Lecture Notes on the Principles and Methods of Applied Mathematics

Michael (Misha) Chertkov
(lecturer)
and **Colin Clark**
(recitation instructor for this and other core classes)

Graduate Program in Applied Mathematics,
University of Arizona, Tucson

March 17, 2023

Contents

Chapter 1

Applied Math Core Courses

Every student in the Program for Applied Mathematics at the University of Arizona takes the same three core courses during their first year of study. These three courses are called Methods (Math 581), Theory (Math 584), and Algorithms (Math 589). Each course presents a different expertise, or 'toolbox' of competencies, for approaching problems in modern applied mathematics. The courses are designed to discuss many of the same topics, often synchronously, (Fig. 1.1). This allows them to better illustrate the potential contributions of each toolbox, and also to provide a richer understanding of the applied mathematics. The material discussed in the courses include topics that are taught in traditional applied mathematics curricula (like differential equation) as well as topics that promote a modern perspective of applied mathematics (like optimization, control and elements of computer science and statistics). All the material is carefully chosen to reflect what we believe is most relevant now and in the future.

The essence of the core courses is to develop the different toolboxes available in applied

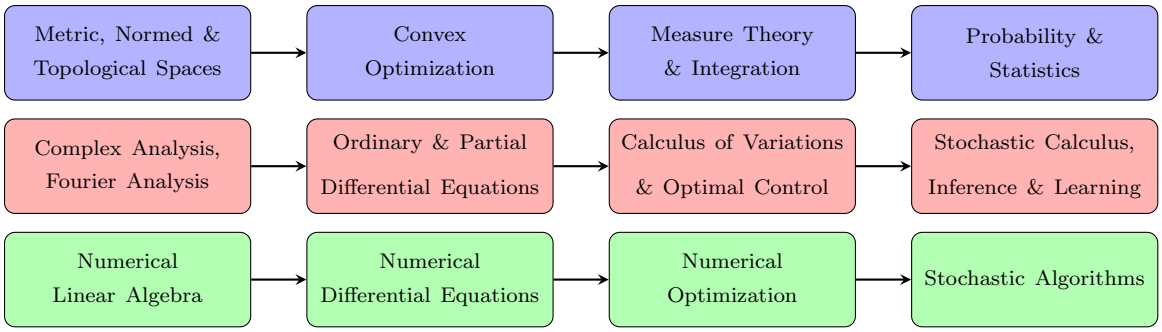


Figure 1.1: Topics covered in Theory (blue), Methods (red) and Algorithms (green) during the Fall semester (columns 1 & 2) and Spring semester (columns 3 & 4)

mathematics. When we're lucky, we can find exact solutions to a problem by applying powerful (but typically very specialized) techniques, or methods. More often, we must formulate solutions algorithmically, and find approximate solutions using numerical simulations and computation. Understanding the theoretical aspects of a problem motivates better design and implementation of these methods and algorithms, and allows us to make precise statements about when and how they will work.

The core courses discuss a wide array of mathematical content that represents some of the most interesting and important topics in applied mathematics. The broad exposure to different mathematical material often helps students identify specific areas for further in-depth study within the program. The core courses do not (and cannot) satisfy the in-depth requirements for a dissertation, and students must take more specialized courses and conduct independent study in their areas of interest.

Furthermore, the courses do not (and cannot) cover all subjects comprising applied mathematics. Instead, they provide a (somewhat!) minimal, self-consistent, and admittedly subjective (due to our own expertise and biases) selection of the material that we believe students will use most during and after their graduate work. In this introductory chapter of the lecture notes, we aim to present our viewpoint on what constitutes modern applied mathematics, and to do so in a way that unifies seemingly unrelated material.

What is Applied Mathematics?

We study and develop mathematics as it applies to model, optimize and control various physical, biological, engineering and social systems. Applied mathematics is a combination of (1) mathematical science, (2) knowledge and understanding from a particular domain of interest, and often (3) insight from a few 'math-adjacent' disciplines (Fig. 1.2). In our program, the core courses focus on the mathematical foundations of applied math. The more specialized mathematics and the domain-specific knowledge are developed in other coursework, independent research and internship opportunities.

Applying mathematics to real-world problems requires mathematical approaches that have evolved to stand up to the many demands and complications of real-world problems. In some applications, a relatively simple set of governing mathematical expressions are able to describe the relevant phenomena. In these situations, problems often require very accurate solutions, and the mathematical challenge is to develop methods that are efficient (and sometimes also adaptable to variable data) without losing accuracy. In other applications, there is no set of governing mathematical expressions (either because we do no know them, or because they may not exist). Here, the challenge is to develop better mathematical

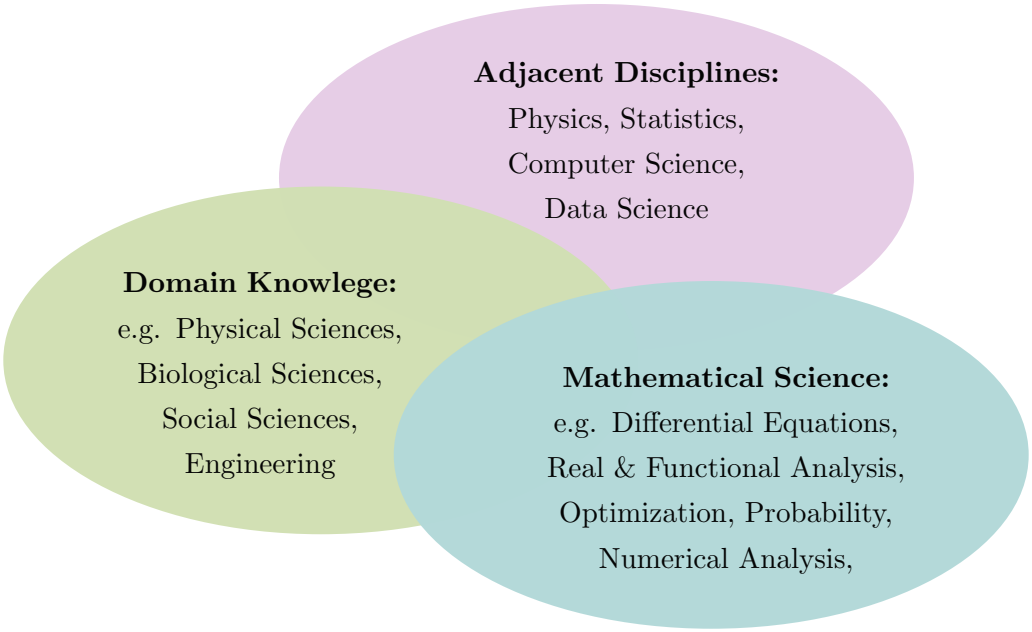


Figure 1.2: The key components studied under the umbrella of applied mathematics: (1) mathematical science, (2) domain-specific knowledge, and (3) a few 'math-adjacent' disciplines.

descriptions of the phenomena by processes, interpreting and synthesizing imperfect observations. In terms of the general methodology maintained throughout the core courses, we devote considerable amount of time to:

1. Formulating the problem, first casually, i.e. in terms standard in sciences and engineering, and then transitioning to a proper mathematical formulation;

2. Analyzing the problem by "all means available", including theory, method and algorithm toolboxes developed within applied mathematics;

3. Identifying what kinds of solutions are needed, and implementing an appropriate method to find such a solution.

Making contributions to a specific domain that are truly valuable requires more than just mathematical expertise. Domain-specific understanding may change our perspective for what constitutes a solution. For example, whenever system parameters are no longer 'nice' but must be estimated from measurement or experimental data, it becomes more the difficult to finding meaning in the solutions, and it becomes more important, and challenging, to estimate the uncertainty in solutions,. Similarly, whenever a system couples many subsystems at scale, it may be no longer possible to interpret the exact expressions, (if they

can be computed at all) and approximate, or 'effective' solutions may be more meaningful. In every domain-specific application, it is important to know what problems are most urgent, and what kinds of solutions are most valuable.

Mathematics is not the only field capable of making valuable contributions to other domains, and we think specifically of physics, statistics and computer science as other fields that have each developed their own frameworks, philosophies, and intuitions for describing problems and their solutions. This is particularly evident with the recent developments in data science. The recent deluge of data has brought a wealth of opportunity in engineering, and in the physical, natural and social sciences where there have been many open problems that could only be addressed empirically. Physics, statistics, and computer science have become fundamental pillars of data science, in part, because each of these 'math-adjacent' disciplines provide a way to analyze and interpret this data constructively. Nonetheless, there are many unresolved challenges ahead, and we believe that a mixture of mathematical insight and some intuition from these adjacent disciplines may help resolve these challenges.

Problem Formulation

We will rely on a diverse array of instructional examples from different areas of science and engineering to illustrate how to translate a rather vaguely stated scientific or engineering phenomenon into a crisply stated mathematical challenge. Some of these challenges will be resolved, and some will stay open for further research. We will be refering to instructional examples, such as the Kirchoff and the Kuramoto-Sivashinsky equations for power systems, the Navier-Stokes equations for fluid dynamics, network flow equations, the Fokker-Plank equation from statistical mechanics, and constrained regression from data science.

Problem Analysis

We analyze problems extracted from applications by all means possible, which requires both domain-specific intuition and mathematical knowledge. We can often make precise statements about the solutions of a problem without actually solving the problem in the mathematical sense. **Dimensional analysis** from physics is an example of this type of preliminary analysis that is helpful and useful. We may also identify certain properties of the solutions by analyzing any underlying **symmetries** and establishing the correct **principal behaviors** expected from the solutions, some important example involve oscillatory behavior (waves), diffusive behavior, and dissipative/decaying vs. conservative behaviors. One can also extract a lot from analyzing the different **asymptotic regimes** of a problem, say when a parameter becomes small, making the problem easier to analyze. Matching different

asymptotic solutions can give a detailed, even though ultimately incomplete, description.

Solution Construction

As previously mentioned, one component of applied mathematics is a collection of specialized techniques for finding analytic solutions. These techniques are not always feasible, and developing **computational intuition** should help us to identify proper methods of numerical (or mixed analytic-numerical) analysis, i.e. a specific toolbox, helping to unravel the problem.

Part I

Applied Analysis

Chapter 2

Complex Analysis

The real number system is somewhat "deficient" in the sense that not all operations are allowed for all real numbers. For example, taking arbitrary roots of negative numbers is not allowed in the real number system. This deficiency can be remedied by defining the *imaginary unit*, $i := \sqrt{-1}$. A number that is a real multiple of the imaginary unit, for example $3i$, $i/2$ or $-\pi i$, is called an *imaginary number*. A number that has both a real and an imaginary component is called a *complex number*.

Complex analysis is the branch of mathematics that investigates functions of complex variables. A fundamental premise of complex analysis is that *most* operations between real numbers have natural extensions to complex numbers, and that *most* real-valued functions have natural extensions to complex-valued functions. Interestingly, the natural extensions of even the most elementary functions can lead to a richness that often admits new techniques for problem solving.

Complex analysis provides useful tools for many other areas of mathematics, (both pure and applied), as well as for physics (including the branches of hydrodynamics, thermodynamics, and particularly quantum mechanics), and engineering fields (such as aerospace, mechanical and electrical engineering).

2.1 Complex Variables and Complex-valued Functions

2.1.1 The Cartesian Representation of Complex Variables

For two complex numbers, $z_1 = a_1 + ib_1$, and, $z_2 = a_2 + ib_2$, we have

- Addition: $z_1 + z_2 = (a_1 + ib_1) + (a_2 + ib_2) = (a_1 + a_2) + i(b_1 + b_2)$,
- Multiplication: $z_1 z_2 = (a_1 + ib_1)(a_2 + ib_2) = a_1 a_2 + i(a_1 b_2 + b_1 a_2) + i^2 b_1 b_2 = (a_1 a_2 - b_1 b_2) + i(a_1 b_2 + b_1 a_2)$.

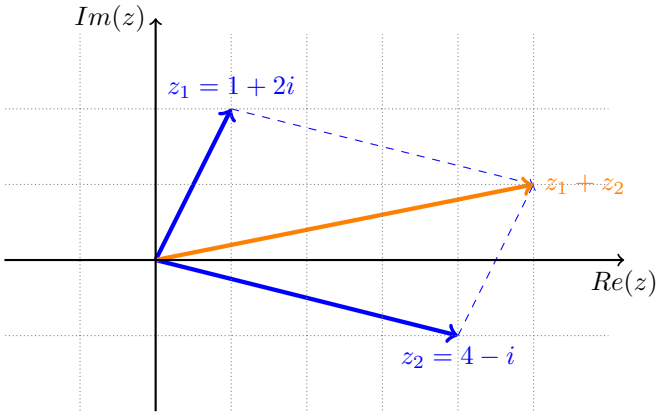


Figure 2.1: Complex numbers can be visualized as vectors in $\mathbb{R}^2$. By convention, the real component is plotted on the horizontal axis, and the imaginary component is plotted on the vertical axis. The addition of two complex numbers is reminiscent of vector addition in $\mathbb{R}^2$.

The addition and subtraction of complex numbers are direct generalizations of their real-valued counterparts.

Example 2.1.1. Let $z_1 = 1 + 2i$ and $z_2 = 4 - i$. Compute (a) $z_1 + z_2$ and (b) $z_1 - z_2$.

Solution.

(a) $z_1 + z_2 = (1 + 2i) + (4 - i) = (1 + 4) + (2 - 1)i = 5 + i$

(b) $z_1 - z_2 = (1 + 2i) - (4 - i) = (1 - 4) + (2 + 1)i = -3 + 3i$

Because the behavior of addition and subtraction is reminiscent of translating vectors in $\mathbb{R}^2$, we often visualize complex numbers as points on a Cartesian plane by associating the the real and imaginary components of the complex number with the x- and y-coordinates respectively.

Definition 2.1.2. The *complex conjugate* of a complex number z, denoted by z^* or $\bar{z}$, is the complex number with an equal real part and an imaginary part equal in magnitude but opposite in sign. That is, if $z = x + iy$ then $z^* := x - iy$.

The multiplication and division of complex numbers are also direct generalizations of their real-valued counterparts with the additional definition, $i^2 = -1$.

Example 2.1.3. Let $z_1 = -1 + 2i$ and $z_2 = 4 - 3i$. Compute (a) $z_1 z_2$, (b) z_1/z_2.

Solution.

(a) $z_1 z_2 = (-1+2i)(4-3i) = -4+3i+8i-6i^2 = 2+11i.$

(b) To compute z_1/z_2, we first multiply it by z_2^*/z_2^*, so that the denominator, $z_2 z_2^*$, is a real number,

$$\frac{z_1}{z_2} = \frac{z_1}{z_2}\left(\frac{z_2^*}{z_2^*}\right) = \frac{-1+2i}{4-3i}\left(\frac{4+3i}{4+3i}\right) = \frac{-4-3i+8i+6i^2}{16-12i+12i+9} = \frac{-10+5i}{25} = -\frac{2}{5}+\frac{1}{5}i.$$

Complex conjugates

Theorem 2.1.4. For algebraic operations including addition, multiplication, division and exponentiation, consider a sequence of algebraic operations over the n complex numbers $z_1, \ldots, z_n$ with the result w. If the same actions are applied in the same order to $z_1^*, \ldots, z_n^*$, then the result will be w^*.

Example 2.1.5. Let us illustrate the Theorem 2.1.4 on the example of a quadratic equation, $az^2+bz+c=0$, where the coefficients, a, b and c are real. Direct application of the Theorem 2.1.4 to this example results in the fact that if the equation has a root, then its complex conjugate is also a root, which is obviously consistent with the roots of quadratic equations formula, $z_{1,2} = (-b \pm \sqrt{b^2-4ac})/(2a)$.

Exercise 2.1. Use Theorem 2.1.4 to show that the roots of a polynomial with real-valued coefficients of *arbitrary* order occur in complex conjugate pairs.

Example 2.1.6. Find all the roots of the polynomial, $p(z) = z^4 - 6z^3 + 11z^2 - 2z - 10$, given that one of its roots is $2-i$.

Solution. We observe that $p(z)$ has real-valued coefficients, so its roots occur in conjugate pairs; given that $z_1 = 2-i$ is a root, then $z_2 = 2+i$ must also be a root, which we verify by evaluation. We factorize $p(z)$ as $p(z) = (z-z_1)(z-z_2)r(z)$, where we find $r(z)$ by polynomial division, giving $r(z) = (z^2-2z-2)$. Therefore, the four roots of $p(z)$ are found by solving

$$0 = z^4 - 6z^3 + 11z^2 - 2z - 10 = (z-z_1)(z-z_2)(z^2-2z-2).$$

Solving $z^2-2z-2=0$ by the quadratic formula gives $z_{3,4} = 1 \pm \sqrt{3}$. Thus, the four roots of $p(z)$ are:

$$z_1 = 2-i, \qquad z_2 = 2+i, \qquad z_3 = 1+\sqrt{3}, \qquad z_4 = 1-\sqrt{3}.$$

Example 2.1.7. Let $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$. Show that if $\omega = z_1/z_2$, then $\omega^* = z_1^*/z_2^*$.

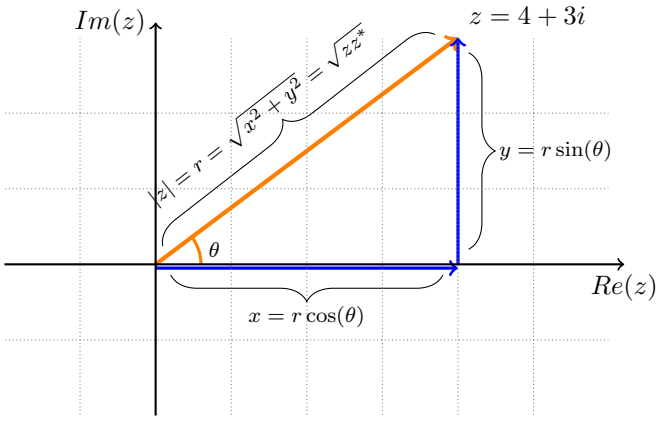


Figure 2.2: A complex number, z, has both a Cartesian representation (shown in blue) and a polar representation (shown in orange). Its modulus, denoted by $|z|$ or r, is non-negative and satisfies $|z|^2 = r^2 := zz^* = x^2 + y^2$. Its argument, denoted by θ, is the angle measured modulo 2π, counter-clockwise from the positive real axis.

Solution. From the definition of a complex conjugate, we have

$$z_1^* = x_1 - iy_1, \quad \text{and} \quad z_2^* = x_2 - iy_2.$$

We must find ω^* and verify that it is equivalent to z_1^*/z_2^*. First compute ω,

$$\omega = \frac{x_1 + iy_1}{x_2 + iy_2} = \frac{x_1 + iy_1}{x_2 + iy_2}\left(\frac{x_2 - iy_2}{x_2 - iy_2}\right) = \frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2} + i\frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2}.$$

Now compute ω^*,

$$\omega^* = \frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2} - i\frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2} = \frac{(x_1 - iy_1)(x_2 + iy_2)}{(x_2 - iy_2)(x_2 + iy_2)} = \frac{x_1 - iy_1}{x_2 - iy_2},$$

which is equivalent to z_1^*/z_2^*, as required.

2.1.2 The Polar Representation of Complex Variables

In addition to their Cartesian representation, complex numbers can also be represented by their polar representation with components r and θ. Here r is called the modulus of z and satisfies $r^2 = |z|^2 := zz^* = x^2 + y^2 \geq 0$, and θ is called the argument of z or sometimes the polar angle. Note that $\theta = \arg(z)$ is defined only for $|z| > 0$, and modulo 2π:

$$x + iy \;=\; r\cos\theta + ir\sin\theta, \qquad \text{for} \qquad r = x^2 + y^2, \text{ and } \theta = \arctan(y, x).$$

The application of trigonometric identities shows that the product of two complex numbers is the complex number whose modulus is the product of the moduli of its factors, and

whose argument is the sum of the arguments of its factors. That is, if $z_1 = r_1\cos\theta_1 + ir_1\sin\theta_1$, and $z_2 = r_2\cos\theta_2 + ir_2\sin\theta_2$, then $z_1z_2 = r_1r_2\cos(\theta_1+\theta_2) + ir_1r_2\sin(\theta_1+\theta_2)$. This summation of arguments whenever two functions are multiplied together is reminiscent of multiplying exponential functions. The polar representation is simplified by defining the complex-valued exponential function. When expressed in their polar representation, the multiplication of two complex numbers is simplified by elementary trigonometric identities. For $z_1 = r_1\cos\theta_1 + ir_1\sin\theta_1$ and $z_2 = r_2\cos\theta_2 + ir_2\sin\theta_2$, their product is

$$\begin{aligned} z_1z_2 &= r_1r_2\cos\theta_1\cos\theta_2 + r_1r_2\sin\theta_1\sin\theta_2 + ir_1r_2\cos\theta_1\sin\theta_2 + ir_1r_2\sin\theta_1\cos\theta_2 \\ &= r_1r_2\cos(\theta_1+\theta_2) + ir_1r_2\sin(\theta_1+\theta_2). \end{aligned}$$

We make two observations: first, the modulus of the product of two complex numbers is the product of their moduli, that is, $|z_1z_2| = |z_1|\,|z_2|$, and second the argument of the product is the *sum* of arguments, that is, $\arg(z_1z_2) = \arg(z_1)+\arg(z_2)$. These observations, which are reminiscent of the multiplication of real-valued exponential functions, motivate the definition of the complex-valued exponential function.

Definition 2.1.8. The *exponential function* is defined for imaginary arguments by

$$re^{i\theta} := r\cos(\theta) + ir\sin(\theta) = x + iy. \tag{2.1}$$

Euler's (famous) formula, $e^{i\pi} = -1$, follows directly from this definition.

Example 2.1.9. Convert $z_1 = -1+2i$ and $z_2 = 4-3i$ to their polar representations and compute (a) their product and (b) their quotient. Compare you answer to Example 2.1.3.

Solution. The polar representations of z_1 and z_2 are:

$$\begin{aligned} r_1 &= z_1z_1^* = \sqrt{5}, & \theta_1 &= \tan^{-1}(10/-5) \approx 2.03, & z_1 &= \sqrt{5}e^{2.03i}; \\ r_2 &= z_2z_2^* = 5, & \theta_2 &= \tan^{-1}(-3/4), \approx -0.64, & z_2 &= 5e^{-0.64i}. \end{aligned}$$

Their product and quotient are:

(a) $z_1z_2 \approx \left(\sqrt{5}e^{2.03i}\right)\left(5e^{-0.64i}\right) = 5\sqrt{5}e^{1.39i} = 5\sqrt{5}\cos(1.39) + i5\sqrt{5}\sin(1.39) = 2+11i.$

(b) $z_1/z_2 \approx \sqrt{5}e^{2.03i}/5e^{-0.64i} = 5\sqrt{5}e^{2.67i} = 1/\sqrt{5}\cos(2.67)+i/\sqrt{5}\sin(2.67) = -0.4+0.2i.$

Sometimes it is convenient to express a complex number using a mixture of Cartesian and polar representations.

Example 2.1.10. Find $\tilde{r}$ and $\tilde{\theta}$ such that the point, $\omega = 1+5i$, can be written as, $\omega = -1 + \tilde{r}e^{i\tilde{\theta}}$.

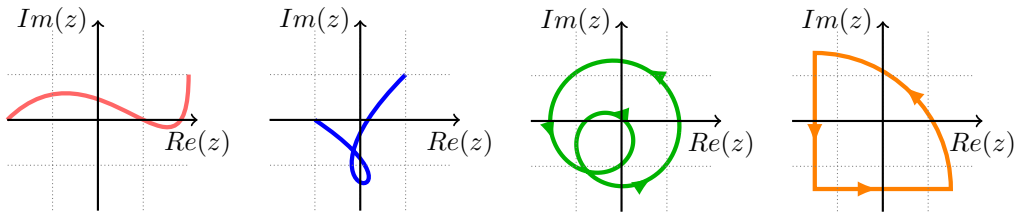


Figure 2.3: Examples of curves in the complex plane. The first two curves (red and blue) are open and the last two curves (green and orange) are closed. The first and fourth curves (red and orange) are simple and the second and third curves (blue and green) are not simple because they self-intersect at points other than the end points.

Solution. Given that $1 + 5i = -1 + \tilde{r}e^{i\tilde{\theta}}$, solves for $\tilde{r}e^{i\tilde{\theta}}$ to get $2 + 5i = \tilde{r}e^{i\tilde{\theta}}$. Solve for $\tilde{r}$ and $\tilde{\theta}$ to get $\tilde{r} = (2 + 5i)(2 - 5i) = \sqrt{29} \approx 5.39$ and $\tilde{\theta} = \tan^{-1}(5/2) \approx 1.19$rad. Therefore, $w \approx -1 + 5.39e^{1.19i}$.

Example 2.1.11. Express $z := (2 + 2i)e^{-i\pi/6}$ by its (a) Cartesian and (b) polar representations.

Solution.

(a) $z = (2+2i)(\cos(-\frac{\pi}{6}) + i\sin(-\frac{\pi}{6})) = \big(2\cos(-\frac{\pi}{6}) + 2\sin(-\frac{\pi}{6})\big) + i\big(2\cos(-\frac{\pi}{6}) + 2\sin(-\frac{\pi}{6})$
$= (1 + \sqrt{3}) + i(\sqrt{3} - 1)$.

(b) $z = (2 + 2i)e^{-i\pi/6} = 2\sqrt{2}e^{\pi/4}e^{-i\pi/6} = 2\sqrt{2}e^{i\pi/12}$.

2.1.3 Parameterization of Curves in the Complex Plane

Definition 2.1.12. A *curve* in the complex plane is a set of points $z(t)$ where, $a \le t \le b$, for some $a \le b$. We say that the curve is *closed* if, $z(a) = z(b)$, and *simple* if it does not self-intersect, except possibly at the end-points. That is the curve is simple if, $z(t) \neq z(t')$ for $t \neq t'$ and $a < t, t' < b$. A curve is called a *contour* if it is continuous and piecewise smooth. By convention, all simple, closed contours are parameterized to be traversed counter-clockwise, unless stated otherwise.

Example 2.1.13. Parameterize the following curves:

(a) The infinite 'vertical' line passing through $\pi/2$.

(b) The semi-infinite ray extending from the point $z = -1$ and passing through $\sqrt{3}i$.

(c) The circular arc of radius ε centered at 0.

