\{v_1, v_2, \ldots, v_k\}$, can be straight-line embedded on the $2k-4$ by $k-2$ grid, for every $k \geq 3$. Let f_3 be the following embedding of G_3:

$$f_3(v_1) = (0,0), f_3(v_2) = (2,0), f_3(v_3) = (1,1).$$

Suppose now that for some $k \geq 4$ we have already found an embedding $f_{k-1}(v_i) = (x_{k-1}(v_i), y_{k-1}(v_i))$, $1 \leq i \leq k-1$, with the following properties:

(*a*) $f_{k-1}(v_1) = (0,0)$, $f_{k-1}(v_2) = (2k-6, 0)$;

(*b*) If $u_1 = u$, u_2, $\ldots$, $u_m = w$ denote the vertices of the exterior face of G_{k-1} in cyclic order, then

$$x_{k-1}(u_1) < x_{k-1}(u_2) < \ldots < x_{k-1}(u_m);$$

(*c*) The segments $f_{k-1}(u_i)f_{k-1}(u_{i+1})$, $1 \leq i < m$, all have slope $+1$ or -1.

Note that (*c*) implies that the Manhattan distance $|x_{k-1}(u_j) - x_{k-1}(u_i)| + |y_{k-1}(u_j) - y_{k-1}(u_i)|$ between the image of any two vertices u_i and u_j on the exterior face of G_{k-1} is even. Consequently, if we take a line with slope $+1$ through u_i and a line with slope -1 through u_j, then they always intersect at a grid point $P(u_i, u_j)$.

Let $u_p, u_{p+1}, \ldots, u_q$ be the neighbours of v_k in G_k ($1 \leq p < q \leq m$). Clearly, $P(u_p, u_q)$ is a good candidate for $f_k(v_k)$, except that we may not be able to connect it to e.g. $f_{k-1}(u_p)$ by a segment avoiding $f_{k-1}(u_{p+1})$. To resolve this problem, we have to modify f_{k-1} before embedding v_k. We shall move the image of $u_{p+1}, u_{p+2}, \ldots, u_m$ one unit to the right, and then move the images of $u_q, u_{q+1}, \ldots, u_m$ to the right by an additional unit. That is, let

$$\tilde{x}_k(u_i) = \begin{cases} x_{k-1}(u_i), & \text{for } 1 \leq i \leq p, \\ x_{k-1}(u_i) + 1, & \text{for } p < i < q, \\ x_{k-1}(u_i) + 2, & \text{for } q \leq i \leq m, \end{cases}$$

$$y_k(u_i) = y_{k-1}(u_i), \quad \text{for } 1 \leq i \leq m\,,$$

and let $f_k(v_k)$ be the point of intersection of the lines of slope $+1$ and -1 through $f_k(u_p)$ and $f_k(u_q)$, respectively. Of course, $f_k(v_k)$ is a grid point that can be connected by disjoint segments to the points $f_k(u_i) = (x_k(u_i), y_k(u_i))$, $p \leq i \leq q$, without intersecting the polygon $f_k(u_1)f_k(u_2)\ldots f_k(u_m)$. However, as we move the image of some u_i, it may be necessary to move some other points (not on the exterior face) as well, otherwise we may create crossing edges.

In order to tell exactly which set of points has to move together with the image of a given exterior vertex u_i, we define recursively a total order '$\prec$' on $\{v_1, v_2, \ldots, v_n\}$. Originally, let $v_1 \prec v_3 \prec v_2$. If the order has already been defined on $\{v_1, v_2, \ldots, v_{k-1}\}$, then insert v_k just before u_{p+1}. According to this rule, obviously

$$u_1 \prec u_2 \prec \cdots \prec u_m\,.$$

Now we can extend the definition of f_k to the interior vertices of G_{k-1}, as follows. For any $1 \leq i \leq k-1$, let

$$\tilde{x}_k(v_i) = \begin{cases} x_{k-1}(v_i), & \text{if } v_i \prec u_{p+1}, \\ x_{k-1}(v_i)+1, & \text{if } u_{p+1} \preceq v_i \prec u_q, \\ x_{k-1}(v_i)+2, & \text{if } u_q \preceq v_i, \end{cases}$$

$$y_k(v_i) = y_{k-1}(v_i)\,.$$

Evidently, f_k satisfies conditions (a),(b) and (c).

To complete the proof, it remains to verify that f_k is a straight-line embedding, i.e., no two segments cross each other. A slightly stronger statement follows by straightforward induction.

Lemma 9.1.5. *Let $f_{k-1} = (x_{k-1}, y_{k-1})$ be the straight-line embedding of $G_k - 1$, defined above, and let $\alpha_1, \alpha_2, \ldots, \alpha_m \geq 0$. For any $1 \leq i \leq k-1$, $1 \leq j \leq m$, let*

$$x(v_i) = x_{k-1}(v_i) + \alpha_1 + \alpha_2 + \cdots + \alpha_j \text{ if } u_j \preceq v_i \prec u_{j+1}\,,$$

$$y(v_i) = y_{k-1}(v_i)\,.$$

Then $f'_{k-1} = (x, y)$ is also a straight-line embedding of G_{k-1}.

The claim is trivial for $k = 4$. Assume that it has already been confirmed for some $k \geq 4$, and we want to prove the same statement for G_k. The vertices of the exterior face of G_k are $u_1, \ldots, u_p, v_k, u_q, \ldots, u_m$. Fix now any nonnegative numbers $\alpha(u_1), \ldots, \alpha(u_p), \alpha(v_k), \alpha(u_q), \ldots, \alpha(u_m)$. Applying the induction hypothesis to G_{k-1} with $\alpha_1 = \alpha(u_1), \ldots, \alpha_p = \alpha(u_p), \alpha_{p+1} = \alpha(v_k) + 1, \alpha_{p+2} = \cdots = \alpha_{k-1} = 0, \alpha_q = \alpha(u_q) + 1, \alpha_{q+1} = \alpha(u_{q+1}), \ldots, \alpha_m = \alpha(u_m)$, we obtain that the restriction of f'_k to G_{k-1} is a straight-line embedding. To see that the edges of G_k incident to v_k do not create any crossing, it is enough to notice that f_k and f'_k map $\{u_{p+1}, \ldots, u_{q-1}\}$ into congruent sets. □

9.2 Online competitive algorithms

9.2.1 Baby example

As an example, consider the following problem. When a baby is born, the parents need a *baby scale* to measure how much she eats. To get a baby scale, they have three options.
1) Buy one for 30€.
2) Rent one for 5€/month.
3) Borrow one from a friend.

Let us rule out the mathematically less fascinating third option and suppose they only have the first two options. It is not hard to decide which to choose *if* they know for how long they need the scale; for less than five months rent and for more months buy. (Here we suppose that the scale will have no value for them later - we could easily modify this condition by subtracting the price for which they can sell it later from the initial price.) But what if they have no clue at all? One option would be to guess and calculate some expected values from

the probabilities. However, there can be too many factors (how well the baby is gaining weight, number of future children) to make any reasonable estimates. Another option is to try to minimize their later regrets, to make sure they could not have done much better.

For example if they decide to buy one (30€) and need it for only one month (5€), then their *competitive ratio* is 6:1 (compared to the best possible choice they could have made). However, if they decide to rent and are blessed with many children and, say, 18 months of scale usage (90€), their ratio becomes 3:1 and could be even worse. So what should they do to minimize the competitive ratio?

The answer first might seem counterintuitive but the best is to mix the above strategies - first rent for a while, then buy. After a little thinking, we can realize that in fact this is the only thing that makes sense and turns out to be not that a crazy idea after all. Now the only thing left to decide is for how long to rent before buying.

Suppose we rent for R month and then buy if we still need the scale. This way we spend $5i$ if we need it for $i \leq R$ months and $5R + 30$ if we need it for at least $R + 1$. The best option would be either to rent the whole time (for $5i$ if we need it for $i \leq 6$ months) or to buy immediately (for 30 if we need it for at least 6 months). Our ratio against the renting option is worst if we need the scale for exactly $R + 1$ months, in this case we get $\frac{5R+30}{5R+5}$. Our ratio against the buying option is of course also worst if we need the scale for at least $R + 1$ months, in this case we get $\frac{5R+30}{30}$. So our goal is to minimize $max(\frac{5R+30}{5R+5}, \frac{5R+30}{30})$ by suitably choosing R. For this we solve $\frac{5R+30}{5R+5} = \frac{5R+30}{30}$ which gives $R = 5$, so we have to buy in the sixth month, which is exactly what we would have done with my wife, but we needed the scale for only five month. So with the next child, we buy from the start...

9.2.2 k-server problem

In the *k-server problem* we control k servers each of which occupies one point in a given finite metric space from which it can move to another one for the cost of the distance between them. There is a series of requests, each of which is a point where we have to move a server (if there is no server present there at the moment). Our goal is to keep our total cost as small as possible. Since we do not know anything against the requests, the best we can try is to minimize the competitive ratio of our algorithm against the cost of what would have been the best sequence of moves, known as the *offline optimum*. It is conjectured that there is an algorithm that is k-competitive[1] but the best known algorithm is only $2k - 1$-competitive. An interesting special case is when all distances are the same is called the *k-paging problem*. First we show that already in this case we cannot hope to have a $< k$-competitive algorithm.

Claim 9.2.1. *No online algorithm can achieve a better competitive ratio than k for the k-paging problem if the metric space has at least $k + 1$ points.*

Proof. Suppose that the space has exactly $k + 1$ points (if it has more, we never request them). Every time we request the point that has no server on it (no optimal

[1]Here in the definition of the ratio we are interested in the asymptotic behavior and ignore additive constants.

algorithm would put two servers to the same point, so we can suppose that there is exactly one such point). This way after R requests, the cost of the algorithm is R. However, the best choice would be at each step to move the server whose location would be requested the latest, so after at least $k-1$ further requests. Thus the offline optimum is at most $\lceil \frac{R}{k} \rceil$. □

Next we present an algorithm that for a space with $k+1$ points achieves a competitive ratio of k. Denote by $D(i)$ the distance traveled by server i before the request and by $d(i)$ the distance of server i from the requested location. The algorithm called $BALANCE$ has the following simple rule:

Always move the server for which $D(i)+d(i)$ is minimized.

So if e.g. we have three servers, the first has traveled 3, the second 4 and the third 6 units until the query, which is at distance 4 from the first, at distance 2 from the second and at distance 1 from the third server. In this case $BALANCE$ moves the second server, as that gives a minimum distance of 6 after the move, while the other two would give 7.

Proof. We can suppose that the request is always the only unoccupied location. First we need to make some definitions. Define $d(i,j)$ as the distance between location i and j. Let R^t be the t-th request. Let opt_i^t be the offline optimum of the first t requests that has no server on location i (if $R^t = i$, i.e. the last request was i, then an extra move must be made after it to move away the server from it). Finally, let D_i^t be the distance traveled by the server at location i after t requests (if $i \neq R^{t+1}$, since that place is unoccupied).

Observation 9.2.2. *$opt_i^{t+1} = opt_i^t$ if $i \neq R^{t+1}$ and*
$opt_{R^{t+1}}^{t+1} = min_{i \neq R^{t+1}} opt_i^t + d(i, R^{t+1})$.

Lemma 9.2.3. *For every $i \neq R^{t+1}$ we have $D_i^t \leq opt_i^t$.*

Proof. We prove this by induction on t. Let $h = R^{t+1}$ and m denote the location for which $D_m^t + d(m,h)$ is minimal (in fact $m = R^{t+2}$). If $i \neq m, h$, then $D_i^{t+1} = D_i^t$ and since $opt_i^{t+1} \geq opt_i^t$, we are done. Otherwise, we have $D_h^{t+1} \leq D_i^t + d(i,h)$ for all $i \neq h$, by the choice of the server we moved to the empty position. But using induction we have $D_i^t + d(i,h) \leq opt_i^t + d(i,h)$ and using the previous observation there is an $i \neq h$ for which $opt_i^t + d(i,h) = opt_h^{t+1}$. Putting the inequalities together, we get exactly what we wanted, $D_h^{t+1} \leq opt_i^t + d(i,h)$. □

From here the proof of the theorem follows from $\sum_i D_i^t \leq \sum_i opt_i^t \leq k \cdot$ (*offline optimum* + *largest distance*). □

9.2.3 Randomization

Another, very interesting problem emerges if we allow *randomized* online algorithms. Here we can measure the competitiveness depending on what kind of offline optimum we take. We imagine that the requests are given by some adversary and we distinguish the three following types.

Oblivious: The requests are generated in advance.

Adaptive Online: The requests are generated depending on the moves so far but the adversary must also make its moves online.

Adaptive Offline: The requests are generated depending on the moves so far and the adversary can decide its moves after all the requests are made.

By definition, we have the following relations among the respective competitive ratios: $\mathcal{C}_{OB} \leq \mathcal{C}_{ADON} \leq \mathcal{C}_{ADOFF} \leq \mathcal{C}_{DET}$.

While $\mathcal{C}_{OB}$ can be much smaller than $\mathcal{C}_{DET}$ (e.g. about $\log k$ for the k-paging problem with $k+1$ locations), the other two quantities are not that far. We can prove this through a few simple statements.