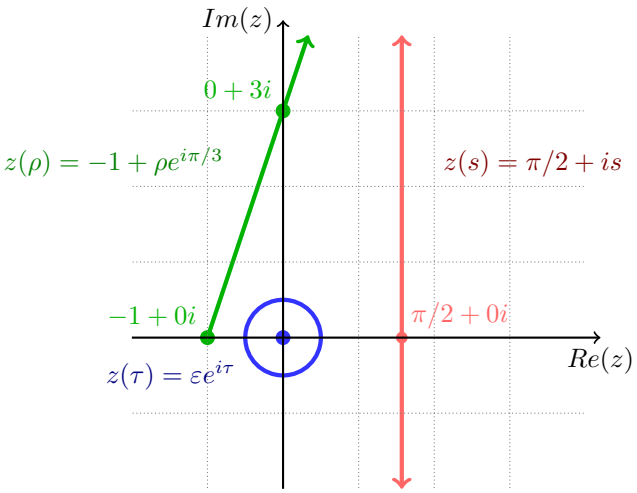


Figure 2.4: Parameterized curves for example 2.1.13. Red: The infinite 'vertical' line passing through $\pi/2$ is parameterized by $z(s) = \pi/2 + is$ for $-\infty < s < \infty$. Green: The semi-infinite ray extending from the point $z = -1$ and passing through $\sqrt{3}i$ is parameterized by $z(\rho) = -1 + \rho e^{i\pi/3}$ for $0 \leq \rho < \infty$. Blue: The circular arc of radius ε centered at 0 is parameterized by $z(\tau) = \varepsilon e^{i\tau}$ for $0 \leq \tau \leq 2\pi$.

Solution.

(a) $z(s) = \pi/2 + is$ for $-\infty < s < \infty$.

(b) $z(\rho) = -1 + \rho e^{i\pi/3}$ for $0 \leq \rho < \infty$.

(c) $z(\tau) = \varepsilon e^{i\tau}$ for $0 \leq \tau \leq 2\pi$.

The Complex Number System

Complex numbers can be considered as the resolution of the notation for numbers that are closed under all possible algebraic operations. What this means is that any algebraic operation between two complex numbers is guaranteed to return another complex number. This is not generally true for other classes of numbers, for example,

i. The addition of two positive integers is guaranteed to be another positive integer, but the subtraction of two positive integers *is not necessarily* a positive integer. Therefore, we say that the positive integers are closed under addition but are *not* closed under subtraction.

ii. The class of all integers is closed under subtraction and also multiplication. However the integers are not closed under division because the quotient of two integers is not necessarily another integer.

iii. The rational numbers are closed under division. However the process of taking limits of rational numbers may lead to numbers that are not rational, so real numbers are needed if we require a system that is closed under limits.

iv. Taking non-integer powers of negative numbers does not yield a real number. The class of complex numbers must be introduced to have a system that is closed under this operation.

Moreover one finds that the class of complex numbers is also closed under the operations of finding a root of algebraic equations, of taking logarithms, and others. We conclude with a happy statement that the class of complex numbers is closed under all the operations.

2.1.4 Functions of a Complex Variable

A function of a complex variable, $w = f(z)$, maps the complex number z to the complex number w. That is, f maps a point in the z-complex plane to a point (or points) in the w-complex plane. Since both z and w have a Cartesian representation, this means that every function of a complex variable can be expressed as two real-valued functions of two real variables, $f(z) := u(x, y) + iv(x, y)$.

2.1.5 Complex Exponentials

In Eq. (2.1) we motivated the definition of the exponential function, $f(z) = e^z$, with the intention to preserve the property that, $e^{z_1+z_2} = e^{z_1}e^{z_2}$, and incidentally that, $e^1 = 2.718\ldots$. This is not the only property we could have chosen to motivate the definition of e^z. We could have chosen to preserve any of the following properties:

- the function's Taylor series $\sum_{n=0}^{\infty} z^n/n!$;
- the function's limiting expression $\lim_{n\to\infty}(1 + z/n)^n$;
- the fact that $f(z) = e^z$ solves the Ordinary Differential Equation, $f'(z) = f(z)$, subject to $f(0) = 1$.

We encourage the reader to verify that all these properties are preserved for the complex exponential, and that any one of them could have motivated our definition and yielded the same results.

An immediate consequence of this observation is that the natural definitions of the complex-valued trigonometric functions are

$$\cos(z) := \frac{e^{iz} + e^{-iz}}{2} \quad \text{and} \quad \sin(z) := \frac{e^{iz} - e^{-iz}}{2i}. \tag{2.2}$$

Example 2.1.14. Let $f(z) = \exp(iz)$ where $z = x + iy$. Express $f(z)$ as the sum, $u(x, y) + iv(x, y)$, where u and v are real-valued functions of x and y.

Solution.

$$\begin{aligned} f(z) &= \exp(i(x + iy)) = \exp(ix - y) = \exp(-y)\exp(ix) \\ &= \exp(-y)\cos(x) + i\exp(-y)\sin(x) \end{aligned}$$

□

Example 2.1.15. Evaluate the functions along the curves : (a) $z \mapsto \sin z$ along the infinite horizontal line passing through $\pi/2$. (b) $z \mapsto \exp(z+1)$ along the semi-infinite ray extending from the point $z = -1$ and passing through $\sqrt{3}i$, and (c) $z \mapsto z^2$ along circular arc of radius ε centered at 0. See example 2.1.13 and also Fig. 2.4.

Solution.

(a) Parameterize the vertical line passing through $\pi/2$ by $\pi/2 + is$ for $-\infty < s < \infty$

$$\begin{aligned} f(s + i\pi/2) = \sin(\pi/2 + is) &= \frac{e^{i\pi/2 - s} - e^{-i\pi/2 + s}}{2i} = \frac{ie^{-s} + ie^{s}}{2i} \\ &= \cosh(s) \quad \text{for } -\infty < s < \infty. \end{aligned}$$

(b) Parameterize the semi-infinite ray extending from the point $z = -1$ and passing through $\sqrt{3}i$ by $-1 + \rho e^{i\pi/3}$ for $0 < \rho < \infty$.

$$\begin{aligned} f(-1 + \rho e^{i\pi/3}) &= \exp(\rho e^{i\pi/3}) = \exp(\rho\cos(i\pi/3) + i\rho\sin(i\pi/3)) \\ &= e^{\rho/2}\left(\cos(\rho\sqrt{3}/2) + i\sin(\rho\sqrt{3}/2)\right) \quad \text{for } 0 \le \rho < \infty. \end{aligned}$$

(c) Parameterize the circular arc of radius ε centered at 0 by $\varepsilon e^{i\tau}$ for $0 \le \tau \le 2\pi$.

$$f(\varepsilon e^{i\tau}) = (\varepsilon e^{i\tau})^2 = \varepsilon^2 e^{2i\tau} \quad \text{for } 0 \le \tau \le 2\pi.$$

Exercise 2.2. Investigate the asymptotic behavior at $|z| \to \infty$ of the complex-valued functions (a) $f(z) = \exp(z)$, (b) $f(z) = \sin(z)$, (c) $f(z) = \cos(z)$. *Hint:* There are many different ways that $|z|$ can go to infinity. For example, we could write $z = x + iy$ and let $x \to \pm\infty$ for fixed (or variable) y or let $y \to \pm\infty$ for fixed (or variable) x. We could also write $z = re^{i\theta}$ and let $r \to \infty$ for fixed (or variable) θ. We are asking you to consider each function from whichever perspectives are most informative for determining what happens as $|z| \to \infty$.

2.1.6 Multi-valued Functions and Branch Cuts

Not every complex function is single-valued. We often deal with functions that are multi-valued, meaning that for some z, there exist two or more w_i such that $f(z) = w_i$. Recall the parametrized curve in example 2.1.15 and consider (c)(i) where we evaluated the function $f(z) = z^2$ along the circle of radius ε centered at the origin. Notice, in particular, that the function returns to its original value, that is, $f(\varepsilon e^{0i}) = f(\varepsilon e^{2\pi i}) = \varepsilon^2$. It may be seem surprising, but there are functions where this is not the case.

Example 2.1.16. Consider the example of $\omega(z) = \sqrt{z}$. When z is represented in polar coordinates, $z = r\exp(i\theta)$, we know that θ is defined up to a shift on $2\pi n$, for any integer n. For our example, this translates to $\omega_n(z) = \sqrt{r}\exp(i\theta/2 + i\pi n)$, where different n will result in (two) different values of $\sqrt{z}$, called two branches, $\omega_1 = \sqrt{r}\exp(i\theta/2)$, $\omega_2 = \sqrt{r}\exp(i\theta/2 + i\pi)$. If we choose one branch, say ω_1, and walk in the complex plane around $z = 0$ in a positive counter-clockwise direction (so that $z = 0$ always stays on the left) changing θ from its original value, say $\theta = 0$, to $\pi/2, \pi, 3\pi/2$ and eventually get to 2π, ω_1 will transition to ω_2. Making one more positive 2π swing will return to ω_1. In other words, the two branches transition to each other after one makes a 2π turn. Per definition below, the point $z = 0$ is called a second order branch point of the two-valued function $\sqrt{z}$.

Definition 2.1.17. A multi-valued function $w(z)$ has a *branch point* at $z_0 \in \mathbb{C}$ if $w(z)$ varies continuously along a sufficiently small circuit surrounding z_0, but does not return to its starting values after one full circuit.

Definition 2.1.18. A *branch* of a multi-valued function $w(z)$ is a single-valued function that is obtained by restricting the image of $w(z)$ and disregarding all but one set of values.

A multi-valued function has the property that if we traverse a sufficiently small closed contour around its branch point, we experience a discontinuity. One should note that the location of this discontinuity is entirely dependent on where we choose to start and stop the closed contour. To see this consider the following two closed contours:

$$\alpha(\theta) = e^{i\theta}, \qquad 0 \le \theta \le 2\pi, \tag{2.3}$$

$$\beta(\phi) = e^{i\phi}, \qquad -\pi \le \phi \le \pi. \tag{2.4}$$

If we traverse these two contours with the function $f(z) = \sqrt{z}$, we see that the discontinuity occurs at $\theta = 0, 2\pi$ in the first case, and $\phi = -\pi, \pi$ in the second case. In truth, the location of this discontinuity was dependent on our choice of contour. We can expand on this idea by introducing the notion of a *branch cut*. A branch cut is something *we* pick in

order to separate the branches of a multi-valued function. For most multi-valued functions, this means that when we attempt to traverse a cut with some closed contour, we end up experiencing a discontinuity. Really what we have is a contour that is closed in the domain of f, but maps to an open contour.

Definition 2.1.19. A *branch cut* is a curve in the complex plane along which a branch is discontinuous.

Remark. Branch cuts are usually not unique, and are something that is defined by us, not the multi-valued function in question. One branch is arbitrarily selected as the principal branch. Most software packages employ a set of rules for selecting the principal branch of a multi-valued function.

Example 2.1.20. The generalization of example 2.1.16 to $\omega(z) = z^{1/n}$ is straightforward. This function has n branches, $\omega_1(z), \ldots, \omega_n(z)$ and thus $z = 0$ is called an n^{th} order branch point of the n^{th}-valued function $z^{1/n}$.

Example 2.1.21. Another important example is $\omega(z) = \log(z)$. We can represent z by its polar representation, $z = re^{i(\theta+2\pi n)}$, to show that log is a multi-valued function with infinitely many (but countable number of) roots, $\omega_n = \log(r) + i(\theta + 2n\pi), \quad n = 0, \pm 1, \ldots$. In this case, $z = 0$ is an infinite order branch point.

Definition 2.1.22. (Branch points at $z = \infty$.) Consider a multi-valued function $f(z)$. We say that f has a branch point at $z = \infty$ if the function $g(w) = f(1/w)$ has a branch point at $w = 0$.

Example 2.1.23. Find the branch points of $\log(z-1)$ and sketch a set of possible branch cuts. Choose a branch cut and describe the resulting branches.

Solution. Parameterize the function as follows, $\log(z-1) = \log\rho + i\phi$, where $z - 1 = \rho\exp(i\phi)$ with $\rho > 0$ (non-negative real) and ϕ real. Since ϕ changes by multiples of 2π as we travel on a closed path around $z = 1$, the point $z = 1$ is a branch point of $\log(z-1)$. We can observe that $z = \infty$ is also a branch point (thus infinite branch point) by replacing z with $z = \frac{1}{w}$ and observing that $w = 0$ is a branch point. Therefore a valid branch cut for the function should connect the two branch points as illustrated in Fig. (2.7).

To describe branches of the function, let us choose (for concreteness) the branch cut starting at $z = 1$ and moving along the x axis to $z = +\infty$. Introduce potential branches of z by $z_n = 1 + \rho\exp(i\phi + 2i\pi n),\ n = 0, 1, \cdots$. Given that $f(z) = \log(z-1)$, the family of z_n translates into the following branches: $f_n(z) = \log\rho + i\phi + 2i\pi n$. Observe that each branch is distinct from the others and that each is a single-valued, analytic function in $\mathbb{C}$ excluding the branch cut.

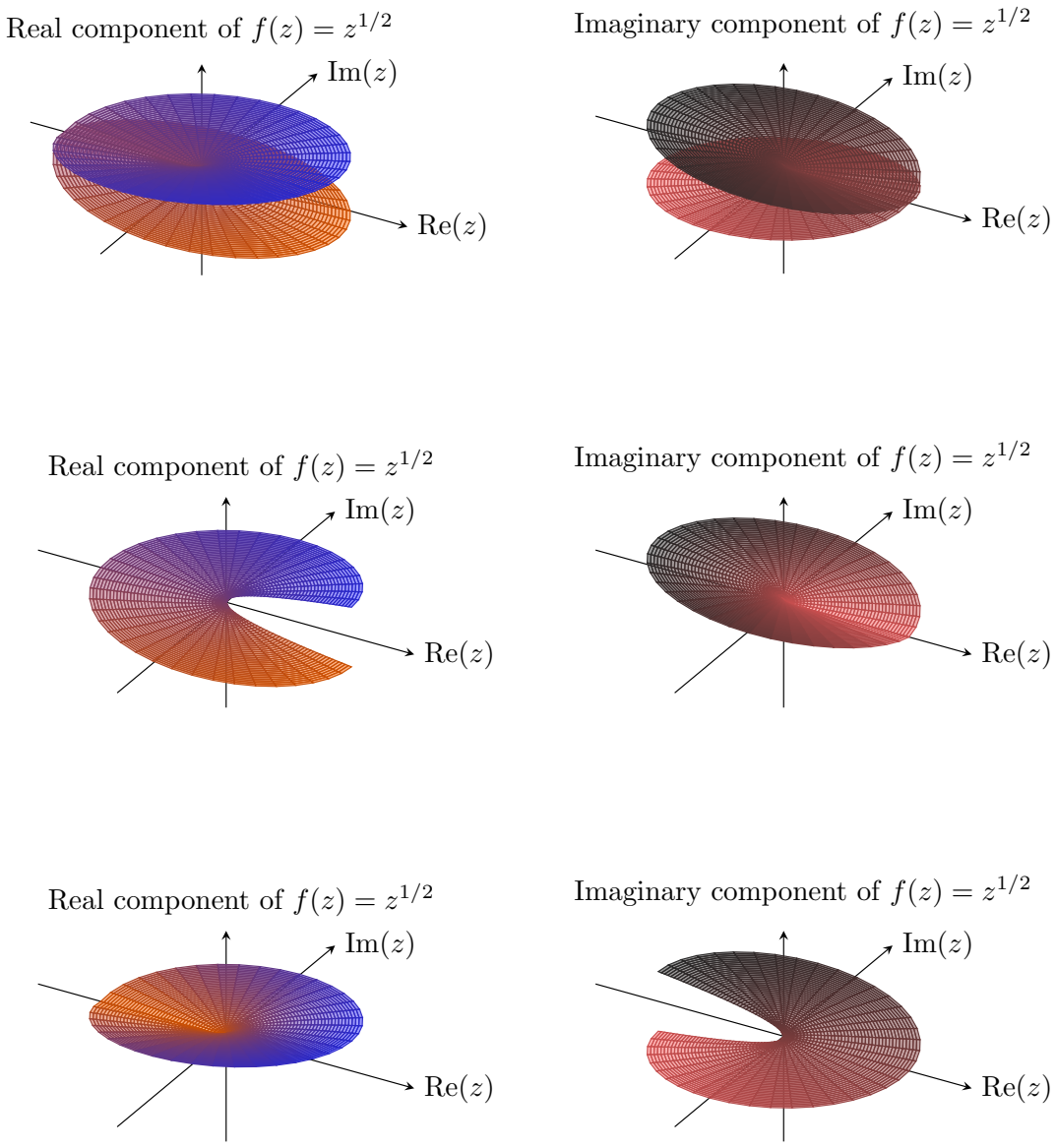


Figure 2.5: (a) Top row: The real (left) and imaginary (right) components of $z \mapsto z^{1/2}$. (b) Middle row: The representation, $z = re^{-i\theta}$, with $0 \leq \theta < 2\pi$, gives a branch cut along the positive real axis and a (single-valued) branch that is analytic everywhere except along the branch cut. (c) Bottom row: The representation, $z = re^{-i\theta}$, with $-\pi \leq \theta < \pi$, gives a branch cut along the negative real axis and a (single valued) branch of $z \mapsto z^{1/2}$ that is analytic everywhere except along the branch cut.

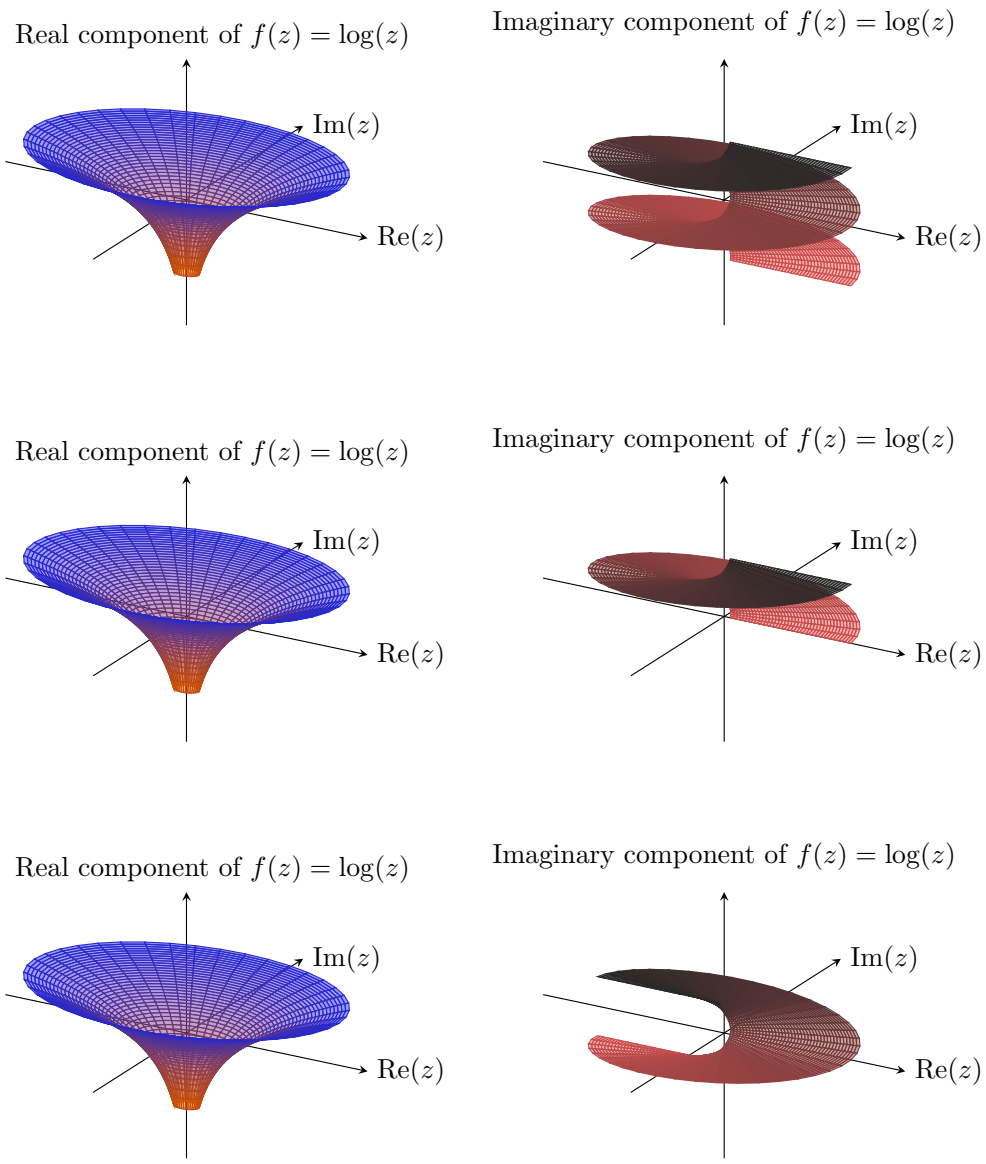


Figure 2.6: (a) Top row: The real (left) and imaginary (right) components of $z \mapsto \log(z)$. (b) Middle row: The representation, $z = re^{-i\theta}$, with $0 \leq \theta < 2\pi$, gives a branch cut along the positive real axis and a (single-valued) branch of $z \mapsto \log(z)$ that is analytic everywhere except along the branch cut. (c) Bottom row: The representation, $z = re^{-i\theta}$, with $-\pi \leq \theta < \pi$, gives a branch cut along the negative real axis and a (single valued) branch of $z \mapsto \log(z)$ that is analytic everywhere except along the branch cut.

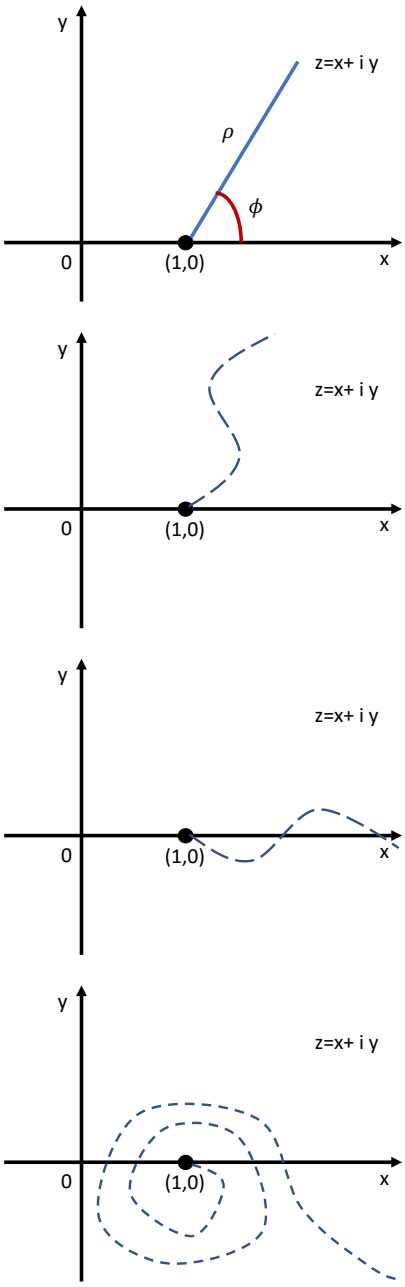


Figure 2.7: Polar parametrization of $\log(z-1)$ (left) and three examples of branch cut for the function connecting its two branch points, at $z = 1$ and at $z = \infty$.

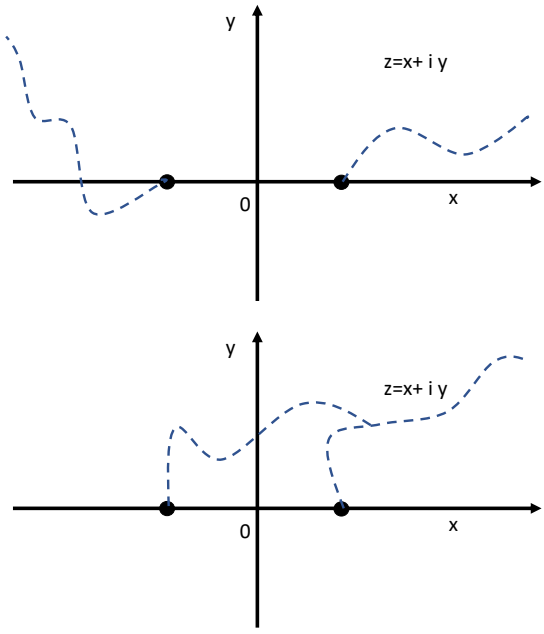


Figure 2.8

Example 2.1.24. Consider $\log(z^2-1) = \log(z-1) + \log(z+1)$. As we travel around $z = 1$, $\log(z-1)$ and also $\log(z^2-1)$ change by 2π. Therefore $z = 1$ is a branch point of $\log(z^2-1)$. Similarly, $z = -1$ and $z = \infty$ are two other branch points of $\log(z^2-1)$. Fig. (2.8) show two branch cut examples of $\log(z^2-1)$.

Two important general remarks are in order.

1. The function $\log(f(z))$ has branch points at the zeros of $f(z)$ and at the points where $f(z)$ is infinite, as well as (possibly) at the points where $f(z)$ itself has branch points. But be careful with this (later possibility): the zeros have to be zeros in the sense of analytic functions and by infinities we mean poles. Other types of (singular) behaviors in $f(z)$ can lead to unexpected results, e.g. check what happens at $z = 0$ when $f(z) = \exp(1/z)$.

2. The fact that a function $f(z)$ or its derivatives may or may not have a (finite) value at some point $z = z_0$, is irrelevant as far as deciding the issue of whether or not z_0 is a branch point of $f(z)$.

Exercise 2.3. Identify the branch points, introduce suitable branch cuts, and describe the resulting branches for the functions (a) $f(z) = \sqrt{(z-a)(z-b)}$, and (b) $g(z) = \log((z-1)/(z-2))$.

The graphs of complex multi-valued functions are in general two-dimensional manifolds in the space $\mathbb{R}^4$. These manifolds are called Riemann surfaces. Riemann surfaces are visualized in three-dimensional space with parallel projection and the image surface in three-dimensional space is rendered on the screen. (See `http://matta.hut.fi/matta/mma/SKK_MmaJournal.pdf` for details and visualization with Mathematica.)

Example 2.1.25. Find all values of $z \in \mathbb{C}$ satisfying the equation, $\sin(z) = 3$.