Claim 9.2.4. *If for G and H algorithms we have $\mathcal{C}_{ADON}(G) \leq \alpha$ and $\mathcal{C}_{OB}(H) \leq \beta$, then $\mathcal{C}_{ADOFF}(G) \leq \alpha\beta$.*

Claim 9.2.5. *If there are finitely many options at each request, then $\mathcal{C}_{ADOFF} = \mathcal{C}_{DET}$.*

Corollary 9.2.6. *If there are finitely many options at each request, then $\mathcal{C}_{DET} \leq (\mathcal{C}_{ADON})^2$.*

9.3 Probabilistically checkable proofs

Note that depending on time and interest, we might completely skip this section and rather to some earlier topics, or just cover parts of it, so the write-up is also more sketchy.

9.3.1 Interactive proofs

A very famous open problem of theoretical computer science is the *graph isomorphism problem*. The input is two graphs, G and H, and our goal is to decide whether they are *isomorphic* or not. Of course special cases are easy to decide, e.g. if they do not have the same number of vertices, or edges, then they cannot be isomorphic. It is also not hard to come up with an algorithm if one of them is a tree. In case they are isomorphic, at least we have a *certificate* for it - an isomorphism. (This means that the problem belongs to NP.) But it is not even known whether we can certify in such a way that they are *not* isomorphic. (This means that we do not know whether the problem belongs to $co-NP$.) However, we have a certain interactive certificate for non-isomorphism.

Suppose that P (*Prover*) can distinguish between G and H and wants to convince V (*Verifier*) of this fact. If his method is something easy to compute (like count the number of vertices whose degree is 7 and they differ), then of course he can just reveal the method to V who can check this fact himself. But maybe what P does is more complicated, maybe he has computational powers that are far bigger than V's (like me submitting a request to Wolfram|Alpha). Of course if V believes P, then his word is sufficient (like in my case) but suppose that V does

not trust P and wants to verify that he is not lying or mistaken. How can he do it?

V can use the following simple *interactive protocol*. He takes one of G or H randomly and permutes its vertices, obtaining the graph R. Then he asks P whether G and R are isomorphic or not. If P knows how to distinguish between G and H, then he can easily answer this. If not, then he can only guess randomly, which means he makes a mistake at least 50% of the time if R was selected randomly. By repeating the above several times, V can be quite sure that P indeed knows how to differentiate between G and H.

Notice that the above protocol also has the nice property that V does not learn anything about P's method, so he can keep it secret. Interactive proofs of this type have a large theory, called *zero knowledge proofs*.

9.3.2 PCP theorem

Suppose someone claims that they have proved an important mathematical conjecture, like $P = NP$. Of course such claims are usually incorrect but it takes a great effort to read through a long paper to find a mistake in it. The PCP theorem (informally) says that every proof can be (algorithmically) transformed into another whose correctness can verified easily - in the new proof reading only a constant number of bits (letters) will reveal any mistake with at least 50% chance! Moreover, repeating this several times (each time with different, randomly chosen letters) we can reduce the probability of the error to arbitrarily small. Also, the length of the new proof won't be much more than the old one (it only grows from n to $O(npolylogn)$).

9.3.3 Hardness of approximation

Using the PCP theorem, it can be shown that many problems cannot be *approximated* efficiently, unless certain surprising fact were true. Here we sketch one of these proofs.

Suppose that there is an efficient (i.e. polynomial time) algorithm that computes an approximation $f(G)$ of the clique number ($\omega(G)$) of any graph G such that $\omega(G) \leq f(G) \leq 2\omega(G)$. Then we can efficiently decide of any statement whether it has a not too long proof (e.g. if we are looking for a proof of length n, then we can decide it in $poly(n)$ time).

Proof. By the PCP theorem, any such statement also has a proof P that is not much longer and must be checked only at a few places, and if the proof is incorrect, it is discovered with probability at least 50%. Let us denote the verifying algorithm by V and suppose that V uses $k \leq O(\log n)$ random bits and reads $b = O(1)$ bits of P.

For any statement, we construct a graph G as follows. The nodes of G will be certain pairs (r, a), where $r \in \{0,1\}^k$ and $a \in \{0,1\}^b$. To decide if such a pair is a node, we start the algorithm V with the given statement and with r as its random bits. After some computation, it tries to look up b entries of P; we give it the bits $a_1, \ldots, a_b$. At the end, it outputs 0 or 1; we take (r, a) as a node of G if it outputs 1.

To decide if two pairs (r, a) and (r', a') are adjacent, let us also remember which positions in P the algorithm V tried to read when starting with r, and also when starting with r'. If there is one and the same position read both times, but the corresponding bits of a and a' are different, we say that there is a *conflict*. If there is no conflict, then we connect (r, a) and (r', a') by an edge. Note that the graph G can be constructed in polynomial time.

Suppose that the statement has a short proof. For every sequence $r = r_1 \dots r_k$ of random bits, we can specify a string $a \in \{0,1\}^b$ such that $(r, a) \in V(G)$, namely the string of bits that the algorithm reads from P when started with the random sequence r. Furthermore, it is trivial that between these there is no conflict, so these 2^k nodes form a clique. Thus in this case $\omega(G_x) \geq 2^k$.

Now suppose that the statement does not have a short proof, and assume that the nodes $(r^1, a^1), (r^2, a^2), \dots, (r^N, a^N)$ form a clique of size $\omega(G)$ in G. The strings $r^1, r^2, \dots$ must be different; indeed, if $r^i = r^j$, then V tries to look up the same positions in P in both runs, so if there is no conflict, then we must have $a^i = a^j$.

We create a string P as follows. Initially all its bits are empty. We run V with the random bits r^1, and we insert the bits of a^1 in the b positions which the algorithm tries to look up. Then we repeat this with random bits r^2; if we need to write in a position we already have a bit in, we do not have to change it since (r^1, a^1) and (r^2, a^2) are connected by an edge. Similarly, we can enter the bits of each a^i in the appropriate positions without having to overwrite previously entered bits.

At the end, certain positions in P will be filled. We fill the rest arbitrarily. It is clear from the construction that for the string P constructed this way, V will accept with probability at least $\frac{N}{2^k}$, what must be less than $\frac{1}{2}$ if the statement has no short proof. Therefore $\omega(G) = N < 2^{k-1}$.

Now if a polynomial time algorithm exists that computes a value $f(G)$ such that $\omega(G) \leq f(G) \leq 2\omega(G)$, then we have $f(G) \geq \omega(G) \geq 2^k$ if the statement has a short proof, but $f(G) \leq 2\omega(G) < 2^k$ if it does not, so we can efficiently decide this. □

computing, in Proc. of 11th ACM Symposium on Theory of Computing, 1979, 209-213.

Chapter 10

Gergely Zábrádi: p-adic numbers and applications

These are the notes for the course 'p-adic numbers and applications' at the Summer School for undergraduates at ELTE, July 2013.

In this course we intend to advertise the usefulness and relevance of the p-adic numbers. Instead of concentrating on the proof of one particular theorem, our goal is to give an idea of 1) how things work in the p-adic world; 2) what questions can be answered using them; 3) what directions of current research there are.

The book [4] that we will mostly follow in motivating p-adics is an excellent introduction. The books [7, 12] are more advanced. The former gives a concise introduction to the theory of p-adic L-functions (and zeta-functions) and the latter contains an elementary proof of the Hasse-Minkowski theorem.

10.1 Why p-adic numbers?

Historically, the main motivation for the developement of algebraic number theory was the attempt to prove Fermat's Last Theorem, ie. when $n \geq 3$ is an integer then there are no integer solutions of $x^n + y^n = z^n$ with $xyz \neq 0$. This was such a problem in mathematics whose solution required the systematic study of several areas and led to the developement of arithmetic geometry, among many others.

Arithmetic algebraic geometry is the area of mathematics dealing with the rational or integer solutions of polynomial equations.

Over the last century, p-adic numbers have played a very important role in arithmetic geometry. They were introduced by Kurt Hensel in 1897 motivated by the analogies of $\mathbb{Z}$ with field of fractions $\mathbb{Q}$ and $\mathbb{C}[t]$ (complex polynomials in 1 variable) with field of fractions $\mathbb{C}(t)$. Note for instance that both $\mathbb{Z}$ and $\mathbb{C}[t]$ are unique factorisation domains, ie. any element can be decomposed (upto units uniquely) as a product of primes. While the primes in $\mathbb{Z}$ are the usual prime numbers, the primes in $\mathbb{C}[t]$ are the linear polynomials $t - \alpha$ ($\alpha \in \mathbb{C}$). Moreover, any rational number can be written as the quotient of two integers; similarly, any rational function can be—by definition—written as a quotient of two polynomials. The analogy, in fact, is much deeper. We may write each polynomial $P(t)$ in the form $P(t) = a_0 + a_1(t-\alpha) + a_2(t-\alpha)^2 + \cdots + a_n(t-\alpha)^n$ (with $a_i \in \mathbb{C}$) and each integer $m \geq 0$ in the form $m = a_0 + a_1 p + \cdots + a_n p^n$ with $a_i \in \{0, 1, \ldots, p-1\}$. The

expansion in $t-\alpha$ will show, for example, whether or not the polynomial vanishes at $t=\alpha$ and if so, to which order. On the other hand, for integers this expansion tells us to what order p divides m. Moreover, in case of quotients of polynomials we can push this further. Taking $f(t)=P(t)/Q(t)$ and $\alpha\in\mathbb{C}$ we may expand

$$f(t)=a_{n_0}(t-\alpha)^{n_0}+a_{n_0+1}(t-\alpha)^{n_0+1}+\cdots=\sum_{i\geq n_0}a_i(t-\alpha)^i\ .$$

This is called the Laurent series expansion of f around α.

- We can have $n_0<0$ here—this happens if and only if the order of the root of $Q(t)$ at α is bigger than the order of the root of $P(t)$ at α. In complex analysis we say in this case that f has a *pole* at α of order $-n_0$.
- The expansion will not be finite. In fact, it will only be finite if $Q(t)$ is constant times a power of $t-\alpha$. One can show that the series will be convergent in a punctured neighbourhood of α, but for now we regard the expression above as a formal Laurent series, ignoring the question of convergence.

Why don't we try the same for the rational numbers? We may write both the numerator and denominator in base p and divide formally. For example, with $p=5$ we obtain

$$\frac{35}{31}=\frac{2p+p^2}{1+p+p^2}=2p+4p^2+3p^3+p^4+4p^5+4p^6+\dots\ .$$

To check that this is indeed correct we multiply both sides by $31=1+p+p^2$ and use $p=5$ to compute (expanding upto p^6)

$$\begin{aligned}(1+p+p^2)(2p+4p^2+3p^3+p^4+4p^5+4p^6+\dots)=\\ =(2p+2p^2+2p^3)+(4p^2+4p^3+4p^4)+(3p^3+3p^4+3p^5)+\\ +(p^4+p^5+p^6)+4p^5+4p^6+4p^6+\cdots=\\ =2p+6p^2+9p^3+8p^4+8p^5+8p^6+\cdots=\\ =2p+p^2+10p^3+8p^4+8p^5+8p^6+\cdots=\\ =2p+p^2+10p^4+8p^5+8p^6+\cdots=2p+p^2\end{aligned}$$

This above is not precise at all (with all those dots) but at least you should get the feeling what is going on. However, it is easy to check that this can always be done and the process gives an infinite expansion

$$\frac{a}{b}=a_{n_0}p^{n_0}+a_{n_0+1}p^{n_0+1}+\dots$$

of any (positive, for now) rational number a/b with $a_i\in\{0,1,\dots,p-1\}$. This even reflects the properties of the rational number a/b "near p" (or "locally at p"), ie. if $(a,b)=1$ then $n_0<0$ if and only if $p\mid b$. You could ask what happens to the negatives? As any negative number is a product of -1 and a positive number, it suffices to expand -1:

$$-1=(p-1)+(p-1)p+(p-1)p^2+\cdots+(p-1)p^n+\dots\ .$$

If we return for the moment to the case of rational functions, each $f(t)\in\mathbb{C}(t)$ can be expanded as Laurent series at each primes $t-\alpha$. However, we have seen

many functions at calculus class having a Laurent (even Taylor) series expansion that are not a quotient of two polynomials, for instance e^t or $\sin t$. We may even ignore convergence and take the field $\mathbb{C}((t))$ of all formal Laurent series (with finite "tail"). The field $\mathbb{C}(t)$ of rational functions is a subfield of this. The field $\mathbb{Q}_p$ of p-adic numbers is the analogue of $\mathbb{C}((t))$, ie. the set

$$\mathbb{Q}_p = \{a_{n_0}p^{n_0} + a_{n_0+1}p^{n_0+1} + \cdots \mid a_i \in \{0, 1, \ldots, p-1\},\ n_0 \leq i\}$$

of finite-tailed (= "finite to the *left*", but usually infinite to the *right*) Laurent series with the above described multiplication and addition. Note that unlike in $\mathbb{C}((t))$ we need to "carry over", e.g. $(2+0\cdot p+\ldots)+((p-1)+0\cdot p+\ldots) = 1+1\cdot p+\ldots$. We denote by $\mathbb{Z}_p$ the subring of those elements in $\mathbb{Q}_p$ with $n_0 \geq 0$. This subset is indeed closed under addition and multiplication.