Solution. Start with the definition of complex valued sin:

$$\frac{e^{iz} - e^{-iz}}{2i} = 3,$$

Multiply each side by $2i\, e^{iz}$

$$(e^{iz})^2 - 6ie^{iz} - 1 = 0.$$

This can be solved using the quadratic formula which, after some algebra, gives

$$e^{iz} = i(3 \pm 2\sqrt{2}).$$

Now, take the natural log of both sides

$$iz = \ln(i) + \ln(3 \pm 2\sqrt{2}).$$

By $\ln(z) = \ln(r) + i(\theta + 2n\pi)$ where $n = 1, 2, 3, ...$, we know that $\ln(i) = \ln(1) + i(\frac{\pi}{2} + 2n\pi)$, so

$$z = \frac{\pi}{2} + 2n\pi \pm i \ln(3 + 2\sqrt{2}).$$

2.2 Analytic Functions and Integration along Contours

2.2.1 Analytic functions

The derivative of a real valued function is defined at a point x via the limiting expression

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x},$$

and we say that the function is differentiable at x if the limit exists and is independent of whether the x is approached from above or below as given by the sign of Δx.

Definition 2.2.1. The derivative of a complex function is defined via a limiting expression:

$$f'(z) = \lim_{\Delta z \to 0} \frac{f(z + \Delta z) - f(z)}{\Delta z}. \tag{2.5}$$

This limit only exists if $f'(z)$ is independent of the direction in the z-plane the limit $\Delta z \to 0$ is taken. (Note: there are infinitely many ways to approach a point $z \in \mathbb{C}$.)

If one sets, $\Delta z = \Delta x$, Eq. (2.5) results in

$$f'(z) = u_x + iv_x,$$

where $f = u + iv$. However, setting $\Delta z = i\Delta y$ results in

$$f'(z) = -iu_y + v_y.$$

A consistent definition of a derivative requires that the two ways of taking the derivative coincide, that is,

$$u_x = v_y, \quad u_y = -v_x, \tag{2.6}$$

and this gives a necessary condition for the following statement.

Theorem 2.2.2 (Cauchy-Riemann Theorem)**.** The function $f(z) = u(x, y) + iv(x, y)$ is differentiable at the point $z = x + iy$ iff (if and only if) the partial derivatives, u_x, u_y, v_x, v_y are continuous and the Cauchy-Riemann conditions (2.6) are satisfied in a neighborhood of z.

Notice that in the explanations which lead us to the Cauchy-Riemann theorem (2.2.2) we only sketched one side of the proof – that it is necessary for the differentiability of $f(z)$ to have the theorem's conditions satisfied. To complete the proof, one needs to also show that Eq. (2.6) is sufficient for the differentiability of $f(z)$. In other words, one needs to show that any function $u(x, y)+iv(x, y)$ is **complex-differentiable** if the Cauchy–Riemann equations hold. The missing part of the proof follows from the following chain of transformations

$$\begin{aligned}
\Delta f = f(z + \Delta z) - f(z) &= \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y + O\left((\Delta x)^2, (\Delta y)^2, (\Delta x)(\Delta y)\right) \\
&= \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right)\Delta z + \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right)\Delta z^* + O\left((\Delta x)^2, (\Delta y)^2, (\Delta x)(\Delta y)\right) \\
&= \frac{\partial f}{\partial z}\Delta z + \frac{\partial f}{\partial z^*}\Delta z^* + O\left((\Delta x)^2, (\Delta y)^2, (\Delta x)(\Delta y)\right) \\
&= \Delta z\left(\frac{\partial f}{\partial z} + \frac{\partial f}{\partial z^*}\frac{\Delta z^*}{\Delta z}\right) + O\left((\Delta x)^2, (\Delta y)^2, (\Delta x)(\Delta y)\right),
\end{aligned} \tag{2.7}$$

where $O\left((\Delta x)^2, (\Delta y)^2, (\Delta x)(\Delta y)\right)$ indicates that we have ignored terms of orders higher or equal than two in Δx and Δy. In transition to the last line of Eq. (2.7) we change variables from (x, y) to (z, z^*), thus using

$$\begin{aligned}
\frac{\partial}{\partial x} &= \frac{\partial z}{\partial x}\frac{\partial}{\partial z} + \frac{\partial z^*}{\partial x}\frac{\partial}{\partial z^*} = \frac{\partial}{\partial z} + \frac{\partial}{\partial z^*}, \\
\frac{\partial}{\partial y} &= \frac{\partial z}{\partial y}\frac{\partial}{\partial z} + \frac{\partial z^*}{\partial y}\frac{\partial}{\partial z^*} = i\frac{\partial}{\partial z} - i\frac{\partial}{\partial z^*},
\end{aligned}$$

and its inverse (known as "Wirtinger derivatives")

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right), \frac{\partial}{\partial z^*} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right).$$

Observe that $\Delta z^*/\Delta z$ takes different values depending on which direction we take the respective, $\Delta z, \Delta z^* \to 0$ limit in the complex plain. Therefore to ensure that the derivative, $f'(z)$, is well defined at any z, one needs to require that

$$\frac{\partial f}{\partial z^*} = 0, \tag{2.8}$$

i.e. that f does not depend on z^*. It is straightforward to check that the "independence of the complex conjugate" Eq. (2.8) is equivalent to Eq. (2.6). □

Definition 2.2.3 ((Complex) Analyticity)**.** A function $f(z)$ is called (a) analytic (or holomorphic) at a point, z_0, if it is differentiable (as a complex function) in a neigborhood of z_0; (b) analytic in a region of the complex plane if it is analytic at each point of the region.

Exercise 2.4. The isolines for a function, $f(x, y) = u(x, y) + iv(x, y)$, are defined to be the curves, $u(x, y) = \text{const}$ and $v(x, y) = \text{const}'$. Show that the iso-lines of an analytic function always cross at a right angle.

Example 2.2.4. Let $f(z) = u(x, y) + iv(x, y)$ be analytic. Given that $u(x, y) = x + x^2 - y^2$ and $f(0) = 0$, find $v(x, y)$.

Solution. We start by utilizing the Cauchy-Riemann conditions:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}\ , \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}.$$

This gives us the two differential equations

$$\frac{\partial v}{\partial y} = 2x + 1\ , \quad \frac{\partial v}{\partial x} = 2y,$$

$$v(x, y) = 2xy + y + C_1(x)\ , \quad v(x, y) = 2xy + C_2(y).$$

Based on these two solutions, $C_1(x) = k$, for some constant k, and $C_2(y) = y + k$. Given the initial condition $f(0) = 0$, $v(0, 0) = 0$, and therefore $k = 0$. So, we find the solution, $v(x, y) = 2xy + y$. Therefore

$$f(z) = (x + x^2 - y^2) + i(2xy + y).$$

Exercise 2.5. Let $f(z) = u(x, y) + iv(x, y)$ be analytic. Given that, $v(x, y) = -2xy$ and $f(0) = 1$, find $u(x, y)$.

Examples of functions that are *not* analytic

Example 2.2.5. Determine whether (and where) $f(z) = z^*$ is analytic and compute its derivative where it exists.

Solution. Recall: If $z = x + iy$, then $z^* := x - iy$. We first compute u_x, v_y, u_y and v_x, and determine whether (and where) they are continuous.

$$u_x = 1 \qquad v_y = -1$$
$$u_y = 0 \qquad v_x = 0$$

We confirm that the partial derivatives are continuous everywhere in $\mathbb{C}$. We now check the Cauchy-Riemann conditions and find that they are *not* satisfied anywhere in $\mathbb{C}$ because $u_x = 1 \neq -1 = v_y$. Intuitively, the complex conjugate fails to be analytic because analytic functions can be locally approximated by rotations and stretches of the complex plane, whereas the complex conjugate function is a reflection.

Example 2.2.6. Determine whether (and where) $f(z) = z^{1/2}$ is analytic and compute its derivative where it exists.

Solution. We leave it to the reader to apply the chain rule and the trigonometic identity $\sin^2(\theta) + \cos^2(\theta) = 1$ to verify that the Cauchy-Riemann equations in polar coordinates are

$$\frac{\partial u}{\partial r} = \frac{1}{r}\frac{\partial v}{\partial \theta}$$
$$\frac{\partial v}{\partial r} = -\frac{1}{r}\frac{\partial u}{\partial \theta}$$

We compute the relevant partial derivatives of $z^{1/2}$ and observe that they are *not* defined at $z = 0$. We also observe that they cannot be continuous at a branch cut because branches of $z^{1/2}$ are not continuous across the branch cut. The Cauchy-Riemann conditions are satisfied everywhere else. In conclusion, a branch of $z^{1/2}$ is analytic in any region of $\mathbb{C} \setminus \{0\}$ that does not contain the branch cut. We leave it to the reader to show that the derivative in polar representation is given by $f'(z) = e^{-i\theta}(u_r + iv_r)$. For our example, this gives $f'(z) = \frac{1}{2}r^{-1/2}e^{-i\theta/2} = \frac{1}{2}z^{-1/2}$.

Example 2.2.7. Determine whether (and where) the function $f(z) = 1/z$ is analytic.

Solution. Note that f is not defined at $z = 0$ and that $\lim_{z\to 0} f(z)$ does not exist. Rationalize the denominator to write $f(z) = x/(x^2+y^2) - iy/(x^2+y^2)$. The relevant partial derivatives are:

$$u_x = \frac{y^2 - x^2}{x^2 + y^2} \qquad v_y = \frac{y^2 - x^2}{x^2 + y^2}$$
$$u_y = \frac{-2xy}{x^2 + y^2} \qquad v_x = \frac{2xy}{x^2 + y^2}$$

The partial derivatives exist and are continuous everywhere (except $z = 0$) and the the Cauchy-Riemann conditions are satisfied on $\mathbb{C} \setminus \{0\}$. We evaluate the derivative $f'(z)$ and observe that $\lim_{z\to 0} f'(z)$ does not exist. We say that f has a *simple pole* at $z = 0$ because $(z - 0)f(z)$ is analytic in a neighborhood of 0. We will revisit this in section 2.2.5.

Example 2.2.8. Determine whether (and where) the functions (a) $\exp(z)$, (b) $z \exp(\bar{z})$ and (c) $(\exp(z) - 1)/z$ are analytic and compute their derivatives where they exist.

The Cauchy-Riemann theorem 2.2.2 has a couple of other complementary, geometrical and from the world of partial differential equations, interpretations discussed below.

"Geometry" of Complex: Conformal Mapping

Let us now make a detour into the subject of conformal map describing a function of two variables, x, y, that locally preserves angles, but not necessarily lengths. We will see, and quite remarkably, that the the rich family of conformal functions (maps) can be analyzed based solely on the Cauchy-Riemann condition (2.6).

Indeed, the Cauchy-Riemann conditions (2.6) can be restated in the following compact form

$$i\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y}.$$

Then the Jacobian matrix of the function $f : \mathbb{R}^2 \to \mathbb{R}^2$, i.e. of the $(x, y) \to (u, v)$ map is

$$J = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix} = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ -\frac{\partial u}{\partial y} & \frac{\partial u}{\partial x} \end{pmatrix}.$$

Geometrically, the off-diagonal (skew-symmetric) part of the matrix represents rotation and the diagonal part of the matrix represents stretching/re-scaling. The Jacobian of a function $f(z)$ takes infinitesimal line segments at the intersection of two curves in z and rotates them to the corresponding segments in $f(z)$. Therefore, a function satisfying the Cauchy-Riemann equations, with a nonzero derivative, preserves the angle between curves in the plane. Transformations corresponding to such functions and functions themselves are called *conformal*. That is, the Cauchy-Riemann equations are not only conditions for the function analyticity, but are also conditions for a function to be conformal.

The following famous theorem of the complex analysis builds a strong theoretical foundation to the conformal maps. (It is due to Bernard Riemann, who has stated it in his Ph.D. thesis in 1851. First proof of the theorem was published in 1912 by Constantin Carathéodory.)

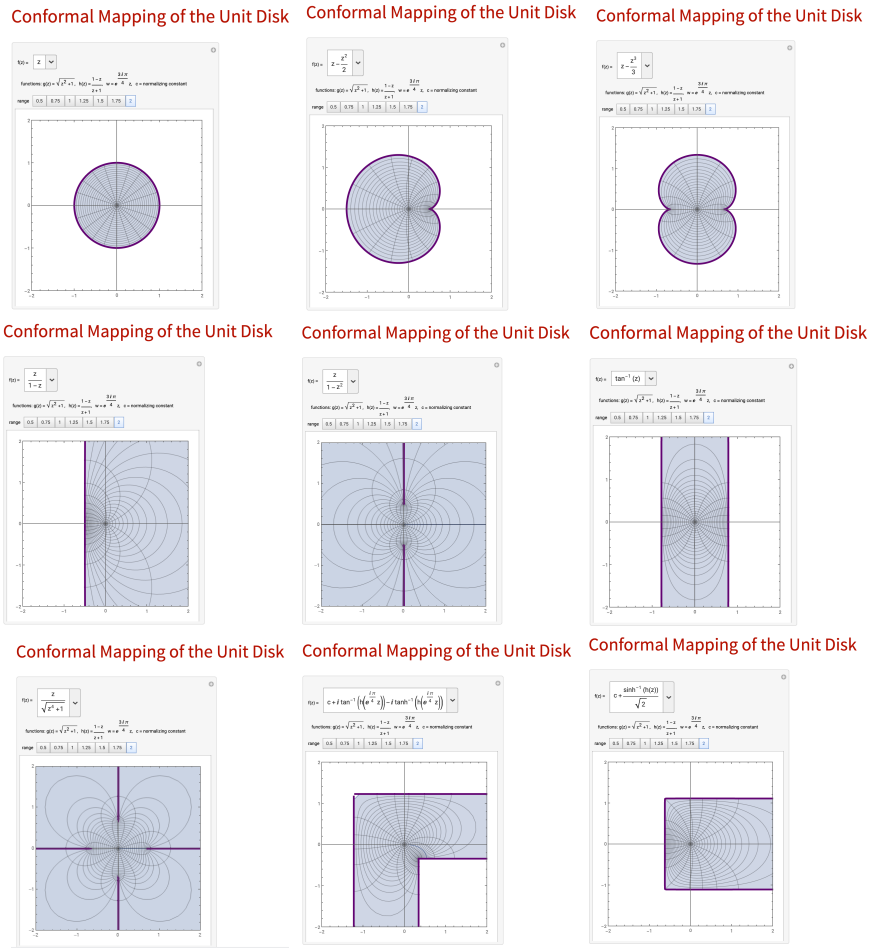


Figure 2.9: Exemplary functions (maps) from unit disk. Screenshots from `https://demonstrations.wolfram.com/ConformalMappingOfTheUnitDisk/`.

Theorem 2.2.9 (Riemann Mapping Theorem)**.** If A is a non-empty single-connected open subset of $\mathbb{C}$ which is not entire $\mathbb{C}$, $A \subset \mathbb{C}$, than there exists a holomorphic (complex analytic) function $f(z)$ mapping A to the unit disk, $D \equiv \{s \in \mathbb{C} \ : \ |z| < 1\}$, $f : A \to D$. Moreover $f^{-1} : D \to A$ is also holomorphic.

The theorem allows to build various conformal maps from one single-connected domain to another by reducing the problem to two, each mapping the respective domain to the unit disk.

It is useful for developing geometrical intuition to consider conformal maps associated with elementary functions. See a number of illustrations in Fig. 2.9. Notice, however that even relatively simple Riemann mappings, such as a map from the unit disk to the interior of a square, may not be expressible in terms of elementary functions and in general one

needs to rely on approximate numerical methods to build the maps. (See ConformalMaps Julia package `https://github.com/sswatson/ConformalMaps.jl` for approximating the Riemann map from a simply connected planar domain to a disk.)

"Physics" of Complex: Harmonic functions

Here we will make a fast jump to the end of the semester where Partial Differential Equations (PDEs) will be discussed in details. Consider solution of the Laplace equation in two dimensions

$$(\partial_x^2 + \partial_y^2) f(x,y) = 0. \tag{2.9}$$

Eq. (2.9) defines the so-called Harmonic functions. We do it now, while studying complex calculus, because, and quite remarkably, an arbitrary analytic function is a solution of Eq. (2.9). This statement is a straightforward corollary of the Cauchy-Riemann theorem (2.2.2). To see it we recall that, $f = u + iv$, and use the following set of transformation following from the Cauchy-Riemann conditions (2.6) also assuming that the function, $f(z)$, is analytic at z (which allows us to differentiate it one more time with respect to x and y)

$$\begin{cases} u_x = v_y \\ u_y = -v_x \end{cases} \Rightarrow \begin{cases} u_{xx} = v_{xy} \\ u_{yy} = -v_{xy} \end{cases} \Rightarrow u_{xx} + u_{yy} = 0$$

$$\begin{cases} u_x = v_y \\ u_y = -v_x \end{cases} \Rightarrow \begin{cases} u_{xy} = v_{yy} \\ u_{xy} = -v_{xx} \end{cases} \Rightarrow v_{xx} + v_{yy} = 0$$

The descriptor "harmonic" in the name harmonic function originates from a point on a taut string which is undergoing periodic motion which is pleasant-sounding, thus coined by ancient Greeks harmonic (!). This type of motion can be written in terms of sines and cosines, functions which are thus referred to as harmonics. Fourier analysis, which we will turn our attention to soon, involves expanding periodic functions on the unit circle in terms of a series over these harmonics. These functions satisfy Laplace equation and over time "harmonic" was used to refer to all functions satisfying Laplace equation.

Laplace equation, and thus harmonic functions, arise most prominently in mathematical physics, in particular in electromagnetism (electrostatics) and fluid mechanics (hydrostatics).

For example, in electrostatics it describes distribution of electrostatic potential within the planar domain cut in a metal – domain which is free of charge, i.e. free of singularities. On the other hand, singular points of the harmonic functions above are expressed as "point charges" and/or continuously distributed "charge densities". Placing a point charge at the origin, $z = x + iy = 0$, results in the solution of the Laplace equation (2.9), which

is singular, i.e. non-analytic at the origin. We will see later in the PDE part of the course, that the analytic function correspondent to a point charge placed at the origin is, $f(z) = C \log z$, where C is a constant related to the value of the charge, and then, $\mathrm{Re}(f(z)) = u = C \log r = (C/2) \log(x^2 + y^2)$, is the corresponding electrostatic potential.

The family of the harmonic (complex) functions is rich. Each harmonic function which satisfies Eq. (2.9) will yield another harmonic function when multiplied by a constant, rotated, and/or has a constant added. The inversion of each function will yield another harmonic function which has singularities which are the images of the original singularities in a spherical "mirror". Also, the sum of any two harmonic functions will yield another harmonic function.

2.2.2 Integration along Contours

Complex integration is defined along an oriented contour C in the complex plane.

Definition 2.2.10 (Complex Integration)**.** Let $f(z)$ be analytic in the neighborhood of a contour C. The integral of $f(z)$ along C is

$$\int_C f(z)\, dz := \lim_{n\to\infty} \sum_{k=0}^{n-1} f(\zeta_k)(\zeta_{k+1} - \zeta_k), \tag{2.10}$$

where for each n, $(\zeta_k | k = 0, \cdots, n)$ describes an ordered sequence of points along the path breaking it into n intervals such that $\zeta_0 = a$, $\zeta_n = b$ and $\max_k |\zeta_{k+1} - \zeta_k| \to 0$ as $n \to \infty$.

Remark. It is now time to utilize parametrization of the complex functions discussed earlier in the course. Let $z(t)$ with $a \le t \le b$ be a parameterization of C, then definition 2.2.10 is equivalent to the Riemann integral of $f(z(t))z'(t)$ with respect to t. Therefore,

$$\int_C f(z)\, dz = \int_a^b f(z(t))\, z'(t)\, dt \tag{2.11}$$

Example 2.2.11. In example 2.1.13 we evaluated the functions $f_1(z) = \sin z$, $f_2(z) = \exp(z + 1)$, and $f_3(z) = z^2$ along the parameterized curves described in example 2.1.15. Now compute (a) $\int_{C_1} f_1(z)\, dz$, (b) $\int_{C_2} f_2(z)\, dz$, and (c) $\int_{C_3} f_3(z)\, dz$ where C_1 is the the vertical line segment from $\pi/2 - iM$ to $\pi/2 + iM$, C_2 is the ray segment extending from the point $z = -1$ and to the point $\sqrt{3}i$, and C_3 is the circular arc of radius ε centered at 0.

Solution.

(a) Let $z = \pi/2 + is$, then $dz = ids$ for $-M < s < M$.

$$\int_{C_1} \sin(z)dz = \int_{-M}^{+M} \frac{e^{i\pi/2-s} - e^{-i\pi/2+s}}{2i} ids = \int_{-M}^{+M} i\cosh(s)ds = 2i\cosh(M).$$

(b) Let $z = -1 + \rho e^{i\pi/3}$ for $0 \le \rho \le 2$. Then $dz = e^{i\pi/3}d\rho$.

$$\int_{C_2} e^{z+1}dz = \int_0^2 e^{\rho e^{i\pi/3}} e^{i\pi/3} d\rho = \left| e^{\rho e^{i\pi/3}} \right|_0^2 = e^{2\cos(\pi/3)+i2\sin(\pi/3)} - 1$$
$$= e^1\cos(\sqrt{3}) - 1 + ie^1\sin(\sqrt{3}).$$

(c) Let $z = \epsilon e^{i\tau}$ for $0 \le \tau < 2\pi$, then $dz = i\epsilon e^{i\tau}d\theta$. Therefore,

$$\int_{C_3} z^2\,dz = \int_0^{2\pi} \left(\epsilon e^{i\tau}\right)^2 i\epsilon e^{i\tau} d\theta = \left| \tfrac{1}{3}\epsilon^3 e^{3i\theta} \right|_0^{2\pi} = \tfrac{1}{3}\varepsilon^3 e^{6\pi i} - \tfrac{1}{3}\varepsilon^3 e^0 = 0.$$

Exercise 2.6. Let C_+ and C_- represent the upper and lower unit semi-circles centered at the origin and oriented from $z = -1$ to $z = 1$. Find the integrals of the functions (a) z; (b) z^2; (c) $1/z$; and (d) $\sqrt{z}$ along C_+ and C_-. For $\sqrt{z}$, use the branch where z is represented by $re^{i\theta}$ with $0 \le \theta < 2\pi$.

Example 2.2.12. Let C be the circular closed contour of radius R centered at the origin. Show that

$$\oint_C \frac{dz}{z^m} = 0, \quad \text{for } m = 2, 3, \ldots \tag{2.12}$$

by parameterizing the contour in polar coordinates.

Solution. One possible parameterization of the contour is $z(\theta) = Re^{i\theta}$ for $0 \le \theta < 2\pi$. Therefore, $dz = iRe^{i\theta}d\theta$. Changing the integral to polar coordinates gives:

$$\oint_C \frac{dz}{z^m} = \int_0^{2\pi} \frac{iRe^{i\theta}}{(Re^{i\theta})^m} d\theta.$$

Because $m = 2, 3, 4, ...$, we know that $m - 1 > 0$. Therefore,

$$\int_0^{2\pi} \frac{iRe^{i\theta}}{(Re^{i\theta})^m} d\theta = \frac{i}{R^{m-1}} \int_0^{2\pi} \frac{1}{(e^{i\theta})^{m-1}} d\theta = \frac{i}{R^{m-1}} \left(\frac{i}{m-1} e^{-i\theta(m-1)} \Big|_0^{2\pi} \right) = 0$$

A simple, but useful, continuation of this example would be confirming that the integral is not 0 when $m = 1$.

Example 2.2.13. Use numerical integration to approximate the integrals in the examples above and verify your results.

2.2.3 Cauchy's Theorem

In general the integral along a path in the complex plane depends on the entire path and not only on the position of the end points. The following fundamental question arrives naturally: is there a condition which makes the integral dependent only on the end points of the path? The question is answered by the following famous theorem.

Theorem 2.2.14 (Cauchy's Theorem, 1825)**.** If $f(z)$ is analytic in a simply connected region $\mathcal{D}$ of the complex plane then for all paths, C, lying in this region and having the same end points, the integral $\int_C f(z)\,dz$ has the same value.

A more compact way of stating the theorem is to say that integrals of analytic functions are *path independent*.

It is important to recognize that for Cauchy's theorem to hold in the case of the multi-valued functions one needs the integrand to be a single-valued function. The cuts introduced in the preceding section are required for exactly this reason – to force the integration path to stay within a single branch of a multi-valued function and thus to guarantee analyticity (differentiability) of the function along the path.

The same theorem can be restated in the following form.

Theorem 2.2.15 (Cauchy's Theorem (closed contour version))**.** Let $f(z)$ be analytic in a simply connected region $\mathcal{D}$ and C be a closed contour that lies in the interior of $\mathcal{D}$. Then the integral of f along C is equal to zero: $\oint_C f(z)\,dz = 0$.

To make the transformation from the former formulation of Cauchy's formula to the latter one, we need to consider two paths connecting two points of the complex plain. From Eq. (2.10), we see that paths are oriented and that changing the direction of the path changes the value of the integral by a factor of -1. Therefore, of the two paths considered, one needs to reverse its direction, then leading us to a closed contour formulation of Cauchy's theorem.

Let us now sketch the proof of the closed contour version of Cauchy's theorem. Consider breaking the region of the complex plane bounded by the contour C into small squares with the contours C_k, as well as the original contour C, oriented in the positive direction (counter-clockwise). Then

$$\oint_C dz f(z) = \sum_k \oint_{C_k} f(z) dz, \tag{2.13}$$

where we have accounted for the fact that integrals over the inner sides of the small contours cancel each other, as two of them (for each side) are running in opposite directions. Next,

pick inside a C_k contour a point, z_k, and then approximate, $f(z)$, expanding it in the Taylor series around z_k,

$$f(z) = f(z_k) + f'(z_k)(z - z_k) + O\left(\Delta^2\right) \tag{2.14}$$

where with Δ-squares, the length of C_k is at most 4Δ, and we have at most $(L/\Delta)^2$ small squares. Substituting Eq. (2.14) into Eq. (2.13) one derives

$$\oint\limits_{C_k} dz f(z) = f(z_k) \oint\limits_{C_k} dz + f'(z_k) \oint\limits_{C_k} dz(z - z_k) + \oint\limits_{C_k} dz O\left(\Delta^2\right) = 0 + 0 + \Delta^3. \tag{2.15}$$

Summing over all the small squares bounded by C one arrives at the estimate $\Delta \to 0$ in the $\Delta \to 0$ limit. $\square$.