10.1.1 Exercises

Exercise 10.1.1. Suppose that $f(t) = P(t)/Q(t)$ is in lowest terms so that $P(t)$ and $Q(t)$ do not have common zeros. Show that the expansion of $f(t)$ in $t - \alpha$ is finite if and only if $Q(t) = a_m(t-\alpha)^m$ for some $0 \leq m \in \mathbb{Z}$ and $0 \neq a_m \in \mathbb{C}$.

Exercise 10.1.2. Consider a p-adic number $x = a_0 + a_1p + \cdots + a_np^n + \ldots$. What is the expansion of $-x$?

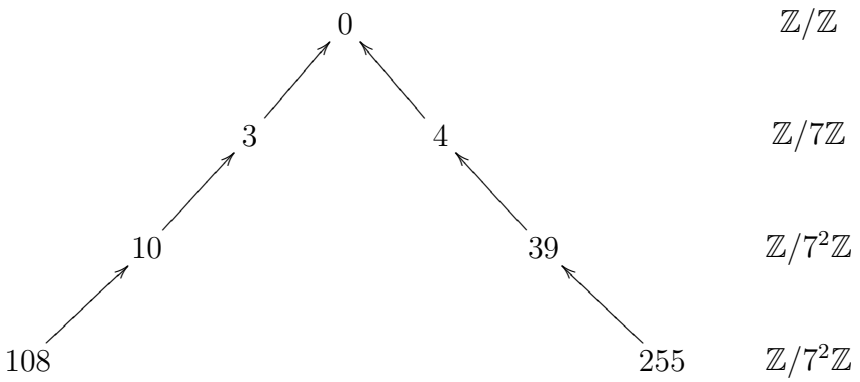
Exercise 10.1.3. Show that $\mathbb{Q}_p$ is indeed a field.

Problem 10.1.4. Prove that the p-adic expansion of an element in $\mathbb{Q}_p$ is eventually periodic if and only if the element is rational (ie. lies in $\mathbb{Q}$). *Hint:* Mimic the proof of the analogous statement in $\mathbb{R}$

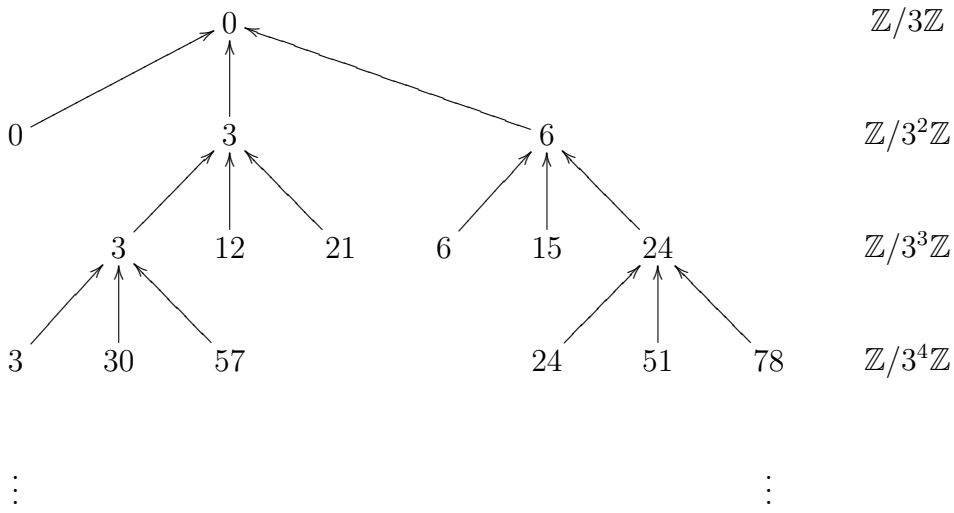
10.2 Solving equations in $\mathbb{Q}_p$

We would like to illustrate how solving equations in the p-adics is related to solving equations modulo p^n. For example, take $p = 7$ and consider the equation $x^2 = 2$. Solve it first mod 7, we find right away that $x \equiv \pm 3 \pmod 7$ is a solution. Then proceed to mod 7^2 and look for the solution in the form $x = 3 + 7x_1$ (or $x = -3 + 7x_1$). $(3 + 7x_1)^2 = 9 + 42x_1 + 49x_1^2 \equiv 9 + 42x_1 \pmod{49}$, so we need $9 + 42x_1 \equiv 2 \pmod{49}$, that is $x_1 \equiv 1 \pmod 7$. Note that in this second step we only need to solve a *linear* equation, not quadratic any more. Now we go on to 7^3 and look for the solution in the form $x = 3 + 1 \cdot 7 + 7^2x_2$. By a similar calculation we obtain $x_2 \equiv 2 \pmod 7$. And so forth we obtain a solution $x = 3 + 7 + 2 \cdot 7^2 + \cdots \in \mathbb{Q}_7$ of the equation $x^2 = 2$. (In particular we see that $\mathbb{Q} \subsetneq \mathbb{Q}_7$.) Similarly, we will also find a solution of the form $x = 4 + 5 \cdot 7 + \cdots \in \mathbb{Q}_7$ starting with the solution $-3 \equiv 4 \pmod 7$. As $mathbbQ_7$ is a field, we have found all the solutions. All this worked out pretty well because 7 does not divide the

discriminant of the polynomial $x^2 - 2$. The tree of solutions in $\mathbb{Z}_7$ looks like



What happens if the prime p does divide the discriminant of our equation? Let us have a look at the equation $x^2 = 9$ in $\mathbb{Q}_3$. Modulo 3 this has only one double root, $x \equiv 0 \pmod 3$. So we are looking for the solution in the form $x = 3x_1$ and $(3x_1)^2 \equiv 9 \equiv 0 \pmod 9$ is satisfied trivially for any $x_1 = 0, 1, 2$, therefore we have 3 solutions of $x^2 = 9$ in $\mathbb{Z}/9\mathbb{Z}$, namely $0, 3$, and 6. Now we look at the equation mod $3^3 = 27$. $(3x_1)^2 \equiv 9 \pmod{27}$ has solutions $x_1 \equiv 1, 2 \pmod 3$. Hence we have $\{x \in \mathbb{Z}/27\mathbb{Z} \mid x^2 = 9\} = \{3, 6, 12, 15, 21, 24\}$. In other words, the solutions $x \equiv 3, 6 \pmod 9$ can be lifted to a solution mod 27 in three ways, but the solution $x \equiv 0 \pmod 9$ cannot be lifted. By proceeding further, it is not hard to see that we will always have either 3 or 0 lifts of each solution mod 3^n to a solution mod 3^{n+1} for all $n \geq 1$ and the tree



of solutions have 2 infinite branches (and many finite) contending to the fact that there are only 2 solutions (namely $x = \pm 3$) in $\mathbb{Q}_3$.

10.2.1 Exercises

Exercise 10.2.1. Give a rigorous proof that the above process gives you a solution of $x^2 = 2$ in $\mathbb{Q}_7$.

Exercise 10.2.2. Prove that $x^2 + 1 = 0$ has a solution in $\mathbb{Q}_5$, but not in $\mathbb{Q}_7$. Can you describe the primes p for which this equation has a solution in $\mathbb{Q}_p$?

Problem 10.2.3. Show that if $f(x) \in \mathbb{Z}[x]$ is a *monic* polynomial and p is a prime then all the solutions of $f(x) = 0$ in $\mathbb{Q}_p$ lie in fact in $\mathbb{Z}_p$. *Hint:* Prove by contradiction and try and compute the first nonzero term of $f(\alpha)$ for $\alpha \in \mathbb{Q}_p \setminus \mathbb{Z}_p$.

Problem 10.2.4. Prove that the field $\mathbb{Q}_p$ is not algebraically closed for any prime number p. Can you construct an irreducible polynomial over $\mathbb{Q}_p$ of any given degree $0 < n \in \mathbb{Z}$?

Problem 10.2.5. Verify that the inclusion $\mathbb{Q} \hookrightarrow \mathbb{Q}_p$ is strict for any prime number p. *Hint:* You could argue by noting that the cardinality of $\mathbb{Q}_p$ is bigger than the cardinality of $\mathbb{Q}$, but there is also an algebraic argument.

10.3 Precise definition of $\mathbb{Q}_p$

Definition 10.3.1. *Let K be a field. We call a function $|\cdot|\colon K \to \mathbb{R}^{\geq 0}$ an absolute value (or multiplicative valuation) on K, if it satisfies*

(1) $|x| = 0 \iff x = 0$;

(2) $|xy| = |x||y|$;

(3) $|x + y| \leq |x| + |y|$ *(triangle inequality).*

The absolute value $|\cdot|$ induces a metric $d(x, y) := |x - y|$ on K. This way K becomes a metric space, in particular, there is a topology on it.

Example 10.3.2. *The* trivial *absolute value: $|x| = 1$ if $x \neq 0$ and $|0| = 0$.*

Definition 10.3.3. *We say that the two absolute values $|\cdot|_1$ and $|\cdot|_2$ on K are* equivalent*, if they induce the same topology.*

Proposition 10.3.4. *$|\cdot|_1$ and $|\cdot|_2$ are equivalent if and only if there exists a real number $s > 0$ such that $|x|_1 = |x|_2^s$ for all $x \in K$.*

Proof. The implication $\Leftarrow$ is trivial. Conversely, note that $|x|_i < 1$ holds if and only if the powers of x tend to zero in the absolute value $|\cdot|_i$ $(i = 1, 2)$. Hence if $|\cdot|_1$ and $|\cdot|_2$ induce the same topology then $|x|_1 < 1 \iff |x|_2 < 1$. Applying this to $x = a/b$ and $x = b/a$ we obtain $|a|_1 \leq |b|_1 \Leftrightarrow |a|_2 \leq |b|_2$ $(a, b \in K)$. In particular, if one of $|\cdot|_1$ and $|\cdot|_2$ is trivial then so is the other. Therefore we may assume that there exists a $y \in K$ such that $|y|_1 > 1$ (whence $|y|_2 > 1$), so we choose $0 < s := \log_{|y|_2} |y|_1 \in \mathbb{R}$ so that we have $|y|_1 = |y|_2^s$. Now for any $0 \neq x \in K$ there is an $\alpha = \alpha(x) \in \mathbb{R}$ with $|x|_1 = |y|_1^\alpha$. We choose the sequence $(\frac{m_i}{n_i})_{i \in \mathbb{N}}$ $(m_i, n_i \in \mathbb{Z},\ n_i \neq 0)$ of rational numbers so that $\lim_{i \to \infty} \frac{m_i}{n_i} = \alpha + 0$. We obtain $|x|_1 = |y|_1^\alpha < |y|_1^{m_i/n_i}$, hence $|x^{n_i}|_1 < |y^{m_i}|_1$, whence $|x^{n_i}|_2 < |y^{m_i}|_2$, ie. $|x|_2 < |y|_2^{m_i/n_i}$. Letting $i \to \infty$ we deduce $|x|_2 \leq |y|_2^\alpha$. The inequality $|x|_2 \geq |y|_2^\alpha$ is proven in a similar fashion, so we have $|x|_1 = |y|_1^\alpha = |y|_2^{s\alpha} = |x|_2^s$ for all $0 \neq x \in K$ (and, of course, also for $x = 0$). □

Definition 10.3.5. *We say that the absolute value $|\cdot|$* non-archimedean *if the set $\{|n \cdot 1|\colon n \in \mathbb{Z}\} \subseteq \mathbb{R}$ is bounded. Otherwise $|\cdot|$ is archimedean.*

Remark. The above definition is equivalent to saying that the ring homomorphism $f\colon \mathbb{Z} \to K$, $f(1) = 1$ has bounded image in K if and only if $|\cdot|$ is non-archimedean.

Example 10.3.6. *1. The trivial absolute value is non-archimedean.*

2. *The usual absolute value (that we denote by $|\cdot|_\infty$ in this note) on $\mathbb{R}$ (or on $\mathbb{C}$, or on any subfield $K \leq \mathbb{C}$) is archimedean.*

3. *Let p be a prime. The p-adic absolute value $|\cdot|_p$ on $\mathbb{Q}$ is defined by $|\frac{a}{b}p^n|_p = p^{-n}$ where $p \nmid a, b \in \mathbb{Z}$ (and $|0|_p = 0$). This is non-archimedean, since whenever $\frac{a}{b}p^n \in \mathbb{Z}$ we have $n \geq 0$ and $|\frac{a}{b}p^n|_p = p^{-n} \leq 1$.*

Proposition 10.3.7. *The absolute value $|\cdot|$ is non-archimedean if and only if the so called* ultrametric *inequality holds:*

(3′) $|x + y| \leq \max(|x|, |y|)$.

Moreover, if $|\cdot|$ is non-archimedean then $\{|n \cdot 1|, n \in \mathbb{Z}\}$ is not only bounded, but bounded by 1.

Proof. If (3′) holds then we have $|n \cdot 1| \leq |1| = 1$. On the other hand, for $0 < k \in \mathbb{Z}$, $|x| \geq |y|$ and $|n \cdot 1| \leq C$ for some $0 < C \in \mathbb{R}$ then we have

$$|x+y|^k = |(x+y)^k| = |\sum_{j=0}^{k} \binom{k}{j} x^j y^{k-j}| \leq$$
$$\leq \sum_{j=0}^{k} |\binom{k}{j} \cdot 1||x|^j |y|^{k-j} \leq \sum_{j=0}^{k} C|x|^k = (k+1)C|x|^k \ .$$

Taking k^{th} root and letting $k \to \infty$ the statement follows. □

Theorem 10.3.8 (Ostrowski)**.** *On $\mathbb{Q}$ any nontrivial absolute value $|\cdot|$ is equivalent to either the real $|\cdot|_\infty$ or the p-adic $|\cdot|_p$ absolute value for some prime p. These valuations are pairwise inequivalent.*

Proof. Case 1: $|\cdot|$ is non-archimedean. If we have $|p| = 1$ for all primes p then the absolute value is trivial (see Exercise 10.3.1). So we may take a prime p such that $\|p\| < 1$. Therefore the set $A := \{a \in \mathbb{Z}\colon \|a\| < 1\}$ contains p and is an ideal in $\mathbb{Z}$ as it is closed under addition by (3′) and also by multiplication by any integer because of (2) (see Proposition 10.3.7). On the other hand, $1 \notin A$ so we have $A = (p)$ as (p) is a maximal ideal in $\mathbb{Z}$. Hence for $p \nmid a, b \in \mathbb{Z}$ we have $|a| = |b| = 1$ and $|\frac{a}{b}p^n| = |p|^n = |\frac{a}{b}p^n|_p^s$ where $s := \log_{1/p}|p|$.

Case 2: $|\cdot|$ is archimedean. Let $1 < m, n \in \mathbb{Z}$ be arbitrary.

Lemma 10.3.9. *We have $|m|^{1/\log m} = |n|^{1/\log n}$. (Here* log *denotes, say, the natural logarithm, in fact the base doesn't matter.)*

Proof. Write m in base n, ie. $m = \sum_{i=0}^{r} a_i n^i$ where $0 \leq a_i < n$ $(0 \leq i \leq r)$. So we have $n^r \leq m$, whence $r \leq \frac{\log m}{\log n}$ and $|a_i| \leq a_i|1| = a_i \leq n$. Therefore we compute

$$|m| = |\sum_{i=0}^{r} a_i n^i| \leq \sum_{i=1}^{r} |a_i||n|^i \ . \tag{10.3.1}$$

Note that $|n| \leq 1$ implies $|m| \leq nr \leq \frac{n \log m}{\log n}$. Applying this to m replaced by m^k and taking k^{th} root we obtain $|m| \leq \sqrt[k]{\frac{kn \log m}{\log n}}$. Letting $k \to \infty$ we get an upper bound for $|m|$ independent of m which contradicts to the assumption $|\cdot|$ being archimedean. So we have $|n| > 1$, and using (10.3.1) we compute

$$|m| \leq \sum_{i=0}^{r} |a_i||n|^i \leq |n|^r \sum_{i=0}^{r} |a_i| \leq |n|^r n(r+1) \leq |n|^{\log m / \log n} n(1 + \frac{\log m}{\log n}) \ .$$

Substituting m^k into m, taking k^{th} root, and letting $k \to \infty$ we obtain $|m| \leq |n|^{\log m / \log n}$. The statement follows by interchanging m and n. □

Put $s := \frac{\log |n|}{\log n}$ for some fixed $1 < n \in \mathbb{Z}$. By the above Lemma $0 < s$ and s does not depend on the choice of n. So we have $|m| = e^{s \log m} = m^s = |m|_\infty^s$ for all $1 < m \in \mathbb{Z}$. The statement follows for all nonnegative rational numbers by taking quotients and by Exercise 10.3.1 for negative rationals. □

Definition 10.3.10. *The field K is said to be* complete *with respect to the absolute value $|\cdot|$ if any Cauchy sequence is convergent.*

Example 10.3.11. *Both $\mathbb{R}$ and $\mathbb{C}$ are complete with respect to $|\cdot|_\infty$, but $\mathbb{Q}$ is only complete with respect to the trivial absolute value.*

In the following we are going to show that any field K with an absolute value $|\cdot|$ can be embedded isometrically as a subfield into a complete field. We define

$$R := \{(a_n)_n \in K^{\mathbb{N}} \colon \forall \varepsilon > 0 \exists N \in \mathbb{N} \text{ s. t. } |a_n - a_m| < \varepsilon \text{ for all } m, n \geq N\}$$

as the ring of Cauchy sequences in K. This is indeed a ring with respect to the pointwise addition and multiplication. Note that K can be embedded into R diagonally, ie. we have a ring homomorphism $\iota \colon K \hookrightarrow R$ defined by $\iota(c) := (c)_n$. Let $I_0 \subset R$ be the set of those sequences that are identically 0 except for finitely many terms. This set is an ideal in R. Let $R_0 := R/I_0$ the quotient. We may think of R_0 as the ring of equivalence classes of Cauchy sequences with respect to the equivalence relation $(a_n)_n \sim (b_n)_n$ if $a_n = b_n$ for all but finitely many $n \in \mathbb{N}$.

Proposition 10.3.12. *R_0 is a local ring. Its unique maximal ideal consists of the those Cauchy sequences that converge to 0 ("zero sequences").*

Remark. In case you just heard this expression for the first time a commutative ring R is said to be a *local ring* if it has a unique maximal ideal.

Proof. Let $M \subset R$ be the set of zero sequences. It is clear that $I_0 \subset M$ and M is an ideal in R. On the other hand, if $(a_n)_n$ is a Cauchy sequence with $a_n \not\to 0$ then $1/a_n$ makes sense if n is large enough and is also a Cauchy sequence. This shows that the equivalence class of $(a_n)_n$ is invertible in R_0. Therefore M is indeed the unique maximal ideal of R containing I_0, or equivalently, M/I_0 is the unique maximal ideal in R_0 by Exercise 10.3.4. □

Definition 10.3.13. *Let K be a field with an absolute value $|\cdot|$. We define $\hat{K} := R/M$ to be the* completion *of K wrt. $|\cdot|$. Note that this is indeed a field as M is a maximal ideal in R.*

Note that the composite map $K \overset{\iota}{\hookrightarrow} R \to \hat{K} = R/M$ is still injective: if $0 \neq c \in K$ the the constant c sequence does not tend to 0 hence does not lie in M. So from now on we identify K with its image in $\hat{K}$. We still need to verify that $\hat{K}$ is indeed complete in order to justify the term "completion". (In fact we also need the universal property of $\hat{K}$ for being the completion, ie. any ismoetric field homomorphism $\varphi\colon K \to F$ into a valued field F factors through $\hat{K}$.) For this we first need to extend $|\cdot|$ from K to $\hat{K}$. Since the topology on K is defined so that the map $|\cdot|\colon K \to \mathbb{R}$ is continuous, it takes any Cauchy sequence in K to a Cauchy sequence in $\mathbb{R}$. As $\mathbb{R}$ is complete, we may define $|(a_n)_n|$ as a limit $\lim_{n\to\infty} |a_n|$. So we obtain a valuation R and M is the set of elements with valuation 0 by definition. Therefore the absolute of $(a_n)_n \in R$ only depends on its class in $R/M = \hat{K}$. This way we obtain an absolute value on $\hat{K}$ which we still denote by $|\cdot|$. We leave the proof of the fact that $\hat{K}$ is indeed complete and has the required universal property to the reader as an exercise (see Exercises 10.3.6 and 10.3.7).

Definition 10.3.14. *The field $\mathbb{Q}_p$ of p-adic numbers is the completion of $\mathbb{Q}$ with respect to the p-adic absolute value $|\cdot|_p$.*

10.3.1 Exercises

Exercise 10.3.1. Show that if $|\cdot|$ is any absolute value on the field K then we have $|1| = 1$ and $|-x| = |x|$.

Exercise 10.3.2. Show that in an ultrametric space all triangles are isosceles.

Exercise 10.3.3. Show that the absolute value $|\cdot|_p$ on $\mathbb{Q}$ satisfies the axioms $(1) - (3)$.

Problem 10.3.4. Show that a commutative ring R is local if and only if it contains an ideal $I \lhd R$ such that all the elements in $R \setminus I$ are invertible in R.

Exercise 10.3.5. Verify the axioms $(1) - (3)$ for the absolute value $|\cdot|$ on $\hat{K}$ if $\hat{K}$ is the completion of a valued field $(K, |\cdot|)$.

Problem 10.3.6. The field $\hat{K}$ is complete wrt. $|\cdot|$. *Hint:* We need to show that any Cauchy sequence of Cauchy sequences converges to a Cauchy sequence. You can construct the limit sequence as taking the $n_i{}^{\text{th}}$ term of the i^{th} sequence for n_i large enough (depending on i and the actual sequnce). It is a usual elementary argument in first year analysis how to choose these n_i.

Exercise 10.3.7. Verify the universal property of $\hat{K}$, ie. all $\varphi\colon K \to F$ isometric field embeddings factor through $\hat{K}$ uniquely. Also show that K is dense in $\hat{K}$. *Hint:* Take a complete field F with respect to the absolute value $|\cdot|$ and an isometric embedding $\varphi\colon K \to F$ of K as subfield of F. Extend φ to R as $\tilde{\varphi}((a_n)_n) := \lim_n \varphi(a_n)$. Since $(a_n)_n$ is a Cauchy sequence and F is complete, this makes sense. The kernel of $\tilde{\varphi}$ is exactly M, in particular it factors through $\hat{K} = R/M$.

Problem 10.3.8. Show that the field $\mathbb{Q}_p$ of p-adic numbers constructed in the previous section is indeed the completion of $\mathbb{Q}$ wrt. the absolute value $|\cdot|_p$.

Exercise 10.3.9. Let $(K, |\cdot|)$ be a—not necessarily complete—non-archimedean valued field. We define $\mathcal{O}_K := \{a \in K \colon |a| \leq 1\}$ to be the *ring of integers* in K. Show that this is indeed a subring, moreover, a local ring with maximal ideal $\mathcal{M}_K := \{a \in K \colon |a| < 1\}$. The field $\mathcal{O}_K/\mathcal{M}_K$ is called the residue class field of K. *Hint:* Use Exercise 10.3.4 and note that the elements with absolute value 1 are invertible in $\mathcal{O}_K$.

Exercise 10.3.10. Show that the ring of integers in $\mathbb{Q}_p$ is $\mathbb{Z}_p$.

Exercise 10.3.11. Show that the image of $|\cdot|_p \colon \mathbb{Q}_p \to \mathbb{R}$ is the same as the image of its restriction to $\mathbb{Q}$, namely $\{0\} \cup p^{\mathbb{Z}} \subset \mathbb{R}$.

Problem 10.3.12. What is the region of convergence of the Taylor series of $\log(1+x)$ and $\exp(x)$ at 0 in $\mathbb{Q}_p$?

10.4 Towards irreducibility criteria for polynomials over $\mathbb{Q}$

Exercise 10.4.1. *a*) Show that the polynomial $x^5 - 2x^2 + 6x - 10 \in \mathbb{Q}[x]$ is irreducible.

b) Show that the polynomial $x^2 + 1 \in \mathbb{Q}[x]$ is irreducible.

We note that the above polynomial in *a*) satisfies Eisenstein's criterion for $p = 2$ as 2 divides all the coefficients except for the leading term and 4 does not divide the constant term. In fact, in this proof we only used the prime 2 so the same proof works over $\mathbb{Q}_2$, as well. On the other hand, the polynomial in *b*) is irreducible even over $\mathbb{R}$, so, in particular, it is irreducible over $\mathbb{Q}$. What is the common in these examples?

In fact, Eisenstein's criterion is really a statement over $\mathbb{Q}_p$, not over $\mathbb{Q}$. Whenever a polynomial in $\mathbb{Q}[x]$ is irreducible over some $\mathbb{Q}_p$ or over $\mathbb{R}$ we may deduce its irreducibility over $\mathbb{Q}$, so the method is basically the same in the two examples, but we used different completions.

Over $\mathbb{R}$ it is easy to describe all the irreducible polynomials. These are the linear polynomials, and those quadratics that do not have a root in $\mathbb{R}$. What about $\mathbb{Q}_p$? Can we describe all the irreducible polynomials? The answer is yes, and we need Newton polygons for that. This will provide us with new irreducibility criteria—similar to Eisenstein's—over $\mathbb{Q}$. However, our job is a little bit harder than over $\mathbb{R}$, as the algebraic closure of $\mathbb{Q}_p$ is not a quadratic extension of $\mathbb{Q}_p$, not even a finite extension.

Exercise 10.4.2. Show that the polynomial $x^5 - 2x^4 + 4 \in \mathbb{Q}[x]$ is irreducible.