Disclaimer: We have just used discretization of the integral. When dealing with integrations of functions in the rest of the course we will always discuss it in the sense of a limit, assuming that it exists, and not really breaking the integration path into segments. However, if any question on the details of the limiting procedure surfaces one should get back to the discretization and analyze respective limiting procedure sorely.

One important consequence of the Cauchy's theorem (there will be more discussed in the following) is that all integration rules known for standard, "interval", integrals apply to the contour integrals. This is also facilitated by the following statement.

Theorem 2.2.16 (Triangle Inequality)**.** (A: From Euclidean Geometry) $|z_1 + z_2| \leq |z_1| + |z_2|$, also with equality iff (if and only if) z_1 and z_2 lie on the same ray from the origin. (B: Integral over Interval) Suppose $g(t)$ is a complex valued function of a real variable, defined on $a \leq t \leq b$, then

$$\left| \int\limits_a^b dt g(t) \right| \leq \int\limits_a^b dt |g(t)|, \tag{2.16}$$

with equality iff (i.e. if and only if) the values of $g(t)$ all lie on the same ray from the origin. (Integral over Curve/Path) For any function $f(z)$ and any curve γ, we have

$$\left| \int\limits_\gamma f(z) dz \right| \leq \int\limits_\gamma |f(z)||dz|, \tag{2.17}$$

where $dz = \gamma'(t)dt$ and $|dz| = |\gamma'(t)|dt$.

Proof. We take the "Euclidean" geometry version (A) of the statement, extended to the sum of complex numbers, as granted and give a brief sketch of proofs for the integral formulations. The interval version (B) of the triangular inequality follows by approximating the integral

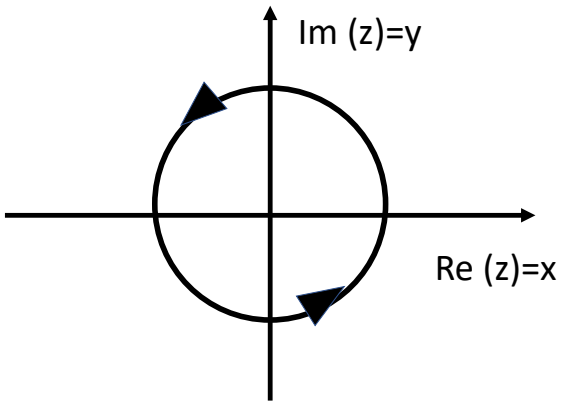


Figure 2.10

as a Riemann sum

$$|g(t)dt| \approx \left|\sum g(t_k)\Delta t\right| \leq \sum |g(t_k)|\Delta t \approx \int_a^b |g(t)|dt,$$

where the middle inequality is just the standard triangular inequality for sums of complex numbers. The contour version (C) of the Theorem follows immediately from the interval version

$$\int_\gamma f(z)dz = \left|\int_a^b f(\gamma(t))\gamma'(t)dt\right| \leq \int_a^b |f(\gamma(t))||\gamma'(t)|dt = \int_\gamma |f(z)||dz|.$$

□

2.2.4 Cauchy's Formula

Recall from definition 2.1.12 that a curve is called *simple* if it does not intersect itself, and is called a *contour* if it is piece-wise smooth.

Theorem 2.2.17 (Cauchy's formula, 1831)**.** Let $f(z)$ be analytic on and interior to a simple closed contour C. Then,

$$f(z) = \frac{1}{2\pi i}\int_C \frac{f(\zeta)d\zeta}{\zeta - z}. \tag{2.18}$$

To illustrate Cauchy's formula consider the simplest, and arguably most important, example of an integral over complex plane, $I = \oint dz/z$. For the integral over closed contour

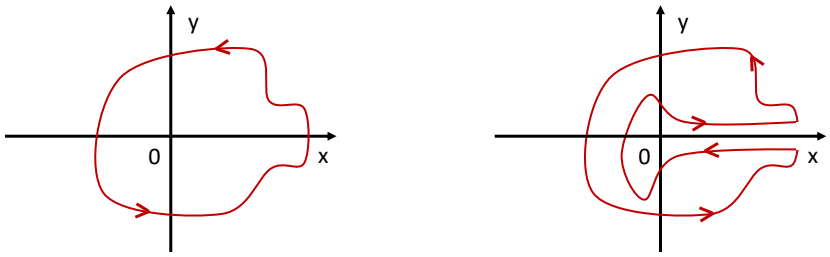


Figure 2.11

shown in Fig. (2.10a), we parameterize the contour explicitly in polar coordinates and derive

$$I = \oint \frac{dz}{z} = \int_0^{2\pi} \frac{r d\exp(i\theta)}{r\exp(i\theta)} = \int_0^{2\pi} \frac{r\exp(i\theta) i d\theta}{r\exp(i\theta)} = i\int_0^{2\pi} d\theta = 2\pi i. \tag{2.19}$$

The integral is not zero.

Next, recall that for the respective standard indefinite integral, $\int dz/z = \log z$. This formula is very naturally consistent with both Eq. (2.19) and with the fact that $\log(z)$ is a multivariate function. Indeed, consider the integral over a path between two points of a complex plain, e.g. $z = 1$ and $z = 2$. We can go from $z = 1$ to $z = 2$ straight, or can do it, for example first making a counter-clockwise turn around 0. We can generalize and do it clockwise and also making as many number of points we want. It is straightforward to check that the integral depends on how many times and in which direction we go around 0. The answers will be different by the result of Eq. (2.19), i.e. $2\pi i$ multiplied by an integer, however it will not depend on the path.

Example 2.2.18. Compute, compare and discuss the difference (if any) between values of the integral $\oint dz/z$ over two distinct paths shown in Fig. (2.11).

Solution. Firstly, note that $1/z$ is not analytic at $z = 0$, and therefore neither contour can be deformed to the other without passing through the non-analytic point. Both curves are simple and closed, therefore Cauchy's formula applies. The curve on the left, C_1 contains $z = 0$, and makes one full turn around the origin, and therefore will be equal to the curve in Figure 2.4, so, $\oint_{C_1} \frac{dz}{z} = 2\pi i$. The curve on the right, C_2, does not contain $z = 0$, and makes a full turn counter-clockwise around the origin, as well as a clockwise turn around the origin on the "inner" part of the curve. So, $\oint_{C_2} \frac{dz}{z} = 0$.

The "small square" construction used above to prove the closed contour version of Cauchy's Theorem, i.e. Theorem 2.2.15, is a useful tool for dealing with integrals over

awkward (difficult for direct computation) paths around singular points of the integrand. However, it should not be thought that all the integrals will necessarily be zero. Consider

$$m = 2, 3, \cdots : \quad \oint \frac{dz}{z^m},$$

where the integral is singular at $z = 0$. The respective indefinite integral (what is sometimes called the "anti-derivative") is $z^{-m+1}/(1-m) + C$, where C is constant. Observe that the indefinite integral is a single-valued function and thus its integral over a closed contour is zero. (Notice that if $m = 1$ the indefinite integral is a multi-valued function within the domain surrounding $z = 0$.)

Cauchy's formula can be extended to higher derivatives

Theorem 2.2.19 (Cauchy's formula for derivatives, 1842)**.** Under the same conditions as in Theorem 2.2.17, higher derivatives are

$$f^{(n)}(z) = \frac{n!}{2\pi i} \int_C \frac{f(\zeta)d\zeta}{(\zeta - z)^{n+1}}. \tag{2.20}$$

Theoretical Implications of Cauchy's Theorem & Cauchy's Formulas

Cauchy's theorem and formulas have many powerful and far reaching consequences.

Theorem 2.2.20. Suppose $f(z)$ is analytic on a region A. Then, f has derivatives of all orders.

Proof. It follows directly from Cauchy's formula for derivatives, Theorem 2.2.19 – that is we have an explicit formula for all the derivatives, so, in particular, the derivatives all exist. □

Theorem 2.2.21 (Cauchy Inequality.)**.** Let C_R be the circle $|z - z_0| = R$. Assume that $f(z)$ is analytic on C_R and its interior, i.e. on the disk $|z - z_0| \leq R$. Finally let $M_R = \max |f(z)|$ over z on C_R. Then

$$\forall n = 1, 2, \cdots : \quad |f^{(n)}(z_0)| \leq \frac{n! M_R}{R^n}.$$

Exercise 2.7. Prove the Cauchy's Inequality Theorem utilizing Theorem 2.2.19. Provide an alternative argument for the Theorem validity on examples of $\exp(z)$ and $\cos(z)$ using a circle that is centered at the origin (you are expected to argue informally and without reference to Theorem 2.2.19 why the inequality holds).

Theorem 2.2.22 (Liouville Theorem.)**.** If $f(z)$ is entire, i.e. analytic at all finite points of the complex plane $\mathbb{C}$, and bounded then f is constant.

Proof. For any circle of radius R around z_0 the Cauchy's inequality (Theorem 2.2.21) states that $f'(z) \leq M/R$, but R can be arbitrarily large, thus $|f'(z_0)| = 0$ for every $z_0 \in \mathbb{C}$. And since the derivative is 0, the function itself is constant. □

Note that $P(z) = \sum_{k=0}^{n} a_k z^k$, $\exp(z)$, $\cos(z)$ are entire but not bounded.

Theorem 2.2.23 (Fundamental Theorem of Algebra)**.** Any polynomial P of degree $n \geq 1$, i.e. $P(z) = \sum_{k=0}^{n} a_k z^k$, has exactly n roots (solutions of $P(z) = 0$).

Proof. The prove consists of two parts. First, we want to show that $P(z)$ has at least one root. (See example below.) Second, assume that P has exactly n roots. Let z_0 be one of the roots. Factor, $P(z) = (z - z_0)Q(z)$. $Q(z)$ has degree $n - 1$. If $n - 1 > 0$, then we can apply the result to $Q(z)$. We can continue this process until the degree of Q is 0. □

Example 2.2.24. Prove that $P(z) = \sum_{k=0}^{n} a_k z^k$ has at least one root.

Solution. We provide a hint and not the full solution: prove by contradiction and utilize the Liouville Theorem 2.2.22.

Theorem 2.2.25 (Maximum modulus principle (over disk))**.** Suppose $f(z)$ is analytic on the closed disk, C_r, of radius r centered at z_0, i.e. the set $|z - z_0| \leq r$. If $|f|$ has a relative maximum at z_0 than $f(z)$ is constant in C_r.

In order to prove the Theorem we will first prove the following statement.

Theorem 2.2.26 (Mean value property)**.** Suppose $f(z)$ is analytic on the closed disk of radius r centered at z_0, i.e. the set $|z - z_0| \leq r$. Then,

$$f(z_0) = \frac{1}{2\pi} \int_0^{2\pi} d\theta f\left(z_0 + r\exp(i\theta)\right).$$

Proof. Call C_r the boundary of the $|z - z_0| \leq r$ set, and parameterize it as $z_0 + re^{i\theta}$, $\quad 0 \leq \theta \leq 2\pi$. Then, according to Cauchy's formula,

$$f(z_0) = \frac{1}{2\pi i} \int_{C_r} \frac{f(z)dz}{z - z_0} = \frac{1}{2\pi i} \int_0^{2\pi} d\theta \frac{f(z_0 + re^{i\theta})}{re^{i\theta}} ire^{i\theta} = \frac{1}{2\pi} \int_0^{2\pi} d\theta f(z_0 + re^{i\theta}).$$

□

Now back to the Theorem 2.2.25. To sketch the proof we will use both the mean value property Theorem 2.2.26 and the triangle inequality Theorem 2.2.16. Since z_0 is a relative

maximum of $|f|$ on C_r we have $|f(z) \leq |f(z_0|)$ for $z \in C_r$. Therefore by the mean value property and the triangle inequality one derives

$$\begin{aligned}
|f(z_0)| &= \left| \frac{1}{2\pi} \int_0^{2\pi} d\theta f(z_0 + re^{i\theta}) \right| \quad \text{(mean value property)} \\
&\leq \frac{1}{2\pi} \int_0^{2\pi} d\theta |f(z_0 + re^{i\theta})| \quad \text{(triangle inequality)} \\
&\leq \frac{1}{2\pi} \int_0^{2\pi} d\theta |f(z_0)|, \quad (|f(z_0 + re^{i\theta})| \leq |f(z_0)|, \quad \text{i.e. } z_0 \text{ is a local maximum}) \\
&= |f(z_0)|
\end{aligned}$$

Since we start and end with $f(z_0)$, all inequalities in the chain are equalities. The first inequality can only be equality if for all θ, $f(z_0 + re^{i\theta})$ lies on the same ray from the origin, i.e. have the same argument or equal to zero. The second inequality can only be an equality if all $|f(z_0 + re^{i\theta}| = |f(z_0)|$. Thus, combining the two observations, one gets that all $f(z_0 + re^{i\theta})$ have the same magnitude and the same argument, i.e. all the same. Finally, if $f(z)$ is constant along the circle and $f(z_0)$ is the average of $f(z)$ over the circle then $f(z) = f(z_0)$, i.e. f is constant on C_r. □

Two remarks are in order. First, based on the experience so far (starting from Theorem 2.2.22) it is plausible to expect that Theorem 2.2.25 generalizes from a disk C_r to any single-connected domain. Second, one also expects that the maximum modulus can be achieved at the boundary of a domain and then the function is not constant within the domain. Indeed, consider example of $\exp(z)$ on the unit square, $0 \leq x, y \leq 1$. The maximum, $|\exp(x + iy)| = \exp(x)$, is achieved at $x = 1$ and arbitrary y, $0 \leq y \leq 1$, i.e. at the boundary of the domain. These remarks and the example suggest the following extension of the Theorem 2.2.25.

Theorem 2.2.27 (Maximum modulus principle (general))**.** Suppose $f(z)$ is analytic on A, which is a bounded, connected, open set, and it is continuous on $\bar{A} = A \cup \partial A$, where ∂A is the boundary of $\bar{A}$. Then either $f(z)$ is a constant or the maximum of $|f(z)|$ on $\bar{A}$ occurs on ∂A.

Proof. Here is a sketch of the proof. Let us cover A by disks which are laid such that their centers form a path from the value where $f(z)$ is maximized to any other points in A, while being totally contained within A. Existence of a maximum value of $|f(z)|$ within A implies, according to Theorem 2.2.25 applied to all the disks, that all the values of $f(z)$ in

the domain are the same, thus $f(z)$ is constant within A. Obviously the constancy of $f(z)$ is not required if the maximum of $|f(z)|$ is achieved at δA. □

Example 2.2.28. Find the maximum modulus of $\sin(z)$ on the square, $0 \leq x, y \leq 2\pi$.

Solution.

$$\sin(z) = \sin(x + iy) = \sin(x)\cosh(y) + i\cos(x)\sinh(y)$$

We want to find the maximum modulus of $\sin(z)$

$$|\sin(x + iy)| = \sqrt{\sin^2(x)\cosh^2(y) + \cos^2(x)\sinh^2(y)}.$$

Using the identities, $\cos^2(x) = 1 - \sin^2(x)$, and $\cosh^2(x) - \sinh^2(x) = 1$, we get

$$|\sin(z)| = \sqrt{\sin^2(x)\cosh^2(y) + (1 - \sin^2(x))\sinh^2(y)} = \sqrt{\sinh^2(y) + \sin^2(x)}.$$

This is maximized when $\sin^2(x) = 1$, so $x = \frac{n\pi}{2}$ where $n = 1, 2, 3, \cdots$, and $y = 2\pi$, as this is the maximum value of y in our square. So,

$$\max(|\sin(z)|) = \sinh^2(2\pi) + 1 = 71688.328285.$$

$z = \frac{n\pi}{2} + i2\pi$, is on the boundary of our square, which satisfies the maximum modulus principle.

2.2.5 Laurent Series

The *Laurent series* of a complex function $f(z)$ about a point a is a representation of that function by a power series that includes terms of both positive and negative degree.

Theorem 2.2.29. A function $f(z)$ that is analytic on the annulus, $R_1 \leq |z - a| \leq R_2$, and within its interior may be represented by a power series, called a *Laurent Series*, that converges on the interior of the annulus:

$$f(z) = \sum_{k=-\infty}^{+\infty} c_k(z - a)^k. \tag{2.21}$$

The coefficients of the Laurent series are given by

$$c_k = \frac{1}{2\pi i} \oint_C \frac{f(z)}{(z - a)^{n+1}} dz, \tag{2.22}$$

where C is any contour that is contained within the annulus and circling a.

Suppose one needs to compute

$$\oint_C f(z)dz,$$

where the contour C circles $z = a$ in the positive (counter-clockwise) direction such that there are no singular points of $f(z)$ in the interior of C, except possibly at $z = a$. If we represent $f(z)$ by its Laurent series, then the only nonzero contribution will come from the $k = -1$ term

$$\oint_C f(z)dz = \oint_C \sum_{k=-\infty}^{\infty} c_k(z-a)^k dz = c_{-1} \oint \frac{dz}{z-a} = 2\pi i c_{-1}. \tag{2.23}$$

Definition 2.2.30. The coefficient corresponding to the $k = -1$ term plays such a significant role in contour integration that it deserved a special name – *residue* of f at $z = a$ – and is denoted by $c_1 = \text{Res}(f; a)$.

Notice, that if $f(z)$ has a simple pole at $z = a$, then

$$c_{-1} = \text{Res}(f, a) = \lim_{z \to a} \left(f(z)(z-a) \right). \tag{2.24}$$

2.3 Residue Calculus

2.3.1 Singularities and Residues

Definition 2.3.1 (Singularity)**.** Let $f : \mathbb{C} \to \mathbb{C}$ and consider $a \in \mathbb{C}$. If f is not analytic at a (meaning that $f'(a)$ does not exist) then we say that a is a *singular point* of f. If f is not analytic at a, but is analytic in the region $0 < |z - a| < R$, then we say that $a \in \mathbb{C}$ is an *isolated singular point* of f.

Definition 2.3.2 (Removable Singularity)**.** Let a be a singular point of a function f, and let c_k be the coefficients of the Laurent expansion of f about a. If $c_k = 0$ for all $k < 0$, then we say that a is a *removable singularity* of f. (Note that f would be analytic if f were redefined at a single point, a.)

Definition 2.3.3 (Simple Pole)**.** Let a be a singular point of a function f, and let c_k be the coefficients of the Laurent expansion of f about a. If $c_{-1} \neq 1$, but $c_k = 0$ for all $k < -1$, then we say that a is a *first order pole* or a *simple pole* of f. See Fig. 2.12(a) for an example of a simple pole.

If f has a simple pole at $z = a$, we can represent f in the form

$$f(z) = \frac{g(z)}{z-a}, \tag{2.25}$$

where $g(z)$ is analytic in a neighborhood of z with $g(a) = c_{-1} \neq 0$.

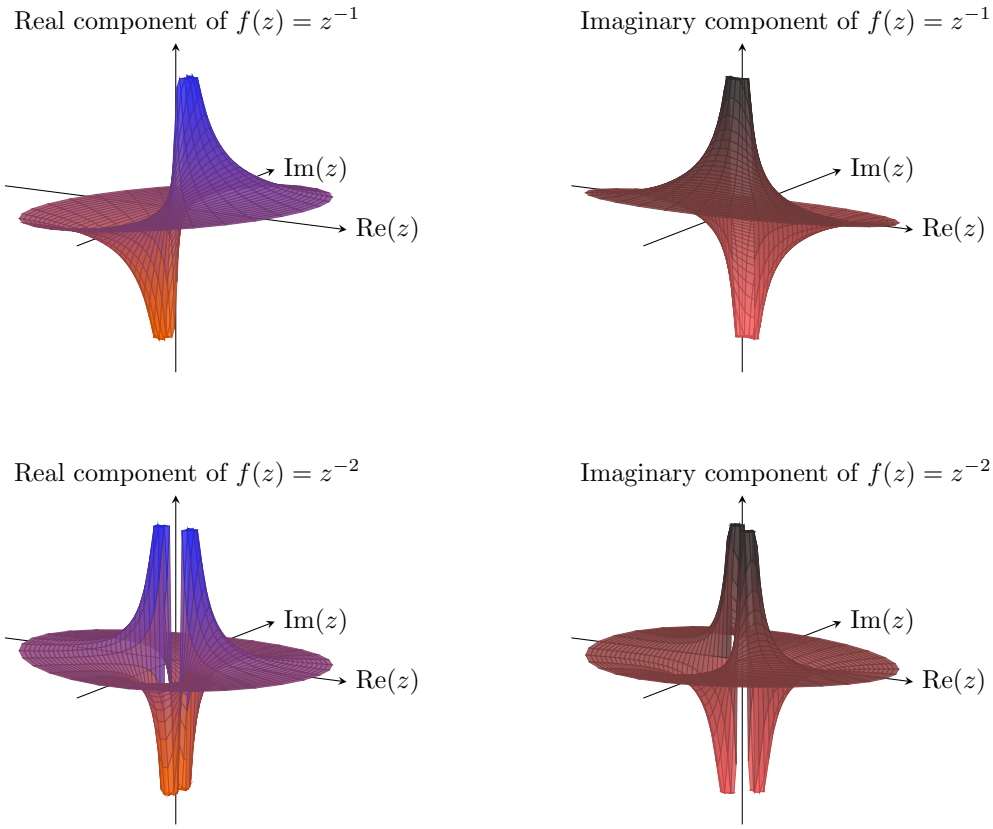


Figure 2.12: (a) Top row: The canonical example of a simple pole. The real (left) and imaginary (right) components of $z \mapsto z^{-1}$. (b) Bottom row: The canonical example of a double pole. The real (left) and imaginary (right) components of $z \mapsto z^{-2}$.

Definition 2.3.4 (Higher Order Pole)**.** Let a be a singular point of a function f, and let c_k be the coefficients of the Laurent expansion of f about a. If, for some positive N, $c_{-N} \neq 0$ but $c_k = 0$ for all $k < -N$, then we say that a is a *N-th order pole* of f. See Fig. 2.12(b) for an example of a double pole.

If f has an N-th order pole at $z = a$, we can represent f in the form

$$f(z) = \frac{g(z)}{(z-a)^N}, \tag{2.26}$$

where $g(z)$ is analytic in a neighborhood of z with $g(a) = c_{-N} \neq 0$.

Example 2.3.5. Find the removable singularity and give the order of the poles in the

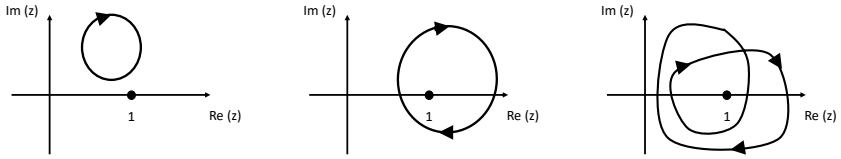


Figure 2.13: See Example 2.3.6.

following function

$$f(z) = \frac{z-1}{(z^4-1)(z+1)} \qquad \left(= \frac{z-1}{(z-1)(z+1)^2(z+i)(z-i)}\right).$$

Solution. Observe that $z = 1$, $z = -1$, $z = i$, and $z = -1$ are singular points of f because f is not defined at these points.

- $z = 1$: For $z \neq 1$, $f(z) = g(z)$ where $g(z) = \frac{1}{(z+1)^2(z+i)(z-i)}$ which is analytic in a neighborhood of $z = 1$. Therefore, $z = 1$ is a removable singularity of f.
- $z = i$: For $z \neq i$, $f(z) = \frac{g(z)}{z-i}$ where $g(z) = \frac{z-1}{(z-1)(z+1)^2(z+i)}$ which is analytic in a neighborhood of $z = 1$. Therefore, $z = i$ is a first order pole (or simple pole) of f.
- $z = -i$: It is similar to $z = i$.
- $z = -1$: For $z \neq -1$, $f(z) = \frac{g(z)}{(z+1)^2}$, where $g(z) = \frac{z-1}{(z-1)(z+i)(z-i)}$ which is analytic in a neighborhood of $z = -1$. Therefore, $z = -1$ is a second order pole (double pole) of f.

In summary, there is a removable singularity at $z = 1$, first-order poles at $z = \pm i$ and a second-order pole at $z = -1$.

Example 2.3.6. Use Cauchy's formula to compute

$$I = \oint \frac{\exp(z^2)dz}{z-1},$$

for three contour examples shown in the Fig. 2.13.

Solution. (a) The integrand is analytic everywhere within the domain surrounded by the contour, therefore $I = 0$. (b) The integrand has a single first-order pole within the domain surrounded by the contour at $z = 1$. Notice, that direction of the contour is negative (clockwise). therefore $I = -2\pi i \mathrm{Res}\left(\exp(z^2)/(z-1), 1\right) = -2\pi i \exp(z^2)|_{z=1} = -2\pi i$. (c) Since the only singularity of the integrand is at $z = 1$ the contour can be reduced to the contour in the case (b), however traveled twice. Therefore, $I = 2 * (-2\pi i) = -4\pi i$.

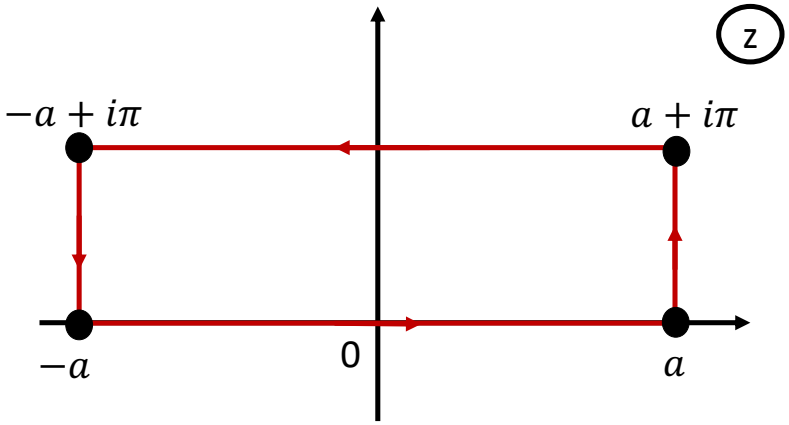


Figure 2.14: See Example 2.3.7.

Example 2.3.7. Use Cauchy's formula to compute

$$I = \oint \frac{dz}{\cosh z},$$

over the contour shown in Fig. 2.14, where a is a positive real.

Solution. To identify possible singularities of the integrand we need to solve $\cosh z_* = 0$, resulting in $z_* = i\pi(n + 1/2)$, where $n = 0, \pm 1, \cdots$. We observe that all the singularities are first-order poles. Only one of the poles, $z_* = i\pi/2$ is within the domain surrounded by the contour. Therefore, according to the Cauchy's formula, the residue formula (2.24), and L'Hôpital's rule,

$$I = 2\pi i \text{Res}\,(1/\cosh z, i\pi/2) = 2\pi i \lim_{z\to i\pi/2} \frac{z - i\pi/2}{\cosh z} = \frac{2\pi i}{\sinh(i\pi/2)} = \frac{2\pi}{\sin \pi/2} = 2\pi.$$

Exercise 2.8. Compute the integral $\oint dz/(e^z - 1)$ over the circle of radius 4 centered around $3i$.

2.3.2 Evaluation of Real-valued Integrals by Contour Integration

Example 2.3.8. Evaluate the following real-valued integral using contour integration:

$$I = \int_{-\infty}^{\infty} \frac{e^{ikx} dx}{x^2 + 1}.$$

Solution. Let $f : \mathbb{C} \to \mathbb{C}$ be given by $f(z) = e^{ikz}/(z^2 + 1)$. Observe that f has simple poles at $z = \pm i$. Let $C = C_1 \cup C_R$ be the contour in the complex plane show in Figure 2.15.

Consider the integral

$$\int_C \frac{e^{ikz}dz}{z^2+1} = \int_{C_1} \frac{e^{ikz}dz}{z^2+1} + \int_{C_R} \frac{e^{ikz}dz}{z^2+1}.$$

Our plan is to use the Cauchy's formula to show that the integral along C is $2\pi i\text{Res}(f;i)$, where $\text{Res}(f;i) = \lim_{z\to i} \frac{e^{ikz}(z-i)}{z^2+1} = \frac{e^{-k}}{2i}$, and then, with the correct parameterization, we can show that the integral along C_1 converges to I as $R \to 0$, and that the integral along C_R converges to 0 as $R \to \infty$.

First, evaluate the integral along C_1. Parameterize C_1 by $z = x + 0i$ for $-R < x < R$. Therefore, $dz = dx$.

$$\int_{C_1} \frac{e^{ikz}dz}{z^2+1} = \int_{-R}^{R} \frac{e^{ikz}dz}{z^2+1} = \int_{-R}^{R} \frac{e^{ikx}dx}{x^2+1} =$$

Observe that in the limit $R \to 0$, this is equivalent to the integral we must find.

Next evaluate the integral along C_2. Parameterize C_2 by $z = e^{Ri\theta} = R\cos(\theta) + iR\sin(\theta)$. Therefore, $dz = Rie^{Ri\theta}d\theta$. This gives

$$\int_{C_2} \frac{e^{ikz}dz}{z^2+1} = \int_0^{\pi} \frac{e^{ikR(\cos(\theta)+i\sin(\theta))}Rie^{Ri\theta}d\theta}{(Re^{i\theta})^2+1}.$$

We must consider what happens to the magnitude of the above integral as $R \to \pm\infty$.

$$\begin{aligned}
\left|\int_0^{\pi} \frac{e^{ikR(\cos(\theta)+i\sin(\theta))}Rie^{Ri\theta}d\theta}{(Re^{i\theta})^2+1}\right| &\le \int_0^{\pi} \left|\frac{e^{ikR(\cos(\theta)+i\sin(\theta))}Rie^{Ri\theta}}{(Re^{i\theta})^2+1}\right| d\theta \quad \text{(by triangle inequality)} \\
&\le R\int_0^{\pi} \left|\frac{e^{ikR\cos(\theta)-kR\sin(\theta)}}{R^2e^{2i\theta}+1}\right| d\theta \\
&\le R\int_0^{\pi} \left|\frac{e^{-kR\sin(\theta)}}{R^2e^{2i\theta}+1}\right| d\theta \quad \text{(because } \left|e^{ikR\cos(\theta)}\right| \le 1\text{)}.
\end{aligned}$$

Observe that for $R \to \infty$, we have $\left|\frac{1}{R^2 e^{2i\theta}+1}\right| \leq \frac{1}{R^2-1}$, because the term $R^2 e^{2i\theta}$ must lie between $-R^2$ and R^2.

$$\leq \frac{R}{R^2-1} \int_0^{\pi} \left|e^{-kR\sin(\theta)}\right| d\theta$$

$$\leq \frac{2R}{R^2-1} \int_0^{\pi/2} \left|e^{-kR\sin(\theta)}\right| d\theta$$

$$\leq \frac{2R}{R^2-1} \int_0^{\pi/2} e^{-kR\frac{2\theta}{\pi}} d\theta \quad \text{(because } \sin(\theta) \geq \tfrac{2\theta}{\pi} \text{ for all } \theta \in [0,\pi]\text{)}.$$

This final integral is one that we can evaluate.

$$\frac{2R}{R^2-1} \int_0^{\pi/2} e^{-kR\frac{2\theta}{\pi}} d\theta = \frac{2R}{R^2-1}\left(\frac{\pi(1-e^{-kR})}{2kR}\right) \to 0 \quad \text{as } R \to \infty.$$

All in all, we see that

$$\int_{C_2} \frac{e^{ikz}dz}{z^2+1} \to 0 \quad \text{as } R \to \infty.$$

From here, we use Cauchy Formula which gives us $\int_C \frac{e^{ikz}dz}{z^2+1} = 2\pi i Res(f,i) = (2\pi i)(e^{-k}/2i) = \pi e^{-k}$. This implies that our final answer is

$$I = \int_{-\infty}^{\infty} \frac{e^{ikx}dx}{x^2+1} = \frac{\pi}{e^k}.$$

Observe that Example 2.3.8 is an application of Jordan's lemma, formally stated in Lemma 2.3.9, where $g(z) = \frac{1}{z^2+1}$.

Lemma 2.3.9. (Jordan's Lemma) Let C_R be a contour of an infinite semicircle in the upper-half of the complex plane (this is the semicircle piece shown in Figure 2.15). Let $f(z)$ be a function of the form $f(z) = e^{iaz}g(z), z \in C_R$ and $\lim_{R\to\infty} |g(Re^{i\theta})| = 0$. Then,

$$\left|\int_{C_R} f(z)dz\right| \leq \frac{\pi}{a} M_R,$$

where $M_R = \max_{\theta\in[0,\pi]} |g(Re^{i\theta})|$.

Example 2.3.10. Evaluate the integral

$$I_1 = \int\limits_{-\infty}^{+\infty} \frac{\cos(\omega x)dx}{1+x^2}, \quad \omega > 0.$$

Note: the respective indefinite integral is not expressible via elementary functions and one needs an alternative way of evaluating the definite integral.

Solution. Observe that

$$\int\limits_{-\infty}^{+\infty} \frac{\sin(\omega x)dx}{1+x^2} = 0,$$

just because the integrand is odd (skew-symmetric) over x. Combining the two formulas above one derives

$$I_1 = \int\limits_{-\infty}^{+\infty} \frac{\cos(\omega x)dx}{1+x^2} + \int\limits_{-\infty}^{+\infty} \frac{\sin(\omega x)dx}{1+x^2} = \int\limits_{-\infty}^{+\infty} \frac{\exp(i\omega x)dx}{1+x^2}.$$

Consider an auxiliary integral

$$I_R = \oint \frac{\exp(i\omega z)dz}{1+z^2}, \quad \omega > 0,$$

where the contour consists of half-circle of radius R and the straight line over real axis from $-R$ to R shown in Fig. 2.15. Since the function in the integrand has two poles of the first order, at $z = \pm i$, and only one of these poles lie within the contour, one derives

$$I_R = 2\pi i \mathrm{Res}\left[\frac{\exp(i\omega z)}{1+z^2}, +i\right] = 2\pi i \frac{\exp(i\omega i)}{2i} = \pi \exp(-\omega).$$

On the other hand I_R can be represented as a sum of two integrals, one over $[-R, R]$, and one over the semi-circle. Sending $R \to \infty$ one observes that the later integral vanishes, thus leaving us with the answer

$$I_1 = \pi \exp(-\omega).$$

Example 2.3.11. Evaluate the integral

$$I = \int\limits_{-\infty}^{+\infty} \frac{dx}{\cosh x},$$

reducing it to a contour integral.

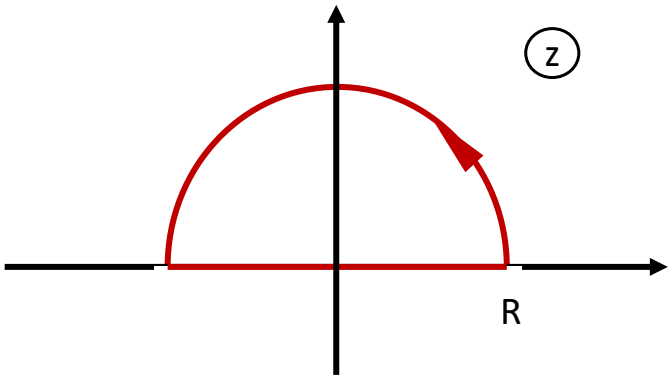


Figure 2.15: See Example 2.3.10.

Solution. Consider contour shown in Fig. 2.14 at $a \to \infty$. Integral along the real axis coincides with the desired integral. Integrals over left (up) and right (down) vertical portions give zero in the $a \to \infty$ limit (because the respective integrands decays to zero exponentially). Note that

$$\cosh(x + i\pi) = \frac{\exp(x + i\pi) + \exp(-x - i\pi)}{2} = -\frac{\exp(x) + \exp(-x)}{2} = -\cosh(x).$$

Therefore the fourth part of the contour becomes, $\int_{i\pi+\infty}^{i\pi-\infty} dx/\cosh(x + i\pi) = I$. Summing up the four pieces and utilizing the result of Example 2.3.7 one derives, $I + 0 + 0 + I = 2\pi$, i.e. $I = \pi$. Obviously the integral can also be evaluated directly (via anti-derivative and definite integral)

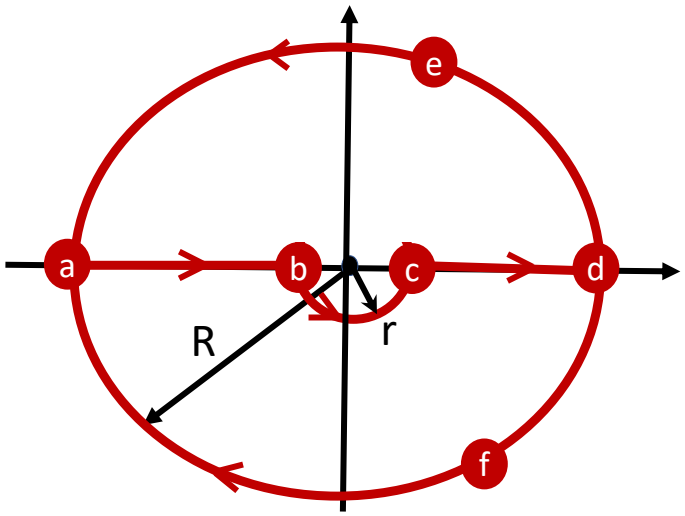
$$I = 2\arctan(\tanh x/2)|_{-\infty}^{+\infty} = \frac{\pi}{2} - \left(-\frac{\pi}{2}\right) = \pi.$$

Exercise 2.9. Evaluate the following integrals reducing them to contour integrals

(a) $\int_0^\infty \frac{dx}{1+x^4}$,

(b) $\int_0^\infty \frac{dx}{1+x^3}$,

(c) $\int_0^\infty \exp\left(ix^2\right) dx$,



0

Figure 2.16

(d) $\int_{-\infty}^{\infty} \frac{\exp(ikx)dx}{\cosh(x)}$,

Cauchy Principal Value

Consider the integral

$$\int_0^{\infty} \frac{\sin(ax)dx}{x}, \tag{2.27}$$

where $a > 0$. As became custom in this part of the course let us evaluate it by constructing and evaluating a contour integral. Since $\sin(az)/z$ is analytic near $z = 0$ (recall or google L'Hôpital rule), we build the contour around the origin as shown in Fig. 2.16. Then going through the following chain of evaluations we arrive at

$$\begin{aligned}
\int_0^{\infty} \frac{\sin(ax)dx}{x} &= \frac{1}{2} \int_{[a\to b\to c\to d]} \frac{\sin(az)}{z} dz \\
&= \frac{1}{4i} \int_{[a\to b\to c\to d]} \left(\frac{\exp(iaz)}{z} - \frac{\exp(-iaz)}{z} \right) dz \\
&= \frac{1}{4i} \int_{[a\to b\to c\to d\to e\to a]} dz \frac{\exp(iaz)}{z} \\
&- \frac{1}{4i} \int_{[a\to b\to c\to d\to f\to a]} dz \frac{\exp(-iaz)}{z} = \frac{1}{4i} (2\pi i - 0) = \frac{\pi}{2}.
\end{aligned} \tag{2.28}$$

(Note that a lot of details in this chain of transformations are dropped. We advise the reader to reconstruct these details. In particular, we suggest to check that the integrals over two semi-circles in Fig. 2.16 decay to zero with $r \to 0$ and $R \to \infty$. For the latter, you may either estimate asymptotic value of the integral yourself, or use the (Jordan's) Lemma 2.3.9.)

The limiting process just explained is often refereed to as the (Cauchy) *principal value* of the integral

$$\mathrm{PV}\int\limits_{-\infty}^{\infty}\frac{\exp(ix)dx}{x}=\lim_{R\to\infty}\int\limits_{-R}^{R}\frac{\exp(ix)dx}{x}=i\pi. \tag{2.29}$$

In general if the integrand, $f(x)$, becomes infinite at a point $x = c$ inside the range of integration, so that the limit on the right of the following expression

$$\int\limits_{-R}^{R}f(x)dx=\lim_{\varepsilon\to 0}\left(\int\limits_{-R}^{c-\varepsilon}dxf(x)+\int\limits_{c+\varepsilon}^{R}dxf(x)\right), \tag{2.30}$$

exists, we call it the principal value integral. (Notice that any of the terms inside the brackets on the right if considered separately may result in a divergent integral.)

Consider another example

$$\int\limits_{a}^{b}\frac{dx}{x}=\log\frac{b}{a}, \tag{2.31}$$

where we write the integral as a formal indefinite integral. However, if $a < 0$ and $b > 0$ the integral diverges at $x = 0$. And we can still define

$$\mathrm{PV}\int\limits_{a}^{b}\frac{dx}{x}:=\lim_{\varepsilon\to 0}\left(\int\limits_{a}^{-\varepsilon}\frac{dx}{x}+\int\limits_{\varepsilon}^{b}\frac{dx}{x}\right)=\lim_{\varepsilon\to 0}\left(\log\frac{\varepsilon}{-a}+\log\frac{b}{\varepsilon}\right)=\log\frac{b}{|a|}, \tag{2.32}$$

excluding ε vicinity of 0. This example helps us to emphasize that the principal value is unambiguous – the condition that the ε-dependent integration limits in $\int^{-\varepsilon}$ and $\int_{\varepsilon}$ are taken with the same absolute value, and say not $\int^{-\varepsilon/2}$ and $\int_{\varepsilon}$, is essential.

If the complex variables were used, we could complete the path by a semicircle from $-\varepsilon$ to ε about the origin (zero), either above or below the real axis. If the upper semicircle were chosen, there would be a contribution, $-i\pi$, whereas if the lower semicircle were chosen, the contribution to the integral would be, $-i\pi$. Thus, according to the path permitted in the complex plane we should have $\int_a^b dz/z = \log(b/|a|) \pm i\pi$. The principal value is the mean of these two alternatives.

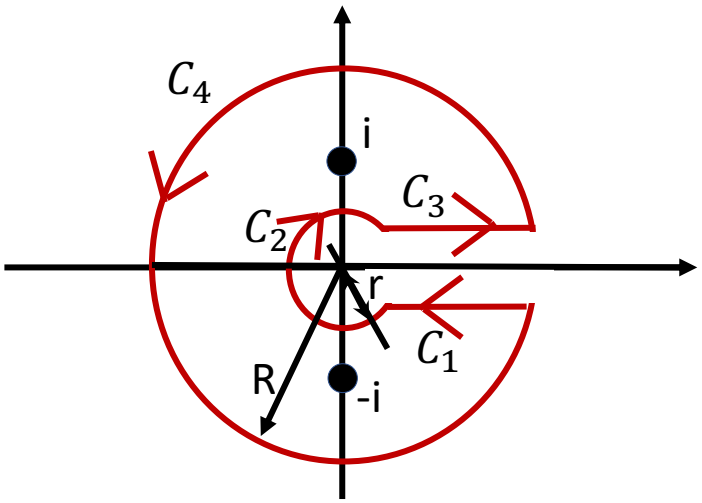


Figure 2.17: See Example 2.3.12.

2.3.3 Contour Integration with Multi-valued Functions

Contour integrals can be used to evaluate certain definite integrals.

Integrals involving Branch Cuts

We discuss below a number of examples of definite integrals which are reduced to contour integrals avoiding branch cuts.

Example 2.3.12. Evaluate the integral

$$\int_0^\infty \frac{dx}{\sqrt{x}(x^2+1)},$$

reducing it to a contour integral.

Solution. The square root in the integrand, $\sqrt{z} = \exp((\log z)/2$, is a multi-valued function, therefore it must be treated with a contour containing a branch cut. Consider

$$\oint \frac{dz}{\sqrt{z}(z^2+1)}$$

where the contour is shown in Fig. 2.17. The contour is chosen to guarantee that

$$r \to 0: \quad \int_{C_2} \frac{dx}{\sqrt{x}(x^2+1)} \to 0,$$

$$R \to \infty: \quad \int_{C_4} \frac{dx}{\sqrt{x}(x^2+1)} \to 0,$$

where the integral is broken in four parts, $\int\limits_{C_1}+\int\limits_{C_2}+\int\limits_{C_3}+\int\limits_{C_4}$. Then resulting (assuming that $r\to 0$ and $R\to\infty$) in

$$\oint \frac{dz}{\sqrt{z}(z^2+1)} = \int\limits_{C_1} \frac{dz}{\sqrt{z}(z^2+1)} + \int\limits_{C_3} \frac{dz}{\sqrt{z}(z^2+1)} = 2\int\limits_0^\infty \frac{dx}{\sqrt{x}(x^2+1)}.$$

On the other hand the full close contour contains two poles of the integrand, at $z=\pm i$, in the interior, therefore

$$\begin{aligned}
\oint \frac{dz}{\sqrt{z}(z^2+1)} &= \pi i \left(\operatorname{Res}\left(\text{at } z=i\right) + \operatorname{Res}\left(\text{at } z=i\right)\right),\\
\operatorname{Res}\left(\text{at } z=i\right) &= \lim_{z\to i}\left(f(z)(z-i)\right) = \lim_{z\to i}\frac{1}{\sqrt{z}(z+i)} = \frac{\exp(3\pi i/4)}{2},\\
\operatorname{Res}\left(\text{at } z=-i\right) &= \lim_{z\to i}\left(f(z)(z+i)\right) = \lim_{z\to i}\frac{1}{\sqrt{z}(z-i)} = \frac{\exp(-3\pi i/4)}{2}.
\end{aligned}$$

Summarizing one arrives at the following answer

$$\int\limits_0^\infty \frac{dx}{\sqrt{x}(x^2+1)} = \pi i\left(\frac{\exp(3\pi i/4)}{2} - \frac{\exp(-3\pi i/4)}{2}\right) = \frac{\pi}{\sqrt{2}}.$$

Exercise 2.10. Evaluate the following integral

$$\int\limits_1^\infty \frac{dx}{x\sqrt{x-1}}.$$

Example 2.3.13. Compute the following integral reducing it to a contour integral

$$I = \int\limits_0^1 \frac{dx}{x^{2/3}(1-x)^{1/3}}. \tag{2.33}$$

Solution. Let us analyze contour integral with almost the same integrand

$$\oint \frac{dz}{z^{2/3}(z-1)^{1/3}} = \oint \frac{dz}{f(z)}, \tag{2.34}$$

and the contour, shown in Fig. (2.18a), surrounding the cut connecting two branching points of $f(z)$, at $z=0$ and $z=1$ (both points are the branching points of the 3rd order).

Recall that the cuts are introduced to make functions which are multi-valued in the complex plain (thus the functions which are not entire, i.e. not analytic within the entire complex plain) to become analytic within the complex plain excluding the cut. Cut also sets the choice for the (originally multi-valued) function branches. In the case under consideration $f(z) := z^{2/3}(z-1)^{1/3}$ has the following parameterization as we go around the cut (in the negative direction):

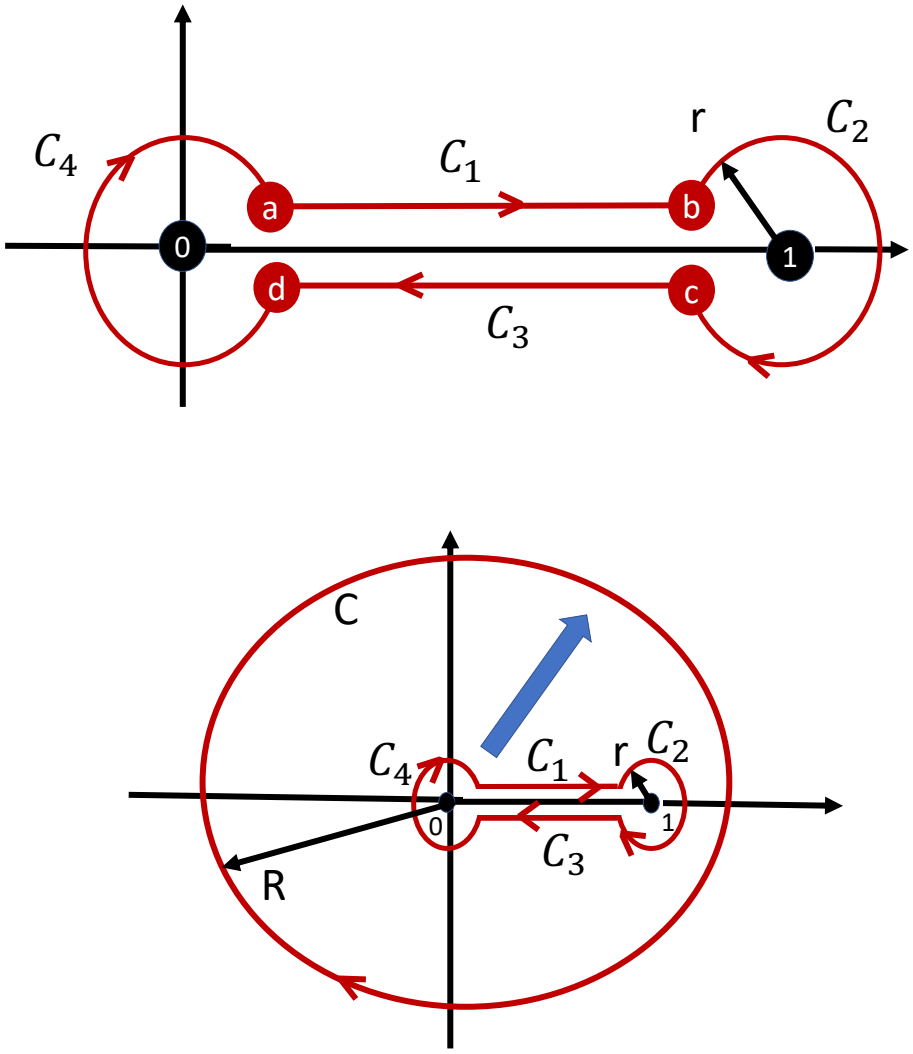


Figure 2.18: See Example 2.3.13.

Sub-contour	Parametrization of z	Evaluation of $f(z)$
$C_1 := [a \to b]$	$x_1,\ x_1 \in [r, 1-r]$	$x_1^{2/3}\lvert 1-x_1\rvert^{1/3}\exp(i\pi/3)$
$C_2 := [b \to c]$	$1 + r\exp(i\theta_2),\ \theta_2 \in [\pi, -\pi]$	$r^{1/3}\exp(i\theta_2/3)$
$C_3 := [c \to d]$	$x_3,\ x_3 \in [1-r, r]$	$x_3^{2/3}\lvert 1-x_3\rvert^{1/3}\exp(-i\pi/3)$
$C_4 := [d \to a]$	$r\exp(i\theta_4),\ \theta_4 \in [2\pi, 0]$	$r^{2/3}\exp(i2\theta_4/3 + i\pi/3)$

Next we compute integrals with the same integrand over the sub-contours, C_1, C_2, C_3, C_4

$$\int\limits_{C_1} \frac{dz}{f(z)} = \int\limits_0^1 \frac{dx_1}{x_1^{2/3}(1-x_1)^{1/3}\exp(i\pi/3)} = \exp(-i\pi/3)I, \tag{2.35}$$

$$\int\limits_{C_2} \frac{dz}{f(z)} = \int\limits_\pi^{-\pi} \frac{ir\exp(i\theta_2)d\theta_2}{(1+r\exp(i\theta_2))^{2/3}(r\exp(i\theta_2)^{1/3}} \to_{r\to 0} 0 \tag{2.36}$$

$$\int\limits_{C_3} \frac{dz}{f(z)} = \int\limits_1^0 \frac{dx_3}{x_3^{2/3}|1-x_3|^{1/3}\exp(-i\pi/3)} = -\exp(i\pi/3)I \tag{2.37}$$

$$\int\limits_{C_4} \frac{dz}{f(z)} = \int\limits_{2\pi}^0 \frac{ir\exp(i\theta_4)d\theta_4}{(r\exp(i\theta_4))^{2/3}(r\exp(i\theta_4)-1)^{1/3}} \to_{r\to 0} 0 \tag{2.38}$$

Taking advantage of $f(z)$ analyticity everywhere outside the $[0,1]$ cut and using Cauchy's integral theorem let us transform the integral over, $C_1 \cup C_2 \cup C_3 \cup C_4$, to the integral with the same integrand over the contour C shown in Fig. (2.18)

$$\int\limits_{C_1} \frac{dz}{f(z)} + \int\limits_{C_2} \frac{dz}{f(z)} + \int\limits_{C_3} \frac{dz}{f(z)} + \int\limits_{C_4} \frac{dz}{f(z)} = \int\limits_C \frac{dz}{f(z)}. \tag{2.39}$$

On the other hand the contour integral over C can be computed in the $R \to \infty$ limit:

$$\int\limits_C \frac{dz}{f(z)} = \int\limits_{2\pi}^0 \frac{iR\exp(i\theta)d\theta}{R^{2/3}\exp(2i\theta/3)(R\exp(i\theta)-1)^{1/3}} \to_{R\to\infty} = -i\int\limits_0^{2\pi} d\theta = -2\pi i. \tag{2.40}$$

Summarizing Eqs. (2.33, 2.34,2.35,2.36,2.37,2.38,2.39,2.40) one arrives at

$$I = \frac{-2\pi i}{-\exp(i\pi/3) + \exp(-i\pi/3)} = \frac{\pi}{\sin(\pi/3)} = \frac{2\pi}{\sqrt{3}}. \tag{2.41}$$

It may be instructive to compare this derivation with an alternative derivation of the integral discussed in [1].