"Solution". This is not an Eisenstein polynomial for $p = 2$ (nor for any other prime) as 4 divides the constant term. What next? The idea is to have a look at the 2-adic absolute values of the roots of this polynomial. Assume we decompose this polynomial $x^5 - 2x^4 + 4 = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3)(x - \alpha_4)(x - \alpha_5)$ over a larger field $\mathbb{Q} < \mathbb{Q}_2 \leq K$ and put $c_i := -\log_2 |\alpha_i|_2 \in \mathbb{R}$ $(1 \leq i \leq 5)$. Then we have $\prod_{i=1}^5 \alpha_i = -4$ hence $\sum_{i=1}^5 c_i = -\log_2 |-4|_2 = 2$. Moreover we compute $|\alpha_i^5|_2 = \frac{1}{2^{5c_i}}$ and $|2\alpha_i^4| = \frac{1}{2^{4c_i+1}}$. Since in the ultrametic world all triangles are isosceles, we have

$5c_i = 4c_i + 1$ or $\min(5c_i, 4c_i + 1) = 2$ by the ultrametric inequality. Note that $5c_i \geq 4c_i + 1$ is impossible as otherwise $\alpha_i^5 - 2\alpha_i^4 = -4$ would be divisible by 2^5 which is nonsense. Therefore we have $c_i = 2/5$ for all $1 \leq i \leq 5$. Now assume that $x^5 - 2x^4 + 4 = g(x)h(x)$ with monic nonconstant $g, h \in \mathbb{Q}_p[x]$. Then $g(x)$ is a product of some $x - \alpha_i$ $(1 \leq i \leq 5)$. Therefore $g(0)$ is a product of some of the α_i (upto sign) therefore its 2-adic absolute value is $|g(0)| = |\alpha_i|^{\deg g} = 2^{2 \deg g/5}$. However, $g(0) \in \mathbb{Q}_p$, so its absolute value is an integer power of 2. So $5 \mid \deg g$ gives us a contradiction as neither g nor h is constant. Therefore $x^5 - 2x^4 + 4$ is irreducible over $\mathbb{Q}_p$ hence also over $\mathbb{Q}$. □

The problem with the above solution is that we have not quite defined the p-adic absolute value of an element of an extension of $\mathbb{Q}_p$. Let alone showing it satisfies the ultrametric inequality. So we are going to do this in the sequel in a precise way.

10.4.1 Hensel's Lemma

There are various forms of Hensel's Lemma. We are going to prove the version that is needed for extending absolute values from K to a finite field extension as this is needed for Newton polygons. It is in some sense the precise generalization of our observations concerning the solutions of $x^2 = 2$ in $\mathbb{Q}_7$. Let K be a complete non-archimedean field with respect to the valuation $|\cdot|$, denote by $\mathcal{O} = \mathcal{O}_K = \{x \in K \mid |x| \leq 1\}$ its ring of integers, by $\mathfrak{p} = \{x \in K \mid |x| < 1\}$ its maximal ideal, and by $k = \mathcal{O}/\mathfrak{p}$ its residue field. We say that a polynomial $f(x) \in \mathcal{O}[x]$ is *primitive* if $f(x) = a_0 + a_1 x + \cdots + a_n x^n$ with $|f| := \max_{0 \leq i \leq n}(|a_i|) = 1$.

Theorem 10.4.1 (Hensel's Lemma)**.** *Let $f(x) \in \mathcal{O}[x]$ be a primitive polynomial and suppose that $\overline{f}(x) := f(x) \pmod{\mathfrak{p}} \in k[x]$ can be written as $\overline{f}(x) = \overline{g}(x)\overline{h}(x)$ with $(\overline{g}(x), \overline{h}(x)) = 1$ in $k[x]$. Then there exist primitive polynomials $g(x), h(x) \in \mathcal{O}[x]$ such that $f(x) = g(x)h(x)$, $\overline{g}(x) = g(x) \pmod{\mathfrak{p}}$, $\overline{h}(x) = h(x) \pmod{\mathfrak{p}}$, and $\deg g = \deg \overline{g}$.*

Proof. Put $d := \deg f$, $m := \deg \overline{g}$. Then we have $d - m \geq \deg \overline{h}$. (Note that we only have $\deg \overline{f} \leq d$ as some of the coefficients in $f(x)$ might reduce to zero modulo $\mathfrak{p}$.) At first we lift $\overline{g}$ and $\overline{h}$ arbitrarily by choosing $g_0, h_0 \in \mathcal{O}[x]$ such that

$$\overline{g} = g_0 \pmod{\mathfrak{p}}, \quad \overline{h} = h_0 \pmod{\mathfrak{p}}$$

and $\deg g_0 = \deg \overline{g}$, $\deg h_0 = \deg \overline{h} \leq d - m$. Since we have $(\overline{g}, \overline{h}) = 1$, there exist polynomials $a(x), b(x) \in \mathcal{O}[x]$ with $a(x)g_0(x) + b(x)h_0(x) \equiv 1 \pmod{\mathfrak{p}}$. So all the coefficients of both $f(x) - g_0(x)h_0(x)$ and $a(x)g_0(x) + b(x)h_0(x) - 1$ are in the maximal ideal $\mathfrak{p}$. Let π be the coefficient with biggest absolute value in these polynomials (in particular, we have $|\pi| < 1$). We are going to construct g and h in the form

$$\begin{aligned} g(x) &= g_0(x) + \pi p_1(x) + \cdots + \pi^n p_n(x) + \ldots \\ h(x) &= h_0(x) + \pi q_1(x) + \cdots + \pi^n q_n(x) + \ldots \end{aligned}$$

such that $\deg p_i < m$ and $\deg q_i \leq d - m$. We construct these polynomials inductively. Let $n \geq 1$ and assume we have constructed

$$\begin{aligned} g_{n-1}(x) &= g_0(x) + \pi p_1(x) + \cdots + \pi^{n-1} p_{n-1}(x) \\ h_{n-1}(x) &= h_0(x) + \pi q_1(x) + \cdots + \pi^{n-1} q_{n-1}(x) \end{aligned}$$

such that $|f - g_{n-1}h_{n-1}| \leq |\pi|^n$. Put $f_n(x) := \frac{f(x)-g_{n-1}(x)h_{n-1}(x)}{\pi^n} \in \mathcal{O}[x]$. We define $p_n(x)$ to be the residue in the Euclidean division of $b(x)f_{n-1}(x)$ by $g_0(x)$, ie. we have $b(x)f_{n-1}(x) = Q_n(x)g_0(x) + p_n(x)$ with some $Q_n \in \mathcal{O}[x]$ and $\deg p_n < \deg g_0 = m$. Note that one can indeed take the euclidean division as the leading coefficient of $g_0(x)$ does not lie in $\mathfrak{p}$ hence it is invertible in $\mathcal{O}$. Now we define $q_n(x) \in \mathcal{O}[x]$ to be the polynomial we obtain by ommiting all the nonzero coefficients of $h_0(x)Q_n(x)+a(x)f_n(x)$ with valuation $\leq |\pi|$ so that we have $|q_n - h_0Q_n - af_n| \leq |\pi|$. On the other hand, we have

$$h_0p_n + g_0(h_0Q_n + af_n) = (h_0b + g_0a)f_n \equiv f_n \pmod{\pi},$$

so $\deg q_n \leq \deg f_n - \deg g_0 \leq d - m$ as we clearly have $\deg(h_0p_n) \leq d$. Moreover, if we put $g_n = g_{n-1} + \pi^n p_n$ and $h_n = h_{n-1} + \pi^n h_n$ then we compute

$$\begin{aligned} f - g_nh_n = f - g_{n-1}h_{n-1} - \pi^n(g_{n-1}q_n + h_{n-1}p_n) - \pi^{2n}p_nq_n \equiv \\ \equiv \pi^n(f_n - g_{n-1}q_n - h_{n-1}bf_n + h_{n-1}Q_ng_0) \equiv \\ \equiv \pi^n(f_n - q_0h_0Q_n - g_0af_n - h_0bf_n + h_0Q_ng_0) \equiv 0 \pmod{\pi^{n+1}} \end{aligned}$$

as we have $g_{n-1} \equiv g_0 \pmod{\pi}$ and $h_{n-1} \equiv h_0 \pmod{\pi}$. The result follows noting that the sums $g(x) = g_0(x) + \sum_{i=1}^{\infty} \pi^i p_i(x)$ and $h(x) = h_0(x) + \sum_{i=1}^{\infty} \pi^i q_i(x)$ both converge to *polynomials* by the bounds on the degree. For these polynomials we have $f(x) = g(x)h(x)$. □

Corollary 10.4.2. *If* $f(x) = a_0 + a_1x + \cdots + a_nx^n \in K[x]$ *is* irreducible *then we have* $|f| = \max(|a_0|, |a_n|)$.

Proof. We prove by contradiction and may assume without loss of generality that $|f| = 1$ (ie. $f(x) \in \mathcal{O}[x]$ primitive). Let $0 < r < n$ be the smallest index such that $|a_i| = 1$. Then $f(x)$ decomposes as $x^r(a_r + \cdots + a_nx^{n-r}) \equiv f(x) \pmod{\mathfrak{p}}$. We obtain a contradiction using Hensel's Lemma. □

10.4.2 Extending valuations

Let K be a complete nonarchimedean field as above. Our goal in this section is to prove the following

Theorem 10.4.3. *Let* L/K *be a finite field extension. Then the valuation* $|\cdot|$ *extends uniquely to an ultrametric valuation on* L. *The extension is given by* $|\alpha| = \sqrt[n]{|N_{L/K}(\alpha)|}$ *for* $\alpha \in L$ *where* $n = |L : K|$ *the degree and* $N_{L/K}(\alpha)$ *is the* norm *of* α, *ie. the determinant of the multiplication by* α *as a* K*-linear map* $L \to L$.

Remark. Note that in case of the archimedean field $\mathbb{R}$ the extension of $|\cdot|_\infty$ to $\mathbb{C}$ is indeed given by $|\alpha|_\infty = \sqrt{|N_{\mathbb{C}/\mathbb{R}}(\alpha)|_\infty} = \sqrt{|\alpha \cdot \overline{\alpha}|}$.

Proof. Let us show the uniqueness first assuming that $\sqrt[n]{|N_{L/K}(\cdot)|}$ is a nonarchimedean absolute value. Suppose we have another extension $|\cdot|'$ to L. Denote by $\mathcal{O}_L = \{\alpha \in L \mid |\alpha| \leq 1\}$ and by $\mathcal{O}'_L = \{\alpha \in L \mid |\alpha|' \leq 1\}$ the rings of integers with respect to the two absolute values and by $\mathfrak{p}_L = \{\alpha \in L \mid |\alpha| < 1\}$ and by $\mathfrak{p}'_L = \{\alpha \in L \mid |\alpha|' < 1\}$ the maximal ideals. Assume that α lies in $\mathcal{O}_L \setminus \mathcal{O}'_L$ and let $f(x) = x^d + a_{d-1}x^{d-1} + \cdots + a_0$ be α's minimal polynomial. Note that the

norm $N_{L/K}(\alpha)$ is a power of a_0 (upto sign). Since $\alpha \in \mathcal{O}_L$, we have $|N_{L/K}(\alpha)| \leq 1$ therefore we also have $|a_0| \leq 1$. By Corollary 10.4.2 we deduce that a_i is in $\mathcal{O}_K$ for all $0 \leq i \leq d-1$. On the other hand, $\alpha \notin \mathcal{O}'_L$ whence $|\alpha|' > 1$ and $|1/\alpha|' < 1$. This means that $1 = |1|' = |-a_{d-1}\ alpha^{-1} - \cdots - a_0\alpha^{-d}|' < 1$ by the ultrametric inequality. This is a condradiction, so we obtain $\mathcal{O}_L \subseteq \mathcal{O}'_L$. Moreover, $\mathfrak{p}'_L \cap \mathcal{O}_L$ is a prime ideal in $\mathcal{O}_L$ therefore it equals $\mathfrak{p}_L$. Hence we have $\mathfrak{p}_L \subseteq \mathfrak{p}'_L$. All in all we obtain $|\alpha| \leq 1 \Rightarrow |\alpha|' \leq 1$ (by $\mathcal{O}_L \subseteq \mathcal{O}'_L$) and also $|\alpha| > 1 \Rightarrow |1/\alpha| < 1 \Rightarrow |1/\alpha|' < 1 \Rightarrow |\alpha|' > 1$ (by $\mathfrak{p}_L \subseteq \mathfrak{p}'_L$) showing that $|\cdot|$ and $|\cdot|'$ are equivalent.

So it remains to show that $\alpha \mapsto |\alpha| = \sqrt[n]{|N_{L/K}(\alpha)|}$ is indeed a nonarchimedean valuation on L. Axioms (1) and (2) are obviously satisfied, so we only need to check (3′). Choose $\alpha, \beta \in L$ and assume (as we may) that $|\beta| \leq |\alpha| \leq 1$. So the statement of (3′) means that we also have $|\alpha + \beta| \leq 1$, in other words we are reduced to proving that $\mathcal{O}_L = \{\alpha \in L \mid N_{L/K}(\alpha) \in \mathcal{O}_K\}$ is a subring (in particular, closed under addition) in L. By Corollary 10.4.2 $\mathcal{O}_L$ is exactly the set of those elements in L whose monic minimal polynomial has coefficients in $\mathcal{O}_K$, ie. the *integral closure* of $\mathcal{O}_K$ in L which is known to be a subring (this is the way one proves that the algebraic integers form a ring). Since the proof is simple, we include it here:

Lemma 10.4.4. *Let B be an integral domain, $A \leq B$ be a subdomain, and $b_1, \ldots, b_k \in B$ arbitrary. The elements b_i ($1 \leq i \leq k$) all have monic minimal polynomials over A (ie. they are* integral *over A) if and only if the subring $A[b_1, \ldots, b_k] \leq B$ generated by $b_1, \ldots, b_k$ over A is finitely generated as a module over A.*

Remark. Note that being finitely generated as a subring is much weaker than being finitely generated as a module over A. In the former we may multiply the generators together but in the latter we can only multiply the generators by constants in A.