Exercise 2.11. Evaluate the integral

$$\int\limits_{-1}^1 \frac{dx}{(1+x^2)\sqrt{1-x^2}}, \tag{2.42}$$

by suggesting and evaluating an equivalent contour integral.

2.4 Extreme-, Stationary- and Saddle-Point Methods *

In this *auxiliary* * Section, we study the family of related methods which allow to approximate integrals dominated by contribution of a special point and its vicinity. Depending on the case it is called extreme-point (which is also called Laplace method), stationary-point or saddle-point method (which is also called steepest-descent method). We start discussing the extreme-point version, corresponding to estimating real-valued integrals over a real domain, then we turn to estimation of oscillatory (complex-valued) integrals over a real interval (stationary-point method) and then generalize to complex-valued integrals over complex path (saddle-point method, or steepest-descent method).

Extreme- (or maximal-) point method applies to the integral

$$I_1 = \int_a^b dx \exp\left(f(x)\right), \tag{2.43}$$

where the real-valued, continuous function $f(x)$ achieves its maximum at a point $x_0 \in]a,b[$. Then one approximates the function by the first terms of its Taylor series expansion around the maximum

$$f(x) = f(x_0) + \frac{(x-x_0)^2}{2} f''(x_0) + O\left((x-x_0)^3\right), \tag{2.44}$$

where we assume $f'(x_0)$=0. Since x_0 is the maximum, $f'(x_0) = 0$ and $f''(x_0) \leq 0$, and we consider the case of a general position, $f''(x_0) < 0$. One substitutes Eq. (2.44) in Eq. (2.43) and then drops the $O((x-x_0)^3)$ term and extends the integration over $[a,b]$ to $]-\infty,\infty[$. Evaluating the resulting Gaussian integral one arrives at the following extreme-point estimation

$$I_1 \to \sqrt{\frac{2\pi}{-f''(x_0)}} \exp\left(f(x_0)\right). \tag{2.45}$$

This approximation is justified if $|f''(x_0)| \gg 1$.

Example 2.4.1. Estimate the following integral

$$I = \int_{-\infty}^{+\infty} dx \exp(f(x)), \quad f(x) = \alpha x^2 - x^4/2,$$

at sufficiently large positive α using the extreme-point method.

Solution. Let us find all stationary points of $f(x)$ (extreme point of the integrand). Solving $f'(x_s) = 0$, one gets that either $x_s = 0$ or $x_s = \pm\sqrt{\alpha}$. Values of f at the extreme-points

*Here and below we will mark *auxiliary* Sections with *. These Sections can be dropped at the first reading. Material form the auxiliary Sections will not contribute midterm and final exams.

are $f(0) = 0$ and $f(\pm\sqrt{\alpha}) = \alpha^2/2$, and we thus choose the dominating extreme point, $x_s = \pm\sqrt{\alpha}$, for further evaluations. In fact, and since the two (dominant) extreme-points are fully equivalent, we pick one of them and then multiply estimation for the integral by two:

$$I \approx 2\exp(\alpha^2/2)\int\limits_{-\infty}^{+\infty} dx\exp(f''(\sqrt{\alpha})x^2/2) = 2\exp(\alpha^2/2)\int\limits_{-\infty}^{+\infty} dx\exp(-2\alpha x^2)$$
$$= \exp(\alpha^2/2)\sqrt{\frac{2}{\alpha\pi}},$$

where we also took into account that $f''(\pm\sqrt{\alpha}) = -4\alpha$.

The same idea, known under the name of the stationary-point method, works for highly oscillatory integrals of the form

$$I_2 = \int\limits_a^b dx\exp\left(if(x)\right), \tag{2.46}$$

where real-valued, continuous $f(x)$ has a real stationary point x_0, $f'(x_0) = 0$. Integrand oscillates least at the stationary point, thus guaranteeing that the stationary point and its vicinity make dominant contribution to the integral. The statement just made may be a bit confusing because the integrand, considered as a function over x is oscillatory making, formally, integral over x to be highly sensitive to positions of the ends of interval. To make the statement sensible consider shifting the contour of integration into the complex plain so that it crosses the real axis at x_0 along a special direction where $if''(x_0)(x-x_0)^2$ shows maximum at x_0 then making the resulting integrand to decay fast (locally along the contour) with $|x - x_0|$ increase. One derives

$$I_2 \approx \exp\left(if(x_0)\right)\int dx\exp\left(if''(x_0)/2(x-x_0)^2\right)$$
$$= \sqrt{\frac{2\pi}{|f''(x_0)|}}\exp\left(if(x_0) + i\mathrm{sign}(f''(x_0))\pi/4\right),$$

where dependence on the interval's end-points disappear (in the limit of sufficiently large $|f''(x_0)|$).

Example 2.4.2. Estimate the following integral

$$I_2 = \int\limits_{-\infty}^{+\infty} dx\exp(if(x)), \quad f(x) = \alpha x^2 - x^4/2,$$

at sufficiently large positive α using the stationary-point method.

Solution. We can re-use here results of the Example 2.4.1. The stationary points of the integrand are the same: $x_s = 0$ and $x_s = \pm\sqrt{\alpha}$. Values of f at the stationary points are $f(0) = 0$ and $f(\pm\sqrt{\alpha}) = \alpha^2/2$ resulting in 1 and $\exp(i\alpha^2/2)$ contributions to the integrand. Therefore, in the asymptotic (large α) estimation we should keep all three contributions to the integral. Computing second derivatives at the three stationary points, $f''(0) = 2\alpha$ and $f''(\pm\sqrt{\alpha}) = -4\alpha$, estimating the three contributions to I_2 according to Eqs. (2.48), and finally summing them up we arrive at

$$\begin{aligned} I_2 &\approx 2\exp(i\alpha^2/2)\int dx \exp(-2i\alpha x^2) + \int dx \exp(i\alpha x^2) \\ &= 2\exp(i\alpha^2/2 - i\pi/4)\sqrt{\frac{\pi}{2\alpha}} + \sqrt{\frac{2\pi}{\alpha}}\exp(i\pi/4). \end{aligned}$$

Now in the most general case (of the saddle-point method, also called the steepest-descent method) we consider the contour integral

$$I_3 = \int\limits_C dz \exp\left(f(z)\right), \tag{2.47}$$

assuming that $f(z)$ is analytic along the contour, C, and also within a domain, $\mathcal{D}$, of the complex plain, the contour is embedded in. Let us also assume that there exists a point, z_0, within $\mathcal{D}$ where $f'(z_0) = 0$. This point is called a saddle-point because iso-lines of $f(z)$ in the vicinity of z_0 show a saddle – minimum and maximum along two orthogonal directions. Deforming C such that it passes z_0 along the "maximal" path (where $f(z)$ reaches maximum at z_0) one arrives at the following saddle-point estimation

$$I_3 \to \sqrt{\frac{2\pi}{-f''(z_0)}}\exp\left(f(z_0)\right), \tag{2.48}$$

where the square-root sign stand for its main (standard) branch, i.e. $\forall\theta \in [0, 2\pi]$: $\sqrt{\exp(i\theta)} = \exp(i\theta/2)$. In what concerns applicability of the saddle-point approximation – the approximation is based on truncating the Taylor expansion of $f(z)$ around z_0, which is justified if $f(z)$ changes significantly where the expansion applies, i.e. $|f''(z_0)|R^2 \gg 1$, where R is the radius of convergence of the Taylor series expansion of $f(z)$ around z_0.

Two remarks are in order. First, let us emphasize that $f(z_0)$ and $f''(z_0)$ can both be complex. Second, there may be a number (more than one) of saddle points in the region of the $f(z)$ analyticity. In this case one picks the saddle-point achieving maximal value (of $f(z_0)$). In the case of degeneracy, i.e. when multiple saddle-points achieves the same value, as was the case both in the Example 2.4.1 and Example 2.4.2, one deforms the contour to pass through all the saddle-points then replacing right hand side in Eq. (2.48) by the sum of the saddle-point contributions.

Exercise 2.12. * Estimate the following integrals

$$(a) \quad \int_{-\infty}^{+\infty} dx \cos\left(\alpha x^2 - x^3/3\right),$$

$$(b) \quad \int_{-\infty}^{+\infty} dx \exp\left(-x^4/4\right) \cos(\alpha x).$$

at sufficiently large positive α through the saddle-point approximation.

In summary, the important lesson (take away) of extreme-, stationary-, saddle-point analysis of the complex integrals is in our principal ability to search for regions dominating the integrals enabled by analyticity of the integrand. We achieve it shifting the integration contour – (a) making it go through the point(s) where the absolute value of the integrand max out and (b) forcing the contour to ascent and descent the point(s) along the steepest direction, therefore exploring the fact that any of the max-points are saddle-points. The approach allows to extract asymptotic behavior of the integral in the regime where we have a parameter making the ascent/descent infinitely steep in the limit.

Two Homework Assignments associated with the Chapter 2 are:

- **HW1:** Exercises 2.1-2.6.
- **HW2:** Exercises 2.7-2.12.

Chapter 3

Fourier Analysis

Fourier analysis is the study of how functions may be represented or approximated by their oscillatory components. Decomposing a function into its oscillatory components (or basis functions), which requires computing the correct coefficients for each component, is achieved by computing an integral. Similarly, recomposing the function from its orthogonal basis functions is achieved by computing a sum or an integral. When the oscillatory components take a continuous range of wave-numbers (or frequencies), the decomposition and recomposition are referred to as the Fourier transform and inverse Fourier transform. When the oscillatory components take a discrete range of wave-numbers (or frequencies), the decomposition and recomposition are referred to as a Fourier Series.

Fourier analysis grew from the study of Fourier series, which is credited to Joseph Fourier for showing that the study of heat transfer is greatly simplified by representing a function as a sum of trigonometric basis functions. The original concept of Fourier analysis has been extended over time and now applies to more general and more abstract situations. The field is often called harmonic analysis.

3.1 The Fourier Transform and Inverse Fourier Transform

Certain functions $f(\boldsymbol{x})$ can be expressed by the representation, known as the Fourier integral,

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} d\boldsymbol{k} \, \exp\left(i\boldsymbol{k}^T\boldsymbol{x}\right) \hat{f}(\boldsymbol{k}), \tag{3.1}$$

where $\boldsymbol{k} = (k_1, \cdots, k_d)$ is the "wave-vector", $d\boldsymbol{k} = dk_1 \cdots dk_d$, and $\hat{f}(\boldsymbol{k})$ is the Fourier transform of $f(\boldsymbol{x})$, defined according to

$$\hat{f}(\boldsymbol{k}) := \int_{\mathbb{R}^d} d\boldsymbol{x} \, \exp\left(-i\boldsymbol{k}^T\boldsymbol{x}\right) f(\boldsymbol{x}). \tag{3.2}$$

Eq. (3.1) and Eq.(3.2) are inverses of each other (meaning, for example, that substituting Eq. (3.2) into Eq. (3.1) will recover $f(\boldsymbol{x})$), and it is for this reason that the Fourier integral is also called the Inverse Fourier Transform. Proofs that they are inverses, as well as other important properties of the Fourier Transform, rely on Dirac's δ-function which in d-dimensions can be defined as

$$\delta(\boldsymbol{x}) := \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} d\boldsymbol{k}\, \exp(i\boldsymbol{k}^T\boldsymbol{x}). \tag{3.3}$$

We will discuss Dirac's δ-function in Section 3.3, primarily for $d = 1$.

At first glance, it might appear that the appropriate class of functions for which Eq. (3.1) is defined is one where both $f(\boldsymbol{x})$ and $\hat{f}(\boldsymbol{k})$ are integrable. We will demonstrate how the definition of the δ-function permits Eq. (3.1) to be defined over a wider class of functions in Section 3.4. More careful consideration of the function spaces to which $f(\boldsymbol{x})$ and $\hat{f}(\boldsymbol{k})$ belong will be addressed in the Theory course (Math 584).

In the interest of maintaining compact notation and clear explanations, important properties for the Fourier Transform will be presented for the one dimensional case (Section 3.2), but each property applies to the more general d-dimensional Fourier transform. There are only a few functions for which their Fourier transform can be expressed by a closed-form representation, see Section 3.4.

Remark. There are alternative definitions for the Fourier transform and its inverse; some authors place the multiplicative constant of $(2\pi)^{-d}$ in the definition of $\hat{f}(\boldsymbol{k})$, other authors prefer the 'symmetric' definition where both $f(\boldsymbol{x})$ and $\hat{f}(\boldsymbol{k})$ are multiplied by $(2\pi)^{-d/2}$, and still others place a 2π in the complex exponential. It is important to read widely during graduate school, but be warned that the specific results you find will depend on the exact definitions used by the author.

3.2 Properties of the 1-D Fourier Transform

In the $d = 1$ case, x may play the role of the spatial coordinate or of time. When x is the spatial coordinate, the spectral variable k is often called the wave number, which is the one dimensional version of the wave vector. When x is time, k is often called frequency and given the symbol ω. The spatial and temporal terminologies are interchangeable.

Linearity: Let $h(x) = af(x) + bg(x)$, where $a, b \in \mathbb{C}$, then

$$\begin{aligned}\hat{h}(k) &= \int_{\mathbb{R}} dx\, h(x)e^{-ikx} = \int_{\mathbb{R}} dx\, \left(af(x) + bg(x)\right) e^{-ikx} \\ &= a\int_{\mathbb{R}} dx\, f(x)e^{-ikx} + b\int_{\mathbb{R}} dx\, g(x)e^{-ikx} = a\hat{f}(k) + b\hat{g}(k).\end{aligned} \tag{3.4}$$

Spatial/Temporal Translation: Let $h(x) = f(x - x_0)$, where $x_0 \in \mathbb{R}$, then

$$\hat{h}(k) = \int_{\mathbb{R}} dx\, h(x)e^{-ikx} = \int_{\mathbb{R}} dx\, f(x - x_0)e^{-ikx} = \int_{\mathbb{R}} dx'\, f(x')e^{-ikx'-ikx_0} = e^{-ikx_0}\hat{f}(k). \tag{3.5}$$

Frequency Modulation: For any real number k_0, if $h(x) = \exp(ik_0x)f(x)$, then

$$\hat{h}(k) = \int_{\mathbb{R}} dx\, h(x)e^{-ikx} = \int_{\mathbb{R}} dx\, f(x)e^{ik_0x}e^{-ikx} = \int_{\mathbb{R}} dx\, f(x)e^{-i(k-k_0)x} = \hat{f}(k - k_0). \tag{3.6}$$

Spatial/Temporal Rescaling: For a non-zero real number a, if $h(x) = f(ax)$, then

$$\hat{h}(k) = \int_{\mathbb{R}} dx\, h(x)e^{-ikx} = \int_{\mathbb{R}} dx\, f(ax)e^{-ikx} = |a|^{-1}\int_{\mathbb{R}} dx'\, f(x')e^{-ikx'/a} = |a|^{-1}\hat{f}(k/a). \tag{3.7}$$

The case $a = -1$ leads to the time-reversal property: if $h(t) = f(-t)$, then $\hat{h}(\omega) = \hat{f}(-\omega)$.
Complex Conjugation: If $h(x)$ is a complex conjugate of $f(x)$, that is, if $h(x) = (f(x))^*$, then

$$\hat{h}(k) = \int_{\mathbb{R}} dx\, h(x)e^{-ikx} = \int_{\mathbb{R}} dx\, (f(x))^*e^{-ikx} = \int_{\mathbb{R}} dx\, \left(f(x)e^{ikx}\right)^* = \left(\hat{f}(-k)\right)^*. \tag{3.8}$$

Exercise 3.1. Verify the following consequences of complex conjugation:

(a) If f is real, then $\hat{f}(-k) = (\hat{f}(k))^*$ (this implies that $\hat{f}$ is a Hermitian function.)

(b) If f is purely imaginary, then $\hat{f}(-k) = -(\hat{f}(k))^*$.

(c) If $h(x) = \mathrm{Re}(f(x))$, then $\hat{h}(k) = \frac{1}{2}\left(\hat{f}(k) + (\hat{f}(-k))^*\right)$.

(d) If $h(x) = \mathrm{Im}(f(x))$, then $\hat{h}(k) = \frac{1}{2i}\left(\hat{f}(k) - (\hat{f}(-k))^*\right)$.

Exercise 3.2. Show that the Fourier transform of a radially symmetric function in two variables, i.e. $f(x_1, x_2) = g(r)$, where $r^2 = x_1^2 + x_2^2$, is also radially symmetric, i.e. $\hat{f}(k_1, k_2) = \hat{f}(\rho)$, where $\rho^2 = k_1^2 + k_2^2$. (We remind that in polar coordinates (r, θ) a radially symmetric function does not depend on the angle θ.)

Differentiation: If $h(x) = f'(x)$, then under the assumption that $|f(x)| \to 0$ as $x \to \pm\infty$,

$$\begin{aligned}\hat{h}(k) &= \int_{\mathbb{R}} dx\, h(x)e^{-ikx} = \int_{\mathbb{R}} dx f'(x)e^{-ikx} = \left[f(x)e^{-ikx}\right]_{-\infty}^{\infty} - \int_{\mathbb{R}} dx(-ik)f(x)e^{-ikx} \\ &= (ik)\hat{f}(k).\end{aligned} \tag{3.9}$$

Integration: Substituting $k = 0$ in the definition, we obtain $\hat{f}(0) = \int_{-\infty}^{\infty} f(x)\,dx$. That is, the evaluation of the Fourier transform at the origin, $k = 0$, equals the integral of f over all its domain.

Proofs for the following two properties rely on the use of the δ-function (which will not be addressed until Section 3.3), and require more careful consideration of integrability (which is beyond the scope of this brief introduction). The following two properties are added here so that a complete list of properties appears in a single location.

Unitarity [Parceval/Plancherel Theorem]: For any function f such that $\int |f|\,dx < \infty$ and $\int |f|^2 < \infty$,

$$\begin{aligned}\int_{-\infty}^{\infty} dx\,|f(x)|^2 = \int_{-\infty}^{\infty} dx f(x)\,(f(x))^* = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} \frac{dk_1}{2\pi} e^{ik_1 x}\hat{f}(k_1) \int_{-\infty}^{\infty} \frac{dk_2}{2\pi} e^{-ik_2 x}\left(\hat{f}(k_2)\right)^* \\ = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\,|\hat{f}(k)|^2. \qquad (3.10)\end{aligned}$$

Definition 3.2.1. The *integral convolution* of the function f with the function g, is defined as

$$(f * g)(x) := \int_{\mathbb{R}} dy\; g(x-y)f(y), \qquad (3.11)$$

Convolution: Suppose that h is the integral convolution of f with g, that is, $h(x) = (f * g)(x)$, then

$$\hat{h}(k) = \int_{\mathbb{R}} dx\; h(x)e^{-ikx} = \int_{\mathbb{R}} dx \int_{\mathbb{R}} dy\; f(x-y)g(y)e^{-ikx} = \hat{f}(k)\hat{g}(k). \qquad (3.12)$$

The convolution of a function f with a kernel g is defined in Eq. (3.11). Consider whether there exists a convolution kernel g resulting in the projection of a function to itself. That is, can we find a g such that $(g * f) = f$ for arbitrary functions f? If such a g were to exist, what properties would it have?

Heuristically, we could argue that such a function would have to be both localized and unbounded. Localized because for the convolution $\int dy\, g(x-y)f(y)$ to "pick out" $f(x)$, $g(x-y)$ must be zero for all $x \neq y$. Unbounded because we also need $g(x-y)$ to be sufficiently large at $x = y$ to ensure that the integral in the middle of Eq. (3.12) could be nonzero.

Such a degree of 'un-boundedness' over a localized point is impossible under the traditional theory of functions, but nonetheless, an unbound $g(x)$ was introduced by Paul Dirac in the context of quantum mechanics. It was not until the 1940's that Laurent Schwartz

developed a rigorous theory for such 'functions', which became known as the theory of distributions. We usually denote this 'function' by $\delta(x)$ and call it the (Dirac) δ-function. See [1](ch. 4) for more details.

3.3 Dirac's δ-function.

3.3.1 The δ-function as the limit of a δ-sequence

We begin our study of Dirac's δ-function by considering the sequence of functions given by

$$\delta_\epsilon(x) = \begin{cases} 1/\epsilon, & |x| \leq \epsilon/2 \\ 0, & |x| > \epsilon/2 \end{cases}. \tag{3.13}$$

The point-wise limit of δ_ϵ is clearly zero for all $x \neq 0$, and therefore the integral of the limit of δ_ϵ must also be zero, that is,

$$\lim_{\epsilon\to 0} \delta_\epsilon(x) = 0 \quad \Rightarrow \quad \int_{-\infty}^{\infty} dx \, \lim_{\epsilon\to 0} f_\epsilon(x) = 0. \tag{3.14}$$

However, for any $\epsilon > 0$, the integral of δ_ϵ is clearly unity, and therefore the limit of the integral of δ_ϵ must also be unity

$$\int_{-\infty}^{\infty} dx \, \delta_\epsilon(x) = 1 \quad \Rightarrow \quad \lim_{\epsilon\to 0} \int_{-\infty}^{\infty} dx \, \delta_\epsilon(x) = 1. \tag{3.15}$$

Although, Eq. (3.14) suggests that $\delta_\epsilon(x)$ may not be very interesting as a *function*, the behavior demonstrated by Eq. (3.15) motivates the use of $\delta_\epsilon(x)$ as a *functional*[a]. For any sufficiently nice function $\phi(x)$, define the functionals $\delta_\epsilon[\phi]$ and $\delta[\phi]$ by

$$\delta[\phi] := \lim_{\epsilon\to 0} \delta_\epsilon[\phi] := \lim_{\epsilon\to 0} \int_{-\infty}^{\infty} dx \, \delta_\epsilon(x)\phi(x). \tag{3.16}$$

The behavior of $\delta[\phi]$ can be demonstrated by approximating the corresponding integrals, $\delta_\epsilon[\phi]$ for each $\epsilon > 0$:

$$\delta_\epsilon[\phi] = \int_{-\infty}^{\infty} dx \, \delta_\epsilon(x)\phi(x) = \int_{-\epsilon/2}^{\epsilon/2} dx \, \frac{1}{\epsilon}\phi(x). \tag{3.17}$$

[a] In casual terms, a function takes numbers as inputs, and gives numbers as outputs, whereas a functional takes functions as inputs and gives numbers as outputs

Letting m_ϵ and M_ϵ represent the minimum and maximum values of $\phi(x)$ on the interval $-\epsilon/2 < x < \epsilon/2$ gives the bounds

$$m_\epsilon \leq \delta_\epsilon[\phi] \leq M_\epsilon \tag{3.18}$$

If ϕ is continuous at $x = 0$, the limit $\delta_\epsilon[\phi]$ as $\epsilon \to 0$ is given by

$$\delta[\phi] = \lim_{\epsilon\to 0} \delta_\epsilon[\phi] = \phi(0). \tag{3.19}$$

In summary, $\delta[\phi]$ evaluates its argument at the point $x = 0$.

Now compare $\delta_\epsilon(x)$ to the sequence of functions given by

$$\tilde{\delta}_\epsilon(x) = \frac{1}{\pi}\frac{\epsilon}{x^2+\epsilon^2}. \tag{3.20}$$

The point-wise limit $\tilde{\delta}_\epsilon(x)$ is also zero for every $x \neq 0$, so as before, the integral of the limit must be zero

$$\lim_{\epsilon\to 0} \tilde{\delta}_\epsilon(x) = 0 \quad \Rightarrow \quad \int_{-\infty}^{\infty} dx \, \lim_{\epsilon\to 0} \tilde{\delta}_\epsilon(x) = 0. \tag{3.21}$$

A suitable trigonometric substitution shows that the integral of $\tilde{\delta}_\epsilon(x)$ is also unity for each $\epsilon > 0$, and as before, the limit of the integrals must be unity:

$$\int_{-\infty}^{\infty} dx \, \tilde{\delta}_\epsilon(x) = 1 \quad \Rightarrow \quad \lim_{\epsilon\to 0} \int_{-\infty}^{\infty} dx \, \tilde{\delta}_\epsilon(x) = 1. \tag{3.22}$$

As with $\delta_\epsilon(x)$, we can use $\tilde{\delta}_\epsilon(x)$ to define the functionals $\tilde{\delta}_\epsilon[\phi(x)]$ and $\tilde{\delta}[\phi]$ by

$$\tilde{\delta}[\phi] := \lim_{\epsilon\to 0} \tilde{\delta}_\epsilon[\phi] := \lim_{\epsilon\to 0} \int_{-\infty}^{\infty} \tilde{\delta}_\epsilon(x)\phi(x)dx. \tag{3.23}$$

This time it takes a little more thought to find the appropriate bounds, but with some effort, it can be shown that

$$\tilde{\delta}[\phi] = \lim_{\epsilon\to 0} \tilde{\delta}_\epsilon[\phi] = \phi(0) \tag{3.24}$$

That is, $\tilde{\delta}[\phi]$ also evaluates its argument at the point $x = 0$.

The sequences $\delta_\epsilon(x)$ and $\tilde{\delta}_\epsilon(x)$ both have the same limiting behavior as functionals, and are examples of what is known as a δ-sequence. Their limiting behavior leads us to the definition of a δ-function, which is defined as $\delta[\phi] = \int_{\mathbb{R}} dx\delta(x)\phi(x) = \phi(0)$.

Remark. The δ-function only makes sense in the context of an integral. Although it is a common practice to write expressions like $\delta(x)f(x)$, as above, such expression should always be considered as $\int_{\mathbb{R}} dx\, \delta(x)f(x)$

Alternative Definitions of the δ-function

We have defined the δ-function in Eq. (3.3) as the limit of a particular δ-sequence, namely the 'top-hat' function given in Eq. (3.14). One has to wonder whether there may be other δ-sequences which give the same limit. For example, consider

$$\delta(t) = \lim_{\epsilon \to 0} \frac{2t^2\epsilon}{\pi(t^2+\epsilon^2)^2}. \tag{3.25}$$

To validate the suitability of Eq. (3.25) as an alternative definition of the δ-function one needs to check first that $\delta(t) \to 0$ as $\epsilon \to 0$ for all $t \neq 0$, and second that $\int dt\delta(t) = 1$. (It is easy to evaluate this integral as the complex pole integral and closing the contour, for example, over the upper part of the complex plane. Observing that the integrand has pole of the second order at $t = i\epsilon$, expanding it into Laurent series around $i\epsilon$ and keeping the $c = -1$ coefficient, and then using the Cauchy formula for the contour integral, we confirm that the integral is equal to unity.)

Exercise 3.3. Validate the following asymptotic representations for the δ-function

(a) $\delta(t) = \lim_{\epsilon \to 0} \frac{1}{\sqrt{\pi\epsilon}} \exp\left(-\frac{t^2}{\epsilon}\right)$,

(b) $\delta(t) = \lim_{n \to \infty} \frac{1-\cos(nt)}{\pi n t^2}$.

In many applications we deal with periodic functions. In this case one needs to consider relations hold within the interval. In view of the δ-function extreme locality (just explored), all the relations discussed above extend to this case.

Example 3.3.1. Validate the following asymptotic representation for the δ-function on the interval $(-\pi, \pi)$

$$\delta(\theta) = \lim_{r \to 1^-} \frac{1-r^2}{2\pi(1-2r\cos(\theta)+r^2)},$$

where $r \to 1^-$ means $r = 1-\epsilon,\ \epsilon > 0,\ \epsilon \to 0$.

Solution. For $0 < r < 1$, define $\delta_r(\theta) = \frac{1-r^2}{2\pi(1-2r\cos(\theta)+r^2)}$. To show that $\delta_r(\theta)$ is a δ-sequence, we must show:

i. $\lim_{r \to 1^-} \delta_r(\theta) = 0$ for each $\theta \neq 0$,

ii. $\int_{-\pi}^{\pi} \delta_r(\theta)\, d\theta = 1$ for $r < 1$ (i.e. for $\epsilon > 0$ where $r = 1-\epsilon$).

i. To show that the point-wise limit of $\delta_r(\theta)$ is zero for each $\theta \neq 0$, note that for any $\theta \neq 0$, $\lim_{r\to 1^-} 1 - 2r\cos(\theta) + r^2 > 0$ but $\lim_{r\to 1^-} 1 - r^2 = 0$.
ii. To show that $\delta_r(\theta)$ integrates to unity for each r. There is a clever trick that evaluates this integral by a complex-valued contour integral. Let C be the parameterization of the unit circle $z(\theta) = 1e^{i\theta}$ for $-\pi < \theta < \pi$. Therefore $dz = ie^{i\theta}d\theta = izd\theta$. Now consider

$$\begin{aligned}
\int_{-\pi}^{\pi} \delta_r(\theta)d\theta &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{1-r^2}{1-2r\cos(\theta)+r^2}d\theta \\
&= \frac{1}{2\pi}\int_C \frac{1-r^2}{1-r(e^{i\theta}+e^{-i\theta})+r^2}\,\frac{dz}{iz} \\
&= \frac{1}{2\pi i}\int_C \frac{1-r^2}{1-r(z+1/z)+r^2}\,\frac{dz}{z} \\
&= \frac{1}{2\pi i}\int_C \frac{1-r^2}{-r+(1+r^2)z-rz^2}\,dz \\
&= \frac{1}{2\pi i}\int_C \frac{1-r^2}{(1-rz)(z-r)}\,dz
\end{aligned}$$

The integrand has a simple pole inside the contour at $z_1 = r$ with residue equal to 1. (There is also a simple pole at $z_2 = 1/r$ with residue r, but this is irrelevant because it is outside the contour.) Therefore, $\int_{-\pi}^{\pi} \delta_r(\theta)d\theta = 1$. □

3.3.2 Properties of the δ-function

Example 3.3.2. For $b, c \in \mathbb{R}$, show that $c\delta(x-b)f(x) = cf(b)\delta(x-b)$.

Solution. We need to show that (a) the two functions are zero at $x \neq b$ (which is trivial) and (b) that their integrals are equal: $\int_{-\infty}^{\infty} dx\, c\delta(x-b)f(x) = c\int_{-\infty}^{\infty} dy\, \delta(y)f(y+b)$

Example 3.3.3. For $a \in \mathbb{R}$, show that $\delta(ax)f(x) = \delta(x)f(0)/|a|$.

Solution. We need to show that (a) the two functions are zero at $x \neq 0$ (which is trivial) and (b) that their integrals are equal: $\int_{-\infty}^{\infty} dx\, \delta(ax)f(x) = \int_{-\infty}^{\infty} \frac{dy}{|a|}\, \delta(y)f(y/a) = \int_{-\infty}^{\infty} dx\delta(x)f(0)/|a| = f(0)/|a|$.

Example 3.3.4. Show that the Fourier transform of a δ-function is a constant.

Solution. $\hat{\delta}(k) = \int_{-\infty}^{\infty} dx\, \delta(x)e^{-ikx} = e^{-ik\,0} = 1$.

Example 3.3.5. Show that

(a) $\delta(x) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} \exp(ikx)$

(b) The Fourier transform of a constant is a δ-function.

Solution. (a) We identify the expression on the RHS as the inverse Fourier transform of the function $\hat{f}(k) = 1$: $f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk\, 1e^{ikx}$. Even though the constant function is not integrable in the traditional sense, the theory of distributions (which are functionals of the δ-function type) allows us to give meaning to this integral. We know that the δ-function is defined so that for any suitable function $\phi(x)$, $\int dx\, \delta(x)\phi(x) = \phi(0)$. Even if we cannot integrate $f(x)$ directly, but can show that, $\int dx\, f(x)\phi(x) = \phi(0)$, for every suitable test function, $\phi(x)$, then we can assert that, $f(x) = \delta(x)$.
(b)

$$\begin{aligned} f[\phi(x)] &= \int_{-\infty}^{\infty} dx\, \phi(x) \int_{-\infty}^{\infty} \frac{dk}{2\pi}\, 1e^{-ikx} = \int_{-\infty}^{\infty} \frac{dk}{2\pi} \int_{-\infty}^{\infty} dx\, \phi(x) e^{-ikx} = \int_{-\infty}^{\infty} \frac{dk}{2\pi}\, \hat{\phi}(k) \\ &= \int_{-\infty}^{\infty} \frac{dk}{2\pi}\, \hat{\phi}(k) e^{ik0} = \phi(0). \end{aligned}$$

And since, $f[\phi] = \phi(0)$, for every suitable test function ϕ, we say that $f(x) = \delta(x)$.

3.3.3 Using δ-functions to Prove Properties of Fourier Transforms

We now return to proving (1) that the Fourier Transform and the inverse Fourier Transfrom are indeed inverses of each other, (2) Plancherel's theorem and (3) the convolution property.

Proposition 3.3.6. The Fourier Transform of the convolution of the function f with the function g is the product $\hat{f}(k)\hat{g}(k)$

$$\begin{aligned} \widehat{(f * g)}(k) &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, g(x-y) f(y) e^{-ikx} \\ &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} \frac{dk_1}{2\pi} \int_{-\infty}^{\infty} \frac{dk_2}{2\pi}\, \hat{f}(k_1)\hat{g}(k_2) e^{-ikx + ik_1(x-y) + ik_2 y} \\ &= \int_{-\infty}^{\infty} dk_1 \int_{-\infty}^{\infty} dk_2\, \hat{f}(k_1)\hat{g}(k_2) \int_{-\infty}^{\infty} \frac{dx}{2\pi} e^{-ikx + ik_1 x} \int_{-\infty}^{\infty} \frac{dy}{2\pi}\, e^{-ik_1 y + ik_2 y} \\ &= \int_{-\infty}^{\infty} dk_1 \hat{f}(k_1)\delta(k - k_1) \int_{-\infty}^{\infty} dk_2\, \hat{g}(k_2)\delta(k_1 - k_2) = \hat{f}(k)\hat{g}(k) \end{aligned}$$

where in transition from the first to the second lines we exchange order of integrations assuming that all the integrals involved are well-defined.

Proposition 3.3.7. Unitarity [Parceval/Plancherel Theorem]:

$$\begin{aligned}\int_{-\infty}^{\infty} dx|f(x)|^2 &= \int_{-\infty}^{\infty} dx f(x)\left(f(x)\right)^* = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} \frac{dk_1}{2\pi} e^{ik_1 x}\hat{f}(k_1) \int_{-\infty}^{\infty} \frac{dk_2}{2\pi} e^{-ik_2 x}\left(\hat{f}(k_2)\right)^* \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} dk_1 \int_{-\infty}^{\infty} dk_2\, \hat{f}(k_1)\left(\hat{f}(k_2)\right)^* \frac{1}{2\pi}\int_{-\infty}^{\infty} dx \exp(ix(k_1-k_2)) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} dk_1 \int_{-\infty}^{\infty} dk_2\, \hat{f}(k_1)\ \left(\hat{f}(k_2)\right)^* \delta(x-y) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} dk|\hat{f}(k)|^2.\end{aligned}$$

Remark. Using the δ-function as the convolution kernel yeilds the self-convolution property:

$$f(x) = \int dy\delta(x-y)f(y). \tag{3.26}$$

Consider δ-function of a function, $\delta(f(x))$. It can be transformed to the following sum over zeros of $f(x)$,

$$\delta\left(f(x)\right) = \sum_n \frac{1}{|f'(y_n)|}\delta(x-y_n).$$

To prove the statement, one, first of all, recall that δ-function is equal to zero at all points where its argument is nonzero. Just this observation suggest that the answer is a sum of δ-functions and what is left is to establish weights associated with each term in the sum. Pick a contribution associated with a zero of $f(x)$ and integrating the resulting expression over a small vicinity around the point, make the change of variable

$$\int dx\delta(f(x)) = \int \frac{df}{f'(x)}\delta(f(x)).$$

Because of the $\delta(f(x))$ term in the integrand, which is nonzero only at the zero point of $f(x)$, we can replace $f'(x)$ by f' evaluated at the zero and move it out from the integrand. The remaining integral obviously depends on the sign of the derivative. □

3.3.4 The δ-function in Higher Dimensions

The d-dimensional δ-function, which was instrumental for introducing d-dimensional Fourier transform in Section 3.1, is simply a product of one dimensional δ-functions, $\delta(\boldsymbol{x}) = \delta(x_1)\cdots\delta(x_n)$.

Example 3.3.8. Compute the δ-function in polar spherical coordinates.

Solution. Functional expressions for the δ-function in the Cartesian frame and Polar frame are

$$f(\boldsymbol{r}) = \int \delta(\boldsymbol{r}-\tilde{\boldsymbol{r}})f(\tilde{\boldsymbol{r}})d\tilde{\boldsymbol{r}} = \int \delta(x-\tilde{x})\delta(y-\tilde{y})\delta(z-\tilde{z})f(\tilde{x},\tilde{y},\tilde{z})d\tilde{x}d\tilde{y}d\tilde{z}, \tag{3.27}$$

$$= \int \delta(\theta-\tilde{\theta})\delta(\phi-\tilde{\phi})\delta(r-\tilde{r})f(\tilde{r},\tilde{\theta},\tilde{\phi})d\tilde{r}d\tilde{\theta}d\tilde{\phi}. \tag{3.28}$$

On the other hand the volume element transformation from the Cartesian frame to the Polar frame is

$$d\tilde{\boldsymbol{r}} = d\tilde{x}d\tilde{y}d\tilde{z} = \tilde{r}^2\sin\tilde{\theta}d\tilde{r}d\tilde{\theta}d\tilde{\phi}. \tag{3.29}$$

Combining Eqs. (3.27,3.28,3.29) we derive

$$\delta(\boldsymbol{r}-\tilde{\boldsymbol{r}}) = \frac{\delta(\theta-\tilde{\theta})\delta(\phi-\tilde{\phi})\delta(r-\tilde{r})}{\tilde{r}^2\sin\tilde{\theta}}.$$

3.3.5 Formal Differentiation: The Heaviside Function and the Derivatives of the δ-function

The δ-function is not technically a well-defined function, as it only exists in the context of being integrated against a well-defined function. However, formally, using integration techniques, we can write down a well-defined notion for a "derivative" of the δ-function. In fact, we can "differentiate" discontinuous or classically non-differentiable functions using the same notion. Once again, we stress that this is not true differentiation, but rather something that looks like differentiation in form. This technique is often referred to as *formal differentiation*.

Substituting, $f(x) = 1$ into Eq. (3.26) we derive, $\int_{-\infty}^{\infty} dx\delta(x) = 1$. This motivates introduction of a function associated with an incomplete integration of the $\delta(x)$

$$\theta(y) := \int_{-\infty}^{y} dx\delta(x) = \begin{cases} 0, & y<0 \\ 1, & y>0, \end{cases} \tag{3.30}$$

called Heaviside- or step-function.

Exercise 3.4. Prove the relation

$$\left(\frac{d^2}{dt^2} - \gamma^2\right)\exp\left(-\gamma|t|\right) = -2\gamma\delta(t). \tag{3.31}$$

Hint: Yes, the step function will be useful in the proof.

One gets, differentiating Eq. (3.30), that $\theta'(x) = \delta(x)$. We can also differentiate the δ-function. Indeed, integrating Eq. (3.26) by parts, and assuming that the respective anti-derivative is bounded, we arrive at

$$\int dy\delta'(y-x)f(y) = -f'(x) \tag{3.32}$$

Substituting in Eq. (3.32), $f(x) = xg(x)$ we derive

$$x\delta'(x) = -\delta(x). \tag{3.33}$$

Expanding $f(x)$ in the Taylor series around $x = y$, ignoring terms of the second order (and higher) in $(x - y)$, and utilizing Eq. (3.34) one arrives at

$$f(x)\delta'(x-y) = f(y)\delta'(x-y) - f'(y)\delta(x-y). \tag{3.34}$$

Notice that $\delta'(x)$ is skew-symmetric and $f(x)\delta'(x-y)$ is not equal to $f(y)\delta'(x-y)$.

We have assumed so far that $\delta'(x)$ is convolved with a continuous function. To extend it to the case of piece-wise continuous functions with jumps and jumps in derivative, one need to be more careful using integration by parts at the points of the function discontinuity. An exemplary function of this type is the Heaviside function just discussed. This means that if a function, $f(x)$, shows a jump at $x = y$, its derivative allows the following expression

$$f'(x) = (f(y+0) - f(y-0))\delta(x-y) + g(x), \tag{3.35}$$

where, $f(y+0) - f(y-0)$, represents value of the jump and $g(x)$ is finite at $x = y$. Similar representation (involving $\delta'(x)$) can be build for a function with a jump in its derivative. Then the $\delta(x)$ contribution is associated with the second derivative of $f(x)$,

Exercise 3.5. Express $t\delta''(t)$ via $\delta'(t)$.

3.4 Closed form representation for select Fourier Transforms

There are a few functions for which the Fourier transforms can be written in closed form.

3.4.1 Elementary examples of closed form representations

Example 3.4.1. Show that the Fourier Transform of a δ-function is a constant.

Solution. See corollary 3.3.4 where we showed $\hat{\delta}(k) = 1$.

Example 3.4.2. Show that the Fourier Transform of a constant is a δ-function.

Solution. In corollary 3.3.5 where we showed that the inverse Fourier transform of unity was $\delta(x)$. A similar calculation shows that $\hat{1}(k) = 2\pi\delta(k)$

Example 3.4.3. Show that the Fourier transform of a square pulse function is a sinc function:

$$f(x) = \begin{cases} b, & |x| < a \\ 0, & |x| > a. \end{cases} \quad \Rightarrow \quad \hat{f}(k) = \frac{2b}{k}\sin(ka)$$

Solution.

$$\begin{aligned} \hat{f}(k) &= \int_{\mathbb{R}} dx f(x) e^{-ikx} = b\int_{-a}^{a} dx e^{-ikx} = \frac{b}{(-ik)} e^{-ikx}\Big|_{-a}^{a} = \frac{b}{-ik}\left(e^{-ika} - e^{ika}\right) \\ &= \frac{2b}{k}\sin(ka). \end{aligned} \tag{3.36}$$

Example 3.4.4. Show that the Fourier transform of a sinc function is a square pulse:

$$g(x) = \frac{\sin(ax)}{ax} \quad \Rightarrow \quad \hat{g}(k) = \begin{cases} \pi/a, & |k| < a \\ 0, & |k| > a. \end{cases}$$

There are a number of different solutions to this problem and it is instructive to look at each one.

Solution.

$$\hat{g}(k) = \int_{-\infty}^{\infty} dx \frac{\sin(ax)}{ax} e^{-ikx} = \int_{-\infty}^{\infty} dx \frac{e^{iax} - e^{-iax}}{2iax} e^{-ikx} = \underbrace{\int_{-\infty}^{\infty} dx \frac{e^{-i(k-a)x}}{2iax}}_{I} - \underbrace{\int_{-\infty}^{\infty} dx \frac{e^{-i(k+a)x}}{2iax}}_{II}$$

$$= \begin{cases} \pi/a, & |k| < a \\ 0, & |k| > a. \end{cases},$$

where the integrals I and II are computed by transforming them to contour integrals analogous to Eq. (2.28) with contours (distinct for the two contributions) shown in Fig. 2.16.

For both integrals, there is a simple pole at $z = 0$ with residue $1/(2ia)$. The contribution from the pole is $\pm(1/2)(2\pi i)/(2ia)$ where the $(1/2)$ arises from the fact that the pole lies on the contour of integration and the $+/-$ is determined by the orientation of contour (Contours that are closed in the upper-half plane do not need to be reversed (+) whereas contours closed in the lower-half plane must be reversed (-)).

When $k < a$, the contour for I must be closed in the upper half plane and clockwise traversal coincides with the orientation of I (i.e $I = +\pi/(2a)$). When $k > a$, the contour for I must be closed in the lower half plane and clockwise traversal must be reversed to coincide with the orientation of I, (i.e. $I = -\pi/(2a)$).

Similarly, when $k < -a$, the contour for II must be closed in the upper half plane and clockwise traversal coincides with the orientation of II (i.e $II = +\pi/(2a)$). When $k > -a$, the contour for II must be closed in the lower half plane and clockwise traversal must be reversed to coincide with the orientation of II, (i.e. $II = -\pi/(2a)$).

Solution. This solution uses the technique of 'differentiating under the integral', more formally known as Leibniz's integration rule. The integral is tricky because of the x in the denominator. If we consider the integrand as a function of two variables, x and a, and differentiate with respect to a, the x will disappear. We can then integrate with respect to x without concern. Taking the anti-derivative of this result with respect to a to 'undo' the earlier differentiation will yield the final answer. Define a function $I(a)$ such that

$$I(a) := \hat{g}(k) = \int_{-\infty}^{\infty} dx \, \frac{\sin(ax)}{ax} e^{-ikx}$$

The

$$\begin{aligned}\frac{d}{da}\left(aI(a)\right) &= \frac{d}{da}\left(\int_{-\infty}^{\infty} dx \, \frac{\sin(ax)}{x} e^{-ikx}\right) = \int_{-\infty}^{\infty} dx \, \frac{\partial}{\partial a}\left(\frac{\sin(ax)}{x}\right) e^{-ikx} = \int_{-\infty}^{\infty} dx \cos(ax) e^{-ikx} \\ &= \pi\delta(a-k) + \pi\delta(a+k)\end{aligned}$$

So we know that $(a\,I(a))' = \pi\delta(a-k)+\pi\delta(a+k)$. Taking the anti-derivative with respect to a to find $a\,I(a)$ will give a square pulse function plus a constant of integration. The constant of integration is determined to be zero by observing that $\lim_{a\to 0} aI(a) = 0$. Therefore

$$I(a) = \frac{1}{a}\int_0^a \pi\delta(\tilde{a}-k) + \pi\delta(\tilde{a}+k)\, d\tilde{a} = \begin{cases} \pi/a & |k| < a \\ 0 & |k| > a \end{cases}$$

Example 3.4.5. Find the Fourier transform of a Gaussian function

$$f(x) = a\exp(-bx^2), \quad a, b > 0.$$

Solution.

$$\begin{aligned}\hat{f}(k) &= \int_{\mathbb{R}} dx f(x) e^{-ikx} = a\int_{-\infty}^{\infty} dx e^{-bx^2} e^{-ikx} \\ &= a\exp\left(-\frac{k^2}{4b}\right)\int_{-\infty}^{\infty} dx \exp\left(-b\left(x + \frac{ik}{2b}\right)^2\right) \\ &= \frac{a}{\sqrt{b}}\exp\left(-\frac{k^2}{4b}\right)\int_{-\infty}^{\infty} dx' e^{-x'^2} \\ &= a\exp\left(-\frac{k^2}{4b}\right)\sqrt{\frac{\pi}{b}}.\end{aligned}$$

Example 3.4.6. Let $a > 0$. Show that

(a) $f(x) = \dfrac{1}{x^2 + a^2} \quad \Rightarrow \quad \hat{f}(k) = \dfrac{\pi}{a} e^{-a|k|}$;

(b) $g(x) := e^{-a|x|} \quad \Rightarrow \quad \hat{g}(k) := \dfrac{2a}{k^2 + a^2}$.

Solution.

(a) If $k < 0$, the integral $\hat{f}(k)$ can be computed with a complex-valued contour integral. Consider the semi-circular contour in the upper-half plane. The contour encloses a simple pole at $z = ia$ with residue $e^{ka}/(2ia)$.

$$\hat{f}(k) = \int_C \frac{1}{z + ia} \frac{1}{z - ia} e^{-ikz} dz = 2\pi i \frac{1}{2ia} e^{ka} = \frac{\pi}{a} e^{ak}.$$

Similarly, if $k > 0$, the integral $\hat{f}(k)$ can be computed with the semi-circular contour in the lower-half plane. The contour encloses a simple pole at $z = -ia$ with residue $-e^{-ka}/(2ia)$. For this contour, we must multiply to -1 to reverse the orientation of the contour.

$$\hat{f}(k) = \int_C \frac{1}{z + ia} \frac{1}{z - ia} e^{-ikz} dz = 2\pi i \frac{1}{2ia} e^{-ka} = \frac{\pi}{a} e^{-ka}.$$

Combining these results gives $\hat{f}(k) = (\pi/a) e^{-a|k|}$.

(b) Split the integral for $x < 0$ and for $x > 0$.

$$\begin{aligned} \hat{g}(k) &= \int_{-\infty}^{\infty} dx\, e^{-a|x|} e^{-ikx} = \int_{-\infty}^{0} dx\, e^{ax} e^{-ikx} + \int_{0}^{\infty} dx\, e^{-ax} e^{-ikx} \\ &= -\frac{1}{ik - a} + \frac{1}{ik + a} = \frac{2a}{k^2 + a^2}. \end{aligned}$$

Exercise 3.6. Find the Fourier transform of

(a) $f(x) = \dfrac{1}{x^4 + a^4}$,

(b) $f(x) = \operatorname{sech}(ax)$.

3.4.2 More advanced examples of closed form representations

We can find closed form representations of other functions by combining the examples above with the properties in Section 3.2.

Example 3.4.7. Let $f(x) = e^{-x^2}$ and define $g(x) = (f * f * f * f)(x)$. Find (a) $\hat{g}(k)$ and (b) $g(x)$.

Solution. From Example 3.4.5, we know that $\hat{f}(k) = \sqrt{\pi}e^{-k^2 4}$. Therefore,

$$\begin{aligned}\hat{g}(k) = \widehat{(f * f * f * f)}(k) &= \hat{f}(k) \cdot \hat{f}(k) \cdot \hat{f}(k) \cdot \hat{f}(k) = \left(\hat{f}(k)\right)^4 \\ &= \left(\sqrt{\pi}e^{-k^2/4}\right)^4 = \pi^2 e^{-k^2}.\end{aligned}$$

We recognize $\hat{g}(k)$ as a Gaussian function (see example 3.4.5), and we know that if $\hat{g}(k)$ is a Gaussian in the form $a\sqrt{(\pi/b)}e^{-k^2/(4b)}$, then $g(x)$ is the Gaussian ae^{-bx^2}. A little algebra shows we can re-write $\hat{g}(k)$ in this form by setting $a = \frac{1}{2}\pi^{3/2}$ and $b = \frac{1}{4}$:

$$g(x) = \frac{1}{2}\pi^{3/2}e^{-x^2/4}. \qquad \square$$

Example 3.4.8. Let $g(x) = \max(0, 1 - |x|)$ (sometimes called a 'tent' function). Compute $\hat{g}(k)$.

Solution. Let $f(x) = 1$ if $|x| < 1/2$ and 0 otherwise. Note that $(f * f)(x) = g(x)$. From example 3.4.3, we know that the Fourier transform of a square pulse is a sinc function. Therefore,

$$\hat{g}(k) = \hat{f}(k)\hat{f}(k) = \left(\frac{2\sin(k/2)}{k}\right)^2 = \operatorname{sinc}^2(k/2). \qquad \square$$

Example 3.4.9. Let $f(t)$ be given by

$$f(t) = \begin{cases} \cos(\omega_0 t), & |t| < A; \\ 0, & \text{otherwise;} \end{cases}$$

where $\omega_0, A \in \mathbb{R}$ with $A > 0$. (a) Compute $\hat{f}(k)$, the Fourier transform of f, as a function of ω_0 and A. (b) Identify the relationship between the continuity of f and ω_0 and A, and discuss how this affects the decay of the Fourier coefficients as $|k| \to \infty$.

Solution.