The proof of the Lemma: $\Rightarrow$: Induction on k. The case $k = 0$ is trivial. Now by induction, the ring $R = A[b_1, \ldots, b_{k-1}]$ is finitely generated as a module over A, say by the generators $x_1, \ldots, x_t$. We are going to show that the set $\{x_j b_k^i \mid 1 \leq j \leq t, 0 \leq i \leq d-1\}$ generate $A[b_1, \ldots, b_k]$ as a module over A where d denotes the degree of the minimal polynomial $f(x) = x^d + a_{d-1}x^{d-1} + \cdots + a_0 \in A[x]$ of b_k over A. Indeed, any element in $A[b_1, \ldots, b_k] = R[b_k]$ can be written as a polynomial in b_k with coefficients in R. The coefficients can be written as an A-linear combination of $x_1, \ldots, x_t$ and the polynomial can be reduced to having degree $< d$ by euclidean division by f as $f(b_k) = 0$ and f is monic.

$\Leftarrow$: Suppose that $A[b_1, \ldots, b_k]$ is finitely generated as a module over A, say by generators $x_1, \ldots x_t$. Then we may take the matrix $M_i \in A^{t \times t}$ of the multiplication by b_i in the basis $x_1, \ldots, x_t$. Note that the matrix M_i is not unique as we may have "relations" between the x_j's. However, it certainly exists since the $x_1, \ldots, x_t$ form a generating system. By the theorem of Cayley and Hamilton, b_i is the root of its own characteristic polynomial which is monic and has coefficients in A. Therefore the minimal polynomial of b_i over A exists and is also monic as it divides the characteristic polynomial. $\square$

10.4.3 Newton polygons

Let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0 \in \mathbb{Q}_p[x]$ be a polynomial. The Newton polygon of f is the (boundary of the) lower convex hull of the points

$$\{(-n, -\log_p |a_n|), \ldots, (-i, -\log_p |a_i|), \ldots, (0, -\log_p |a_0|)\} \subset \mathbb{Z}^2 \subset \mathbb{R}^2$$

on the euclidean plane. That is, take the intersection of all the closed half-planes containing these points and lying above some nonvertical line. We say that the multiplicity of the slope $a/b \in \mathbb{Q}$ is m if we have a segment in the Newton polygon with slope a/b and horizontal width m. The polynomial f has exactly n slopes if counted with multiplicities.

Example 10.4.5. *The Newton polygon of the polynomial* $x^3 + px^2 + px + p^3$ *has vertices* $(-3, 0), (-1, 1)$, *and* $(0, 3)$. *It has slopes* $1/2$ *with multiplicity* 2 *and* 2 *with multiplicity* 1.

The *additive valuation* of $\alpha \in \overline{\mathbb{Q}_p}$ is by definition $-\log_p |\alpha|_p$. Note that α belongs to a finite extension K of $\mathbb{Q}_p$ and we extended $|\cdot|_p$ to K in the previous section.

Theorem 10.4.6. *The multiset of slopes of the Newton polygon of* f *equals the multiset of the additive valuations of the roots of* f *in* $\overline{\mathbb{Q}_p}$.

Proof. We introduce the ρ-norm (Gauss-norm) on $\mathbb{Q}_p[x]$ for each real number $\rho > 0$ by putting $\|a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0\|_\rho := \max_{1 \le i \le n}(|a_i|_p \rho^i)$. The *width* of f under the ρ-norm is the difference between the maximum and minimum values of i for which $\max_i(|a_i|_p \rho^i)$ is achieved. Note that the multiplicity of the slope a/b in the Newton polygon of f is nothing else but the width of f under the ρ-norm with $\rho = p^{-a/b}$. The statement follows from the following

Lemma 10.4.7. *For* $f(x), g(x) \in \mathbb{Q}_p[x]$ *and* $\rho > 0$ *we have* $\|fg\|_\rho = \|f\|_\rho \|g\|_\rho$ *(ie.* $\|\cdot\|_\rho$ *is multiplicative). Moreover, the width of* fg *under the* ρ*-norm equals the sum of the widths of* f *and* g.

Proof. Denote by m_f and M_f the minimum and maximum values of i for which $\max_i(|a_i|_p \rho^i)$ is achieved. The integers m_g, m_{fg}, M_g, and M_{fg} are defined similarly. If we write $g(x) = b_k x^k + \cdots + b_0$ then we have

$$f(x)g(x) = \sum_i \left(\sum_{j+l=i} a_j b_l \right) x^i \ .$$

In the sum $\sum_{j+l=i} a_j b_l$ each summand has absolute value at most $\|f\|_\rho \|g\|_\rho \rho^{-i}$ with equality if and only if $|a_j| = \|f\|_\rho \rho^{-j}$ and $|b_l| = \|f\|_\rho \rho^{-l}$. This cannot occur for $i < m_f + m_g$ and for $i = m_f + m_g$ it occurs only for $j = m_f$ and $l = m_g$. So we have $m_{fg} = m_f + m_g$ and the multiplicativity of $\|\cdot\|_\rho$ also follows. The equality $M_{fg} = M_f + M_g$ is deduced the same way. Therefore the width is indeed additive. □

Corollary 10.4.8. *If the Newton polygon of a polynomial* $f(x) \in \mathbb{Q}[x]$ *wrt. some prime* p *(ie. considered as a polynomial in* $\mathbb{Q}_p[x]$*) is just one line with the only lattice points at the two ends then* $f(x)$ *is irreducible.*

Newton polygons have many more modern applications, too, e.g. in the theory of p-adic differential equations. To read more have a look at [6].

10.4.4 Exercises

Exercise 10.4.3. Show that all the $p-1^{\text{st}}$ roots of unity are contained in $\mathbb{Q}_p$. *Hint:* Try and factor the polynomial $x^{p-1}-1$ using Hensel's Lemma.

Problem 10.4.4. Compute $|1-\varepsilon_m|_p$ for any positive integer m and prime p where ε_m is a primitive m^{th} root of unity. *Hint:* At first do it if $(m,p)=1$ or m is a power of p. For this compute the Newton polygon of a suitable polynomial having $1-\varepsilon_m$ as a root. Finally, write $\varepsilon_m=\varepsilon_{p^h}\varepsilon_j$ where $(j,p)=1$ and $1-\varepsilon_m=(1-\varepsilon_{p^h})+\varepsilon_{p^h}(1-\varepsilon_j)$.

Exercise 10.4.5. Give more details of the proof of Theorem 10.4.6. Verify that the width of f under the ρ-norm is equal to the multplicity of the slope $-\log_p\rho$ in the Newton polygon of f. What is the Newton polygon of the linear polynomial $x-\alpha$?

Exercise 10.4.6. Give a proof of Corollary 10.4.8.

Problem 10.4.7. Show that if Newton polygon of the polynomial $f(x)\in\mathbb{Q}_p$ has two different slopes then f cannot be irreducible. *Hint:* Use the uniqueness of the extension of the absolute value to finite extensions of $\mathbb{Q}_p$ in order to show that the Galois group $\mathrm{Gal}(K/\mathbb{Q}_p)$ acts on any Galois-extension K via isometries. If all the roots of f are Galois-conjugates then they have the same absolute value.

10.5 Applications, research directions, and further reading

If you do not intend to become a number theorist, you may ask why learn the p-adics as they are so different from the "real world". This is, in fact, not quite true. However, let us discuss the most important applications of p-adic methods in Number Theory first, together with a view what the main research directions are. The list below does not intend to be exhaustive—it certainly reflects the interest and the (limited) knowledge of the author.

10.5.1 Hasse's local-global principle

The most important application of the p-adics numbers are through the so-called *local-global* (or Hasse) principle. Roughly speaking the idea is that—as you may have noticed—it is easier to decide whether or not polynomial equations have roots in the fields $\mathbb{R}$ and $\mathbb{Q}_p$ for varying p than deciding it over $\mathbb{Q}$. Clearly, if there are no roots in $\mathbb{Q}_p$ for some p or in $\mathbb{R}$ then there can be no roots in $\mathbb{Q}$ either. The question is up to what extent is the converse true. Unfortunately, this is not always the case. For example, the equation $3x^3+4y^3+5z^3=0$ has a solution in $\mathbb{R}$ and $\mathbb{Q}_p$ for *all* primes p, but not in $\mathbb{Q}$. However, for homogeneous polynomials of degree 2 the local global principle holds. This is the theorem of Hasse and Minkowski (for a detailed and elementary proof see the book [12] by Serre).

There exist certain methods how to "measure" the failure of the Hasse principle. For elliptic curves this is done by the Tate-Shafarevich group which in this case fully accounts for the failure of the principle. The Tate-Shafarevich conjecture

asserts that this group is always finite over for elliptic curves over finite extensions of $\mathbb{Q}$. This is one of the most important open problems in arithmetic geometry. The conjecture of Birch and Swinnerton-Dyer (a millenium prize problem) would imply this and the conjecture has been tested for many numerical examples. There is also some very important theoretical evidence in favour of this conjecture—for instance, the known cases of the BSD conjecture. To read more about elliptic curves Silverman's book [13] is an excellent introduction.

10.5.2 Langlands programme

L-functions play a very important role in Number Theory. These are certain generalizations of the Riemann ζ-function $\zeta(s) = \sum_{n=1}^{\infty} 1/n^s$ ($\mathrm{Re}(s) > 1$). For example, the proof of Dirichlet's Theorem on primes in arithmetic progression is proven using L-functions. The Riemann ζ-function itself (especially the set of its roots) is very much related to the distribution of primes in $\mathbb{Z}$. Further, according to the conjecture of Birch and Swinnerton-Dyer, the L-function of an elliptic curve should vanish to the order of the rank of the curve.

L-functions play the role of the connection between Galois representations and automorphic forms. One can attach L-functions to both types of objects. However, while on the Galois-side it is very natural to write the L-function as Euler product over the primes (for example, we have $\zeta(s) = \prod_{p \text{ prime}} \frac{1}{1-p^{-s}}$), this is not so obvious on the automorphic side. On the other hand, the functional equation and the analytic continuation of L-functions to the whole complex plane—note that a priori $\zeta(s)$ is only defined if $\mathrm{Re}(s) > 1$—is quite standard (well, this is Tate's thesis, in fact), but not at all on the Galois side. In fact, the only method known to show the analytic continuation is via *modularity*, ie. showing that the L-function in question is the L-function of some automorphic form. The Langlands program is the philosophy that one should try to match Galois-representations to automorphic forms having the same L-function. There are not too many known results in this direction. The case when the Galois representation is 1-dimensional, is completely understood via class field theory. The case of Galois-representations coming from elliptic curves was settled by Wiles (and Taylor) when proving Fermat's Last Theorem. More recently, there are other modularity results using Serre's conjectures and the p-adic Langlands correspondence for $\mathrm{GL}_2(\mathbb{Q}_p)$ by Colmez. So one can—rather surprisingly—use p-adic methods to prove the analytic continuation of certain complex functions!

If you are interested in this, you should start out by reading class field theory first for which I recommend the books [11] and [9].

10.5.3 Algebraic geometry

It should be obvious by now that the p-adic numbers are useful when trying to find (or proving that there are no) rational points on algebraic varieties. However, there are several other applications of the p-adics in algebraic geometry. For instance, it is sometimes useful to complete the local ring of a variety at a point, as complete discrete valuation rings have better properties than those that are not complete. Another very important application is in étale cohomology. The étale cohomology is a cohomology theory in algebraic geometry that has better

properties if the coefficients are taken from a finite ring. However, for certain applications, it is necessary to have coefficients with characteristic zero. Therefore one takes the projective limit with coefficients in $\mathbb{Z}/p^n\mathbb{Z}$ to obtain coefficients in $\mathbb{Z}_p$. If you want to learn more on algebraic geometry the best reference is [5].

10.5.4 Group theory

Profinite groups are inverse limits of finite groups. They are naturally compact topological spaces in the inverse limit topology of the finite sets equipped with the discrete topology. For example infinite Galois groups are profinite, but profinite groups also show up as automorphism groups of certain (infinite) rooted trees. The additive group $\mathbb{Z}_p \cong \varprojlim \mathbb{Z}/p^n\mathbb{Z}$ is a profinite group, moreover, it is a *pro-p* group, ie. an inverse limit of finite p-groups. Moreover, it is the unique (upto isomorphism) infinite pro-p group topologically generated by a single element. A pro-p group G is said to have finite rank if all its closed subgroups can be topologically generated by a bounded number of elements. All pro-p groups of finite rank are closed subgroups of $\mathrm{GL}_n(\mathbb{Z}_p)$ for n large enough. If you wish to learn more on pro-p groups, the bible is the book [3].

Another application of the p-adic numbers is in modular representation theory of finite groups. This is because the natural objects to which one can lift up representations in characteristic p to characteristic 0 are complete local integral domains, such as $\mathbb{Z}_p$. For more information on modular representation theory see the book [10].

10.5.5 Dynamical systems

The main result of the groundbreaking paper [1] is the following. We say that a complex number $a \in \mathbb{C}$ is preperiodic for the polynomial $f(z) \in \mathbb{C}[z]$ if the set

$$\{a, f(a), f(f(a)), \dots, f(\dots(f(a))\dots), \dots\}$$

is finite. Fix a positive integer $d > 1$ and complex numbers $a, b \in \mathbb{C}$. The set of parameters $c \in \mathbb{C}$ such that both a and b are preperiodic for $f(z) = z^d + c$ is infinite if and only if $a^d = b^d$. Note that the statement is completely elementary and only concerns complex polynomials. However, the proof requires non-trivial methods in non-archimedean analytic geometry (in the sense of Berkovich [2]).

10.5.6 Algebraic topology

The (still open) Hilbert-Smith conjecture states that if a locally compact group G acts effectively (ie. faithfully) on a topological manifold M then G is a Lie-group. Because of known structural results on locally compact groups the conjecture can be reduced to the case $G \cong \mathbb{Z}_p$ the additive group of the p-adic integers. In other words it would be enough to show that $\mathbb{Z}_p$ cannot act faithfully on a topological manifold M.

The ring $\mathbb{Z}_p$ of p-adic integers is one of the easiest examples of a complete discrete valuation ring (the other one being $k[\![t]\!]$, k field). These are very important in the theory of formal groups which not only show up in algebraic geometry and number theory, but also in algebraic topology. The book [8] is a good introduction to the theory of formal groups.

10.5.7 Physics

The geometry of space-time at small distances seems to be non-archimedean—at least according to some physisists. For instance, the p-adic numbers show up in quantum mechanics, quantum field theory, and string theory, too. I am not an expert on this, so if you are interested, you should consult the book [14] for a start.

10.6 References

[1] M. Baker and L. DeMarco. Preperiodic points and unlikely intersections. *Duke Mathematical Journal*, **159**(1):1–29, 2011.

[2] V. Berkovich. *Spectral theory and analytic geometry over non-Archimedean fields*, volume **33** of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1990.

[3] J. D. Dixon, M. P. F. du Sautoy, A. Mann, and D. Segal. *Analytic pro-p groups*. Cambridge University Press, Cambridge, 1999.

[4] F. Q. Gouvêa. *p-adic Numbers, An Introduction*. Springer, Heidelberg, 1997.

[5] R. Hartshorne. *Algebraic Geometry*, volume **52** of *Graduate Texts in Mathematics*. Springer, 1977.

[6] K. S. Kedlaya. *p-adic Differential Equations*, volume **125** of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2010.

[7] N. Koblitz. *p-adic Numbers, p-adic Analysis, and Zeta-functions*, volume **58** of *Graduate Texts in Mathematics*. Springer, 1984.

[8] M. Lazard. *Commutative formal groups*, volume **443** of *Lecture Notes in Mathematics*. Springer, Berlin, New York, 1975.

[9] J. Neukirch. *Class Field Theory*, volume **280** of *Grundlehren der mathematischen Wissenschaften*. Springer, Heidelberg, 1986.

[10] P. Schneider. *Modular Representation Theory of Finite Groups*. Springer, London, 2013.

[11] J.-P. Serre. *Local Fields*, volume **67** of *Graduate Texts in Mathematics*. Springer, New York, 1980.

[12] J.-P. Serre. *A course in arithmetic*. Springer, New York, 1993.

[13] J. H. Silverman. *The Arithmetic of Elliptic Curves*, volume **106** of *Graduate Texts in Mathematics*. Springer, 1986.

[14] V. S. Vladimirov. *p-adic Analysis and Mathematical Physics*. World Scientific, Singapore, 1994.

Chapter 11

András Zempléni: Extreme Value Modelling

11.1 Introduction

We hear a lot about financial crisis or climate changes, but how can one mathematically tackle the problem of estimating the severity of possible extreme losses or heat waves? Extreme values are in the focus of attention at different areas, like environmental or financial data analysis. We shall overview the classical univariate approaches of block-maxima and peaks-over-threshold and give an introduction to the more recent multivariate approaches and copula methods. Besides the probabilistic background, emphasis will be given to the most important statistical methods, like maximum likelihood estimation or confidence interval construction. Methods for assessing the goodness-of-fit for the chosen models will also be presented, including recently developed bootstrap approaches. The theory will be illu

11.2 Extreme Value Theory

11.2.1 Univariate Extreme Value Theory

First we outline the main probabilistic results providing the basis of parametric modeling of univariate extremes.

To reveal the motivation behind extreme value theory (EVT), let $X_1, ..., X_n$ be a sequence of independent random variables with common distribution function F. In addition let $M_n = \max(X_1, X_2, \ldots, X_n)$ be the maximum of the sequence. The variables X_i often represent hourly or daily values of a process and so M_n represents the maximum of the process over n time units. The distribution function of M_n can be computed in a very elementary way as

$$P(M_n \leq z) = P(X_1 \leq, ..., X_n \leq z) = \prod_{i=1}^{n} P(X_i \leq z) = F^n(z). \qquad (11.2.1)$$

However Equation 11.2.1 is not very useful in practice if F is unknown. Of course, one may suggest to estimate F from the measurements in some way and use this as a plug-in estimate in $\hat{F}^n(z)$. Unfortunately by doing this, even small differences

between F and $\hat{F}$ might be multiplied up, leading to large error in the final estimate of F^n. An alternative solution is proposed by EVT, suggesting to look for approximate distribution families for F^n directly, based on the extreme measurements only. Central limit theory for extreme values (without proofs) is provided below.

Limit for Maxima

For the maximum of univariate i.i.d. variables the theory is well-elaborated. Since analogous statements follow for the minimum as

$$\min(X_1, X_2, \ldots, X_n) = -\max(-X_1, -X_2, \ldots, -X_n),$$

we can limit our attention to the case of maximum. Let

$$z_+ = \sup\{z : F(z) < 1\}$$

denote the upper endpoint of the support of the distribution $F(x)$. It is clear that $M_n \to z_+$ a.s. as $n \to \infty$. Thus, in order to get a non-degenerate limit for M_n, we must consider normalized maxima

$$M_n^* = \frac{M_n - a_n}{b_n},$$

for some sequences of constants $\{a_n\}$ and $\{b_n\} > 0$. The Gnedenko-Fisher-Tippett theorem states that the limit distribution, if exists, is in the class of the so-called extreme value distributions (EVD).

Definition 11.2.1. The extreme value distribution with shape parameter ξ has the following distribution function.
If $\xi \neq 0$,

$$G_\xi(x) = \exp\left[-(1+\xi x)^{-1/\xi}\right]$$

for $1 + \xi x > 0$ (otherwise 0 if $\xi > 0$ and 1 if $\xi < 0$).
If $\xi = 0$,

$$G_\xi(x) = \exp\left[-e^{-x}\right].$$

The $\xi = 0$ case can also be obtained from the $\xi \neq 0$ case by letting $\xi \to 0$. The limit distribution is called *Fréchet* for $\xi > 0$, *Gumbel* or double exponential for $\xi = 0$ and *Weibull* for $\xi < 0$.

One may also define the corresponding location-scale family $G_{\xi,\mu,\sigma}$ by replacing x above by $(x-\mu)/\sigma$ for $\mu \in \mathbb{R}$ and $\sigma > 0$ and changing the support accordingly. It is straightforward to check that Gumbel, Fréchet and Weibull families can be combined into a single family as follows.

Definition 11.2.2. The generalized extreme value (GEV) distribution is defined as

$$G_{\xi,\mu,\sigma}(x) = \exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right\}, \tag{11.2.2}$$

where $1 + \xi\frac{x-\mu}{\sigma} > 0$, $\mu \in \mathbb{R}$ is called the location parameter, $\sigma > 0$ the scale parameter and $\xi \in \mathbb{R}$ the shape parameter.

Theorem 11.2.3. ***[Fisher and Tippett (1928), Gnedenko (1943)]*** *If there exist* $\{a_n\}$ *and* $\{b_n\} > 0$ *sequences such that*

$$P(M_n^* \leq z) = P\Big(\frac{M_n - a_n}{b_n} \leq z\Big) \to G(z) \text{ as } n \to \infty \tag{11.2.3}$$

where G *is a non-degenerate distribution function, then* G *necessarily belongs to the GEV family, defined in Equation 11.2.2. In this case we say that the distribution of* X_i *belongs to the max-domain of attraction of the GEV distribution* G*.*

This theorem is usually used in practical applications for modeling the maxima of observations appearing in consecutive blocks of time (*block maxima*), as e.g. annual/ monthly/ weekly maxima.

Remark 11.2.4. *>From the statistical point of view the apparent difficulty is that the normalizing constants are unknown. This can be easily solved in practice, as the distribution of the non-normalized maxima can be approximated by GEV distribution with different location and scale parameters:*

$$P(M_n \leq z) \sim G\Big(\frac{z - a_n}{b_n}\Big) = G^{\dagger}(z).$$

Limit for Threshold Exceedances

Modeling only the block maxima can be inefficient. As EVT is basically concerned with modeling the tail of an unknown distribution, a natural idea is to model all of those observations X_i, whose values are larger than a considerably high threshold. Due to the results of Balkema and de Haan (1974) it is well-known that if the distribution of X_i lies within the max-domain of attraction of a GEV distribution, then the distribution of the threshold exceedances has a similar limiting representation. The results are summarized in the following theorem.

Theorem 11.2.5. *Let* $X_1, ..., X_n$ *be a sequence of independent random variables with common distribution function* F*. Suppose that* F *belongs to the max-domain of attraction of a GEV distribution for some* ξ, μ *and* $\sigma > 0$*. Then for high thresholds* u

$$P(X_i - u \leq z | X_i > u) \to H(z) = 1 - \Big(1 + \frac{\xi z}{\tilde{\sigma}}\Big)^{-\frac{1}{\xi}} \quad \text{as } u \to z_+, \tag{11.2.4}$$

where $\tilde{\sigma} = \sigma + \xi(u - \mu)$*.*

The family defined in Equation 11.2.4 is called generalized Pareto distribution (GPD).

Remark 11.2.6. *Note, that both of the above limit results are strongly linked in the sense that, as the threshold tends to the right endpoint of the underlying distribution, the conditional distribution of the exceedances converges to GPD if and only if the distribution of the normalized maxima converges to GEV distribution. For graphical illustrations see Figure 11.2.1.*

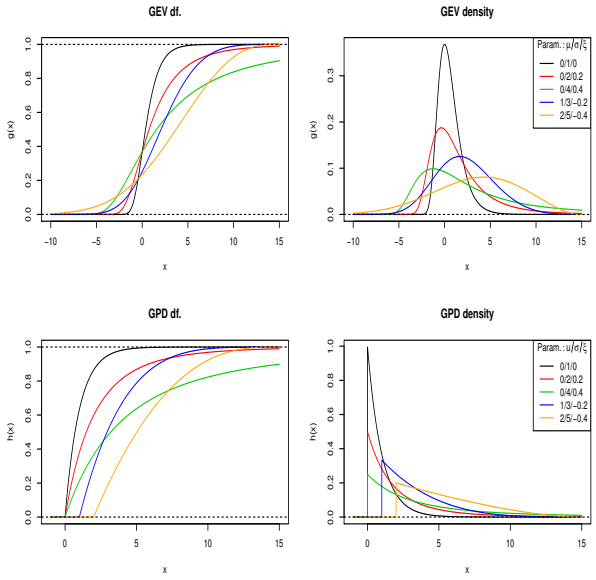


Figure 11.2.1: Distribution and density functions of GEV distribution (Equation 11.2.2) and GPD (Equation 11.2.4).

Conditions for the Limit Theorems

The EVT-based statistical procedures implicitly assume that most distributions of practical interest lie within the max-domain of attraction of a GEV distribution (or equivalently, within a GPD). Therefore, a natural question arises: how general is the class of distributions for which the above limit results hold? Although it is not difficult to find counterexamples (e.g. among discrete distributions), the most well-known continuous distributions belong to this class.

Definition 11.2.7. We say that a distribution tail $\bar{F}$ is regularly varying with index $-\alpha$ for some $\alpha \geq 0$ if for every $x > 0$

$$\lim_{t\to\infty} \frac{\bar{F}(tx)}{\bar{F}(x)} = t^{\alpha}.$$

In addition if $\alpha = 0$, the function $\bar{F}$ is said to be slowly varying.

Theorem 11.2.8 (Max-domain of attraction of the Fréchet distribution)**.** *A distribution function F belongs to the max-domain of attraction of a GEV distribution with $\xi > 0$ (Fréchet-type) if and only if the distribution tail $\bar{F}$ is regularly varying with index $-\xi$.*

This condition is satisfied by e.g. the *Pareto*, the *Cauchy* and the *stable* (for $\alpha < 2$) distributions.

Theorem 11.2.9 (Max-domain of attraction of the Weibull distribution)**.** *A distribution function F belongs to the max-domain of attraction of a GEV distribution with $\xi < 0$ (Weibull-type) if and only if the support of F is bounded to the right (with $x_+ < \infty$ is the right endpoint) and $\bar{F}(x_+ - x^{-1})$ is regularly varying with index $-\xi$.*

In contrast to the heavy-tailed distributions, the Weibull case contains distributions which have a finite right endpoint including e.g. the *uniform* and the *beta* distributions.

The $\xi = 0$ (Gumbel-type) case is more complicated. Although there exist necessary and sufficient conditions here as well, they are hardly used in practice. It can be shown that the max-domain of attraction of the Gumbel distribution covers quite a wide range of families of distribution functions. It contains distributions from heavy-tailed distributions whose all moments are finite (e.g. the *lognormal* distribution) to light-tailed distributions (e.g. the *normal*, the *exponential* or the *gamma* distribution) and even some distributions whose support is bounded to the right are possible. More details and further references can be found about the three above cases e.g. in Section 2.3,2.4 and 2.5 in Beirlant et al. (2004), respectively.

11.2.2 Modeling Multivariate Maxima

Comparing the multivariate problem with the univariate case the new issue that arises is the dependence structure. In such a case - beyond the marginal distributions - we must be able to determine how the individual variables relate to each other. The main question is describing the class of possible dependence structures and then to investigate how can we estimate them. Modeling multivariate extremes typically consists of two distinct steps: modeling univariate margins and then - after the suitable standardization of margins - modeling the dependence. As the first step involves only applying the univariate models of the previous section, here we focus on the second one, namely on characterizing the dependence structures. Order relations on vectors are understood to be component-wise, i.e. for d-dimensional vectors $\mathbf{x} = (x_1, ..., x_d)$ and $\mathbf{y} = (y_1, ..., y_d)$ the relation $\mathbf{x} \leq \mathbf{y}$ is defined as $x_j \leq y_j$ for all $j = 1, ..., d$. In this case the maximum is defined by taking the component-wise maxima, which is defined as

$$\mathbf{x} \vee \mathbf{y} = (x_1 \vee y_1, \ldots, x_d \vee y_d),$$

where $\vee$ stands for the maximum (analogously, $a \wedge b = \min(a, b)$). By using this notation the maximum of a sample of d-dimensional observations
$\mathbf{X}_i = (X_{i,1}, ..., X_{i,d})$ for $i = 1, ..., n$ is defined as

$$\mathbf{M}_n = (M_{n,1}, ..., M_{n,d}) = \Big(\bigvee_{i=1}^{n} X_{i,1}, ..., \bigvee_{i=1}^{n} X_{i,d} \Big).$$

Finally, it should be mentioned that again we can focus on maximum without loss of generality, since the following relation allows us to get the minimum by the help of the maximum of the negatives:

$$\bigwedge_{1}^{n} \mathbf{X_i} = -\bigvee_{1}^{n} (-\mathbf{X_i}).$$

Remark 11.2.10. *The sample maximum is not necessarily a sample point. From a practical point of view this means that the maxima we intend to model need not be simultaneous.*

Limit for Multivariate Maxima

Analogously to the univariate case we assume that $\mathbf{X}$ has distribution function $\mathbf{F}$ and there exist $\mathbf{a}_n$ and $\mathbf{b}_n > \mathbf{0}$ sequences of normalizing vectors, such that

$$P\Big(\frac{\mathbf{M}_n - \mathbf{a}_n}{\mathbf{b}_n} \leq \mathbf{z} \Big) = \mathbf{F}^n(\mathbf{b}_n \mathbf{z} + \mathbf{a}_n) \to \mathbf{G}(\mathbf{z}), \tag{11.2.5}$$

where the G_i margins of the limit distribution $\mathbf{G}$ are non-degenerate distributions. If Equation 11.2.5 holds then $\mathbf{F}$ is said to be in the domain of attraction of $\mathbf{G}$ and $\mathbf{G}$ itself is said to be a multivariate extreme value distribution (MEVD). Since Equation 11.2.3 holds for each margin

$$P\Big(\frac{M_{n,j} - a_{n,j}}{b_{n,j}} \leq z_j \Big) \to G_j(z_j) \text{ as } n \to \infty \tag{11.2.6}$$

for any $j = 1, ..., d$, where the d.f. G_j are non-degenerate by assumption. The margins are necessarily GEV distributions. Hence the essential part of the multivariate extension reduces to handling the dependence structure among the margins. It can be shown that the MEVD cannot be characterized as a parametric family indexed by a finite dimensional parameter vector (in contrary to in the GEV case). Instead, the family of MEVD is usually indexed by the class of the underlying dependence structures.

A useful characterization of MEVD can be given by the next definition.

Definition 11.2.11. A multivariate distribution function $\mathbf{G}$ is called max-stable, if for every positive integer k there exist α_k and $\beta_k > 0$ vectors such that

$$\mathbf{G}^k(\beta_k \mathbf{x} + \alpha_k) = \mathbf{G}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}.$$

It is not difficult to see that the classes of extreme value and max-stable distribution functions coincide (see 8.2.1 in Beirlant et al., 2004).

Exponent Measure

A further consequence of the max-stability is that $\mathbf{G}^{1/k}$ is a valid distribution function for every positive integer k. In such a case we say that the distribution function $\mathbf{G}$ is max-infinitely divisible (Balkema and Resnick, 1977). Specially, there exist a (unique) measure μ on $[\mathbf{q}, \infty) \setminus \{\mathbf{q}\}$, such that

$$\mathbf{G}(\mathbf{x}) = \exp\Big(-\mu([\mathbf{q}, \infty) \setminus [\mathbf{q}, \mathbf{x}])\Big), \tag{11.2.7}$$

where $\mathbf{q} = (q_1, \ldots, q_d)$ and $q_i = \inf\{x \in \mathbb{R} : G_i(x) > 0\}$ is the lower end-point of the ith margin. This μ measure is called the exponent measure. There are quite a few equivalent representations, a few of them will be useful for statistical models.

Positive Association

The following properties show practical cases, where the MEVD may be useful. An MEVD G is necessarily "positively quadrant dependent", namely

$$G(\mathbf{x}) \geq G_1(x_1) \ldots G_d(x_d), \quad \mathbf{x} \in \mathbb{R}^d. \tag{11.2.8}$$

In particular, a random variable Y with distribution function $\mathbf{G}$ as
in Equation 11.2.8 has $cov[f_i(Y_i), f_j(Y_j)] \geq 0$ for any $1 \leq i, j \leq d$ and any pair of non-decreasing functions f_i and f_j such that the relevant expectations exist.

Pickands' Dependence Function

In two dimensions, all information on the dependence structure is covered by another (equivalent) characterization. The joint survivor function with standard exponential margins (denoted by $\star\star$, while $\star$ is kept for unit Fréchet margins) $\bar{\mathbf{G}}_{\star\star}$ is given by

$$\bar{\mathbf{G}}_{\star\star}(z_1, z_2) = P(Z_1 > z_1, Z_2 > z_2) = \exp\left\{-(z_1 + z_2) A\left(\frac{z_2}{z_1 + z_2}\right)\right\}, \tag{11.2.9}$$

where $A(t)$, called the (Pickands) dependence function, is responsible to capture the dependence structure between the margins. It can be shown that the dependence function necessarily satisfies the following properties (P):

1. $(1-t) \vee t \leq A(t) \leq 1$ for $t \in [0,1]$ ($\Rightarrow A(0) = A(1) = 1$);
2. $A(t)$ is convex.

Remark 11.2.12. *In the first property of (P) the lower and upper bounds correspond to the following two limiting cases. If $A(t) = (1-t) \vee t$ then we get complete dependence and if $A(t) = 1$ then independence. For graphical illustration see the left panel in Figure 11.2.2.*

Of course, the representation using exponent measure function and unit Fréchet margins can be written by the dependence function as well:

$$-\log G_\star(y_1, y_2) = V_\star(y_1, y_2) = \left(\frac{1}{y_1} + \frac{1}{y_2}\right) A\left(\frac{y_1}{y_1 + y_2}\right). \tag{11.2.10}$$

Now define a W measure for any B Borel subset of the d dimensional unit simplex by

$$W(B) = \mu_\star\Big(\{\mathbf{y} \in [\mathbf{0}, \infty) : \|\mathbf{y}\|_1 \geq 1, \mathbf{y}/\|\mathbf{y}\|_1 \in B\}\Big).$$

The measure W is called spectral measure. Furthermore, there is a connection between the Pickands dependence function and this spectral measure:

$$A(t) = 1 - t + \int_0^t W([0,\omega])d\omega, \quad t \in [0,1]$$

Conversely W can be computed from A as

$$W([0,\omega]) = 1 + A'(\omega) \text{ if } \omega \in [0,1),$$

and $W([0,1]) = 2$, where A' is the derivative of A. The point masses in the endpoints are

$$W(\{0\}) = 1 + A'(0) \text{ and } W(\{1\}) = 1 - A'(1). \tag{11.2.11}$$

Remark 11.2.13. *If A' is absolutely continuous, then W is absolutely continuous on the interior of the unit interval with density $w = A''$.*

For higher dimensions Equation 11.2.9 could be generalized as

$$\bar{G}_{\star\star}(\mathbf{z}) = \exp\left\{-\left(\sum_{i=1}^d z_i\right) A\left(\frac{z_1}{\sum_{i=1}^d z_i}, ..., \frac{z_{d-1}}{\sum_{i=1}^d z_i}\right)\right\},$$

for some dependence function A, defined on the d-dimensional simplex.

11.3 Statistical inference

11.3.1 Parametric Estimation

In this section we summarize the most important results about maximum likelihood approaches. After the standard results we introduce a method known from the area of spatial statistics providing a very useful tool for estimating models in higher dimensions.

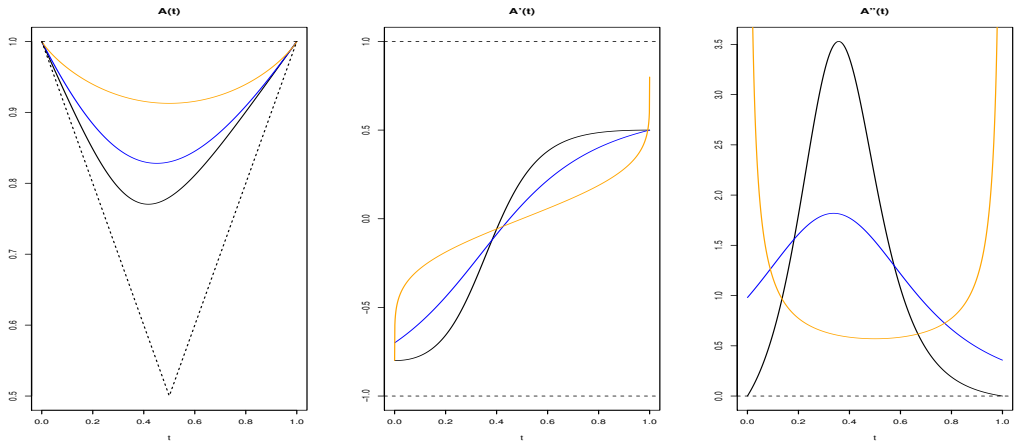


Figure 11.2.2: Differentiable dependence functions, and their derivatives.

Statistical inference

In the univariate case, maximum likelihood method can be used to construct estimation methods for the distribution of maxima or exceedances. It is pointed out in Section 6.3.1 of Embrechts et al. (1997) that there is no explicit solution to the maximum likelihood equations. However in regular cases, when $\xi > -1/2$, there are reliable numerical procedures to find the maximum likelihood estimators. These estimators are efficient, consistent and asymptotically normal. Full discussion about the properties of the estimators, including non-regular ($\xi \leq -1/2$) cases can be found in Smith (1985).

Remark 11.3.1. *For applications in insurance, finance or quite a few environmental data sets the cases of non-negative shape parameter $\xi \geq 0$ are the most relevant, as these data can rarely be supposed to be bounded to the right.*

As it was proven in Smith (1985), maximum likelihood estimators behave regularly in the multivariate case, if the above condition is fulfilled marginally, that is $\xi_i \geq -1/2$ for $i = 1, \ldots, d$. In some cases estimators for the dependence parameters can be supperefficient. For more details, see Section 3.6 in Kotz and Nadarajah (2000).

Model fit

In order to get information about the standard error of estimates, non-parametric bootstrap methods may be applied, but the standard likelihood-based confidence intervals are also possible to calculate. For goodness of fit tests (like the usual Cramer-von Mises test or the Anderson Darling test) the null distribution under estimated parameters is unkown, so we may use parametric bootstrap for critical value calculation. This means that samples are repeatedly simulated from the fitted distribution and the value of the statistics is computed for these samples.

Applications

The free statistical software package R and its add-on packages provide an excellent machinery for applying the models above. The approaches are shown in the books of Coles (2001) and Embrechts et al (1997). We shall investigate

interesting financial and meteorologial data sets, where the extremes play indeed a key role (what loss/gain or rain/extreme temperature can be expected to appear within a given time period). Everybody will have the opportunity to experiment with the data and the software in a computer lab.

11.4 References

[1] Balkema, A. A. and de Haan, L. (1974) Residual lifetime at great age. *Ann. Probab.*, **2**, p.792-804.

[2] Balkema, A. A. and Resnick, S. I. (1977) Max-infinite divisibility. *Journal of Applied Probability*, **14**, p.309-319.

[3] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004) *Statistics of Extremes: Theory and Applications*, Wiley Series in Probability and Statistcs

[4] Coles, S. (2001) *An Introduction to Statistical Modelling of Extreme Values*, Springer-Verlag.

[5] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling Extremal events for Insurance and Finance*, Springer-Verlag

[6] Kotz,S. and Nadarajah,S.(2000) *Extreme Value Distributions: Theory and Applications*, Imperial College Press

[7] Smith, R.L. (1985) Maximum likelihood estimation in a class of nonregular cases, *Biometrika*, **72**, p. 67-90.