(a) By the convolution theorem,

$$f(t) = \cos(\omega_0 t)\text{rect}_A(t) \qquad \Rightarrow \qquad \hat{f}(k) = \widehat{\cos_{\omega_0}}(k) * \widehat{\text{rect}_A}(k),$$

where

$$\widehat{\cos_{\omega_0}}(k) = \pi\delta(k - \omega_0) + \pi\delta(k + \omega_0) \quad \text{and} \quad \widehat{\text{rect}_A}(k) = \frac{2\sin(Ak)}{k}.$$

Therefore,

$$\hat{f}(k) = \widehat{\cos_{\omega_0}}(k) * \widehat{\text{Rect}_A}(k) = \frac{2\pi\sin(A(k - \omega_0))}{k - \omega_0} + \frac{2\pi\sin(A(k + \omega_0))}{k + \omega_0}.$$

Definition	
$\hat{f}(k) = \int_{-\infty}^{+\infty} f(x)e^{-ikx}\,dx$	$\Leftrightarrow \quad f(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \hat{f}(k)e^{ikx}\,dk$
Fourier Transforms Pairs	
$f(x) = \delta(x)$	$\hat{f}(k) = 1$
$f(x) = (2a)^{-1}\,H(1-\lvert x/a\rvert)$	$\hat{f}(k) = \operatorname{sinc}(ak)$
$f(x) = (2a)^{-1}\,\exp(-\lvert x/a\rvert)$	$\hat{f}(k) = (1+(ak)^2)^{-1}$
$f(x) = (\sqrt{2\pi}a)^{-1}\,\exp(-(x/a)^2/2)$	$\hat{f}(k) = \exp(-(ak)^2/2)$
$f(x) = (\sqrt{2\pi}a)^{-1}\operatorname{sech}(\pi x/2a)$	$\hat{f}(k) = \operatorname{sech}(\pi ak/2)$
$f(x) = (\pi a)^{-1}\,(1+(x/a)^2)^{-1}$	$\hat{f}(k) = \exp(-\lvert ak\rvert)$
$f(x) = (\pi a)^{-1}\operatorname{sinc}(x/a)$	$\hat{f}(k) = H(1-\lvert ak\rvert)$
$f(x) = 1$	$\hat{f}(k) = 2\pi\,\delta(k)$

(b) Some basic algebra shows that we can write $\hat{f}(k)$ as follows:

$$\hat{f}(k) = 2\pi\left(\frac{2k\sin(Ak)\cos(A\omega_0) - 2\omega_0\cos(Ak)\sin(A\omega_0)}{k^2-\omega_0^2}\right).$$

In general, the $\hat{f}(k) \sim 1/k$ as $|k| \to \infty$, unless ω_0 and A satisfy $\sin(Ak)\cos(A\omega_0) = 0$, (i.e. $A\omega_0 = n\pi + \frac{\pi}{2}$), then $\hat{f}(k) \sim 1/k^2$.

Exercise 3.7. Let $a \in \mathbb{C}$ with $\mathrm{Re}(a) > 0$ and define $f_a(x) := \dfrac{2a}{a^2+(2\pi x)^2}$. If also $b \in \mathbb{C}$ with $\mathrm{Re}(b) > 0$, show that $(f_a * f_b)(x) = f_{a+b}(x)$.

Exercise 3.8. Show that

(a) $g(x) = \exp(iax)f(bx) \;\Rightarrow\; \hat{g}(k) := \frac{1}{|b|}\hat{f}\left(\frac{k-a}{b}\right)$

(b) $f(x) = \frac{\sin^2(x)}{x} \;\Rightarrow\; \hat{f}(k) = -\frac{i\pi}{2}\Big(\Pi(k-1) - \Pi(k+1)\Big)$.

Here we have that

$$\Pi(k) = \begin{cases} 1, & |k| \le 1; \\ 0, & |k| > 1. \end{cases}$$

3.4.3 Closed form representations in higher dimensions

Example 3.4.10. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, and use the notation $|\boldsymbol{x}|$ to represent $\sqrt{x_1^2 + x_2^2 + \cdots + x_d^2}$. Find the Fourier transform of $g(\boldsymbol{x}) = \exp(-|\boldsymbol{x}|^2)$.

Solution.

$$\begin{aligned}
\hat{g}(\boldsymbol{k}) &= \int\limits_{\mathbb{R}^d} g(\boldsymbol{x}) e^{-i\boldsymbol{k}^T\boldsymbol{x}} d\boldsymbol{x} = \int\limits_{\mathbb{R}^d} e^{-|\boldsymbol{x}|^2} e^{-i\boldsymbol{k}^T\boldsymbol{x}} d\boldsymbol{x} \\
&= \int\limits_{\mathbb{R}^d} e^{-x_1^2 - x_2^2 - \cdots - x_d^2} e^{-i(k_1 x_1 + k_2 x_2 + \cdots + k_d x_d)} d\boldsymbol{x} \\
&= \int\limits_{\mathbb{R}} \int\limits_{\mathbb{R}} \cdots \int\limits_{\mathbb{R}} \left(e^{-x_1^2} e^{-x_2^2} \ldots e^{-x_d^2} \right) \left(e^{-ik_1 x_1} e^{ik_2 x_2} \ldots e^{ik_d x_d} \right) dx_1 dx_2 \ldots dx_d \\
&= \int\limits_{\mathbb{R}} e^{-x_1^2} e^{-ik_1 x_1} dx_1 \int\limits_{\mathbb{R}} e^{-x_2^2} e^{-ik_2 x_2} dx_2 \ \ldots \int\limits_{\mathbb{R}} e^{-x_d^2} \ldots e^{-ik_d x_d} dx_d \\
&= \sqrt{\pi} \exp(-k_1^2/4) \sqrt{\pi} \exp(-k_2^2/4) \ldots \sqrt{\pi} \exp(-k_d^2/4) = \pi^{d/2} \exp(-|\boldsymbol{k}|^2/4).
\end{aligned}$$

3.5 Fourier Series: Introduction

Fourier Series is a version of the Fourier Integral which is used when the function is periodic or of a finite support (nonzero within a finite interval). As in the case of the Fourier Integral/Transform, we will mainly focus on the one-dimensional case. Generalization of the Fourier Series approach to a multi-dimensional case is straightforward.

Consider a periodic function with the period, L. We can represent it in the form of a series over the following standard set of periodic exponentials (harmonics), $\exp(i2\pi nx/L)$:

$$f(x) = \sum_{n=-\infty}^{\infty} f_n \exp\left(2\pi i n x/L\right). \tag{3.37}$$

This, so-called Fourier series, representation of a periodic function immediately shows that the Fourier Series is a particular case of the Fourier integral:

$$\sum_{n=-\infty}^{\infty} f_n \exp\left(2\pi i n x/L\right) \int\limits_{-\infty}^{\infty} dk\, \delta(k-n) = \int\limits_{-\infty}^{\infty} dk\, \exp\left(2\pi i k x/L\right) \sum_{n=-\infty}^{\infty} f_n \delta(k-n). \tag{3.38}$$

Like in the case of the Fourier transform and inverse Fourier transform, we would like to invert Eq. (3.37) and express f_n via $f(x)$. By analogy with the Fourier transform, consider integrating the left hand side of Eq. (3.37) with the oscillating factor, $\exp(-2\pi ikx/L)/L$, where k is an integer, over $x \in [0, L]$. Applying this integration to the right hand side of Eq. (3.37), we run into the following easy to evaluate integral (for each term in the resulting sum)

$$\int\limits_0^L \frac{dx}{L} e^{ikx2\pi/L} e^{-inx2\pi/L} = \frac{e^{i(k-n)2\pi} - 1}{i(k-n)2\pi} = \delta_{k,n} := \begin{cases} 1, & k = n \\ 0, & k \neq n \end{cases}, \tag{3.39}$$

Definition		
$\hat{f}(k) = \int_{-\infty}^{+\infty} f(x) e^{-ikx}\, dx \quad \Leftrightarrow \quad f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(k) e^{ikx}\, dk$		
Property	Original Function	Fourier Transform
	Transformations	
Linearity	$a\, f(x) + b\, g(x)$	$a\, \hat{f}(k) + b\, \hat{g}(k)$
Space-Shifting	$f(x - x_0)$	$e^{-ikx_0}\, \hat{f}(k)$
k-Space-Shifting	$e^{ik_0 x} f(x)$	$\hat{f}(k - k_0)$
Space Reversal	$f(-x)$	$\hat{f}(-k)$
Space Scaling	$f(ax)$	$\lvert a\rvert^{-1} \hat{f}(k/a)$
	Calculus	
Convolution	$g(x) * h(x)$	$\hat{g}(k)\, \hat{h}(k)$
Multiplication	$g(x)h(x)$	$\frac{1}{2\pi}\big(\hat{g}(k) * \hat{h}(k)\big)$
Differentiation	$\frac{d}{dx} f(x)$	$ik\hat{f}(k)$
Integration	$\int_0^x f(\xi)\, d\xi$	$\big(\frac{1}{ik}\big)\, \hat{f}(k)$
	Real-valued vs Complex-valued functions	
Conjugation	$\big(f(x)\big)^*$	$\big(\hat{f}(-k)\big)^*$
Real and Even	$f(x)$ real and even	$\hat{f}(k)$ real and even
Real and Odd	$f(x)$ real and odd	$\hat{f}(k)$ imaginary and odd
Parseval's Theorem / Unitarity		
$\int_{-\infty}^{+\infty} \lvert f(x)\rvert^2\, dx = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \lvert \hat{f}(k)\rvert^2\, dk$		

where expression in the middle is resolved via the LHópitale rule and $\delta_{k,m}$ is the common notation for the so-called Kronecker delta. We observe that only one term in the sum is nonzero, therefore arriving at the desired formula

$$f_n = \int_0^L \frac{dx}{L} f(x) \exp\left(-2\pi i \frac{nx}{L}\right). \tag{3.40}$$

Notice that one may also consider Fourier Transform/Integral as a limit of the Fourier Series. Indeed in the case when a typical scale of the $f(x)$ change is much less than L, many harmonics are significant and the Fourier series transforms to the Fourier integral

$$\sum_{-\infty}^{\infty} \cdots \to \frac{L}{2\pi} \int_{-\infty}^{\infty} dk \cdots . \tag{3.41}$$

Let us illustrate expansion of a function into Fourier series on example of $f(x) = \exp(\alpha x)$ considered on the interval, $0 < x < 2\pi$. In this case the Fourier coefficients are

$$f_n = \int_0^{2\pi} \frac{dx}{2\pi} \exp(-inx + \alpha x) = \frac{1}{2\pi} \frac{1}{\alpha - in} \left(e^{2\pi\alpha} - 1\right). \tag{3.42}$$

Notice that at $n \to \infty$, $f_n \sim 1/n$. As discussed in more details in the following section, the slow decay of the Fourier coefficients is associated with the fact that $f(x)$, when considered as a periodic function over reals with the period 2π, has discontinuities (jumps) at $0, \pm 2\pi, \pm 4\pi, \cdots$.

Exercise 3.9. Let $f(x) = x$ and $g(x) = |x|$ be defined on the interval $-\pi < x < \pi$. (a) Expand both functions as a Fourier series. (b) Compare the dependence of the n-th Fourier coefficient on n for the two functions.

We conclude this Section reminding the reader that our construction of the Fourier Series assumed that the set of harmonic functions forms a complete set of basis functions. Proving this assumption is outside the scope of this course. This proof (and many other proofs) will be discussed in detail in the companion course, Math 584.

3.6 Properties of the Fourier Series

Properties of Fourier series follow their related properties of Fourier transforms discussed in section 3.2. We list the properties in the Table. The presentation is formal, without proofs and made mainly for reference convenience.

Definition

Let $f(x)$ be periodic with period L.

$$f(x) = \sum_{n=-\infty}^{+\infty} f_n \exp\left(2\pi i \frac{nx}{L}\right) \qquad \Leftrightarrow \qquad f_n = \int_0^L \frac{dx}{L} f(x) \exp\left(-2\pi i \frac{nx}{L}\right)$$

Property	Periodic Function	Fourier Series Coefficients
	Scaling	
Linearity	$af(x) + bg(x)$	$af_n + bg_n$
Space-Shifting	$f(x - x_0)$	$e^{2\pi i n x_0/L} f_n$
k-Space-Shifting	$e^{2\pi i k/L} f(x)$	f_{n-k}
Space Reversal	$f(-x)$	f_{-n}
Space Scaling	$f(ax)$ (with period L/a)	f_n
	Calculus	
Periodic Convolution	$\int_0^L f(\xi) g(x-\xi)\, d\xi$	$L f_n g_n$
Multiplication	$f(x)g(x)$	$\sum_{m=-\infty}^{\infty} f_m g_{n-m}$
Differentiation	$\frac{d}{dx} f(x)$	$\frac{2\pi i n}{L} f_n$
Integration	$\int_0^x d\xi f(\xi)$	$\left(\frac{L}{2\pi i n}\right) f_n$
	Real-valued vs Complex-valued functions	
Conjugation	$f^*(x)$	f^*_{-n}
Real and Even	$f(x)$ real and even	f_n real; $f_n = f_{-n}$
Real and Odd	$f(x)$ real and even	f_n imaginary; $f_n = -f_{-n}$

Parseval's Theorem / Unitarity

$$\frac{1}{L}\int_0^L dx |f(x)|^2 = \sum_{n=-\infty}^{\infty} |f_n|^2$$

3.7 Riemann-Lebesgue Lemma

In general, the Fourier series of a function is an infinite series (that is, it contains an infinite number of terms). This means that it is computationally prohibitive to represent an arbitrary function exactly, but a function may be approximated by truncating it's Fourier series. The Riemann-Lebesgue Lemma helps to justify the truncation. The Lemma states that for any integrable function f, the Fourier coefficients f_n must decay as $n \to \infty$.

Theorem 3.7.1 (Riemann-Lebesgue Lemma)**.** If $f(x) \in L^1$, i.e. if the Lebesgue integral of $|f|$ is finite, then $\lim_{n\to\infty} f_n = 0$.

We will not prove the Riemann-Lebesgue lemma here but notice that a standard proof is based on (a) showing that the lemma works for the case of the characteristic function of a finite open interval in $\mathbb{R}^1$, where $f(x)$ is constant within $]a, b[$ and zero otherwise, (b) extending it to simple functions over $\mathbb{R}^1$, that are functions which are piece-wise constant, and then (c) building a sequence of simple functions (which are dense in L^1) approximating $f(x)$ more and more accurately.

Let us mention the following useful corollary of the Riemann-Lebesgue Lemma: For any periodic function $f(x)$ with continuous derivatives up to order m, integration by parts can be performed respective number of times to show that the n-th Fourier coefficient is bounded at sufficiently large n according to $|f_n| \le \frac{C}{|n|^{m+2}}$, where $C = O(1)$.

In particular, and consistently with the example above, we observe that in the case of a "jump", corresponding to continuous anti-derivative, i.e. $m = -1$, $|f_n|$ is $O(1/n)$ asymptotically at $n \to \infty$. In the case of a "ramp", i.e. $m = 0$ with continuous function but discontinuous derivative, $|f_n|$ becomes $O(1/n^2)$ at $n \to \infty$. For the analytic function, with all derivatives continuous, $|f_n|$ decays faster than polynomially as n increases.

Further details of the Lemma, as well as the general discussion of how the material of this Section is related to material discussed in the theory course (Math 584) and also the algorithm course (Math 589), will be given at an inter-core recitation session.

3.8 Gibbs Phenomenon

One also needs to be careful with the Fourier Series truncation, because of the so-called Gibbs phenomenon, called after J. Willard Gibbs, who has described it in 1889. (Apparently, the phenomenon was discovered earlier in 1848 by Henry Wilbraham.) The phenomenon represents an unusual behavior of a truncated Fourier Series built to represent piece-wise continuous periodic function. The Gibbs phenomenon involves both the fact that Fourier sums overshoot at a jump discontinuity, and that this overshoot does not die out as more terms are added to the sum.

Consider the following classic example of a square wave

$$f(x) = \begin{cases} \pi/4, & \text{if} \quad 2n\pi \le x \le (2n+1)\pi, \quad n = 0, 1, 2, \dots \\ -\pi/4, & \text{if} \quad (2n+1)\pi \le x \le (2n+2)\pi, \quad n = 0, 1, 2, \dots \end{cases} \tag{3.43}$$

$$= \sum_{n=0}^{\infty} \frac{\sin((2n+1)x)}{2n+1}, \tag{3.44}$$

where definition of the function is in the first line and the second line describes expression for the function in terms of the Fourier series. Notice that the 2π-periodic function jumps at $2n\pi$ by $\pi/2$.

Let us truncate the series in Eq. (3.44) and thus consider N-th partial Fourier Series

$$S_N(x) = \sum_{n=0}^{N} \frac{\sin((2n+1)x)}{2n+1}. \tag{3.45}$$

Gibbs phenomenon consists in the following observation: as $N \to \infty$ the error of the approximation around the jump-points is reduced in width and energy (integral), but converges to a fixed height. See movie-style visualization (from wikipedia) of how $S_N(x)$ evolves with N. (It is also reproduced in a julia-snippet available at the class D2L repository.)

Let us now back up this simulation by an analytic estimation and compute the limiting value of the partial Fourier Series at the point of the jump. Notice that

$$\frac{d}{d\epsilon} S_N(\epsilon) = \sum_{n=0}^{N} \cos((2n+1)\epsilon) = \frac{\sin(2(N+1)\epsilon)}{2\sin\epsilon}, \tag{3.46}$$

where we have utilized formula for the sum of the geometric progression. Observe that $\frac{d}{d\epsilon} S_N(\epsilon) \to N+1$ at $\epsilon \to 0$, that is the derivative is large (when N is large) and positive. Therefore, $S_N(\epsilon)$ grows with ϵ to reach its (first close to $\epsilon = 0$) maximum at $\epsilon_* = \pi/(2(N+1))$. Now we estimate the value of $S_N(\epsilon_*)$

$$\begin{aligned} S_N(\epsilon_*) &= \sum_{n=0}^{N} \frac{\sin\left(\frac{(2n+1)\pi}{2(N+1)}\right)}{2n+1} = \sum_{n=0}^{N} \frac{\sin\left(\frac{n\pi}{N}\right)}{2n} + O(1/N)\Bigg|_{N\to 0} \\ &\to \frac{1}{2}\int_0^{\pi} \frac{\sin t}{t} dt \approx \frac{\pi}{4} + 0.14, \end{aligned} \tag{3.47}$$

thus observing that at the point of the closest to zero maximum the partial sum systematically overshoots, $f(0^+) = \pi/4$, by an $O(1)$ amount.

Exercise 3.10. Generalize the two functions from Exercise 3.9 beyond the $[-\pi, \pi)$ interval, so they are 2π-periodic functions on $[-5\pi, 5\pi)$. Compute the respective partial Fourier series $S_N(x)$ for select N, and study numerically (or theoretically!) how the amplitude and the width of the oscillations near the points $x = m\pi, m \in \{-5, -4, \ldots, 4\}$ behave as $N \to \infty$.

We complete our discussion of the Fourier Series by mentioning its arguably most significant application in the field of differential equations. Most differential equations of interest can only be solved numerically. Mathematicians often find approximate solutions to a differential equation by representing the solution as a Fourier series and then truncating the series to a finite sum according to the desired accuracy. This method, called the spectral method, will be discussed in the algorithm core course (Math 589).

3.9 Laplace Transform

Recall that to evaluate some Fourier transforms, $\hat{f}(k)$, of the functions, $f(x)$, which do not decay with, $x \to \pm\infty$, like $\exp(ix)$ or $\exp(-x)$, we needed to consider k with nonzero imaginary part to make the resulting Fourier integrals finite.

Let us discuss the two cases, of the oscillatory, $\exp(ix)$, and the exponential, $\exp(-x)$, asymptotic behaviors, separately. The Fourier Series, just discussed, may be considered as a way to avoid the integrability difficulty in the case of the periodic function. Indeed, recall that the coefficients of the Fourier Series, described by Eqs. (3.40), required evaluation of the integral only over a domain of a finite support, like $[0, L]$.

The Laplace transform, we are about to discuss, suggests an elegant way around for another important class of functions, these which decay with $x \to +\infty$, like $\exp(-x)$. Instead of working with the functions defined over all real x, like in the case of the Fourier Transform, or with the periodic functions (or functions defined over a finite support interval) like in the case of the Fourier Series, we turn to discussion of the so-called Laplace Transform (LT) operating on functions defined over the semi-infinite interval, $x \in \mathbb{R}_{\geq 0} = [0, +\infty)$.

Equivalently, we may also introduce the Laplace Transform as a Fourier transform applied to the functions which are nonzero only at $x \geq 0$:

$$\tilde{f}(k) = \int_0^\infty dx \exp(-kx) f(x). \tag{3.48}$$

We consider complex k and require that the integral on the right hand side of Eq. (3.48) is converging (finite) at sufficiently large $\mathrm{Re}(k)$. In other words, $\tilde{f}(k)$ is analytic at $\mathrm{Re}(k) > C$, where C is a positive constant.

Inverse Laplace Transform (ILT) is defined as a complex integral

$$f(x) = \frac{1}{2\pi i} \int_C dk \exp(kx) \tilde{f}(k). \tag{3.49}$$

over the so-called Bromwich contour, C, shown in Fig. (3.1). C can be deformed arbitrarily within the domain, $\mathrm{Re}(k) > 0$, of the $\tilde{f}(k)$ analyticity. Note that by construction, and consistently with the requirement imposed on $f(x)$, the integral on the right hand side of Eq. (3.49) is equal to zero at $x < 0$. Indeed, given that $\tilde{f}(k)$ is analytic at $\mathrm{Re}(k) > 0$ and it approaches zero at $k \to \infty$, contour C can be collapsed to surround ∞, which is also a non-singular point for the integrand thus resulting in zero for the integral.

Properties of the Laplace Transform, following related properties of the Fourier Transform discussed in Section 3.2 are listed formally in the Table below.

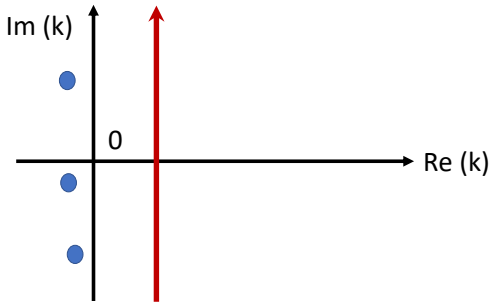


Figure 3.1: The Bromwich integration contour, C, in Eq. (3.48) is shown red. C is often shown as straight line in the complex from $\varepsilon - i\infty$ to $\varepsilon + i\infty$, where ε is an infinitesimally small positive number. Possible singularities of the LT $\tilde{\Phi}(k)$ may only be at the points with negative real number, which are shown schematically as blue dots.

Property	Function	Laplace Transform
	Scaling	
Linearity	$af(x) + bg(x)$	$a\tilde{f} + b\tilde{g}(k)$
Space-Shifting	$f(x - x_0)\theta(x - x_0),\ x_0 \in \mathbb{R}_{>0}$	$e^{-kx_0}\tilde{f}(k)$
k-Space-Shifting	$h(x) = \exp(k_0 x) f(x)$	$\tilde{h}(k) = \tilde{f}(k - k_0)$
Space Scaling	$h(x) = f(ax),\ \forall a \in \mathbb{R}_{>0}$	$\tilde{h}(k) = \frac{1}{a}\tilde{f}\left(\frac{k}{a}\right)$
Differentiation	$h(x) = f'(x)$	$\tilde{h}(k) = k\tilde{f}(k) - f(0^-)$
Integration	$h(x) = \int_0^x f(y)dy$	$\tilde{h}(k) = \frac{1}{k}\tilde{f}(k)$
Convolution	$h(x) = (f * g)(x) = \int_0^x f(y)g(x - y)dy$	$\tilde{h}(k) = \tilde{f}(k)\tilde{g}(k)$

It is instructive to illustrate similarities and differences between the LT and the FT on basic examples.

Consider, first, the one sided exponential

$$f(x) = \theta(x)\exp(-\alpha x), \quad \alpha > 0, \tag{3.50}$$

$$\hat{f}(k) = \frac{\alpha - ik}{\alpha^2 + k^2}, \tag{3.51}$$

$$\tilde{f}(k) = \frac{1}{k + \alpha}, \quad \mathrm{Re}(k + \alpha) > 0. \tag{3.52}$$

Considered in the limit $\alpha \to 0^+$ Eqs. (3.50,3.51,3.52) turn into the following set of

relations for the step function, $\theta(x)$:

$$f(x) = \theta(x), \tag{3.53}$$

$$\hat{f}(k) = \pi\delta(k) - \frac{i}{k}, \tag{3.54}$$

$$\tilde{f}(k) = \frac{1}{k}. \tag{3.55}$$

Shifting and re-scaling the *theta*-function we transform Eqs. (3.53,3.54,3.55) into the following expressions for the Laplace transform of the signature function

$$f(x) = 2\theta(x) - 1 = \text{sign}(x), \tag{3.56}$$

$$\hat{f}(k) = -\frac{2i}{k}, \tag{3.57}$$

$$\tilde{f}(k) = \frac{1}{k}. \tag{3.58}$$

Exercise 3.11. Find the Laplace Transform of (a) $f(x) = \exp(-\lambda x)$ where $\text{Re}(\lambda) > 0$, (b) $f(x) = x^n$ where $n \in \mathbb{Z}$, (c) $f(x) = \cos(\nu x)$ where $\nu \in \mathbb{R}$, (d) $f(x) = \cosh(\lambda x)$ where $\text{Re}(\lambda) > 0$, (e) $f(x) = 1/\sqrt{x}$. Show details.

Exercise 3.12. Find the Inverse Laplace Transform of $1/(k^2 + a^2)$. Show details.

3.9.1 Integral Representations and Asymptotics of Special Functions

The Laplace transform is often used to describe (and to manipulate) integral representations of the special functions. Consider, $f(x) = x^\nu$. Upto an elementary factor its Laplace transform is related to the so-called Gamma, Γ, function (of the parameter ν)

$$\tilde{f}(k) = \int_0^\infty dx x^\nu e^{-kx} = k^{-\nu-1} \int_0^\infty dy y^\nu e^{-y} = \frac{\Gamma(\nu+1)}{k^{\nu+1}},$$

where the integrals are well defined at $k > 0$. Then, the inverse Laplace transform returns "definition" of the Γ function in terms of a contour integral

$$\frac{x^\nu}{\Gamma(\nu+1)} = \frac{1}{2\pi i} \int_{\varepsilon-i\infty}^{\varepsilon+i\infty} dk \frac{e^{kx}}{k^{\nu+1}}, \tag{3.59}$$

where $\varepsilon > 0$ to guarantee that we are on the right of the singularity at $k = 0$ which is a pole if ν is positive integer and a branch point for noninteger ν.

In the general case of the branch point at $k = 0$, the second branch point of the integrand in Eq. (4.24) is at $k = \infty$, and we ought to introduce the branch cut connecting the branch points to guarantee analyticity of the integrand. We choose the branch cut along

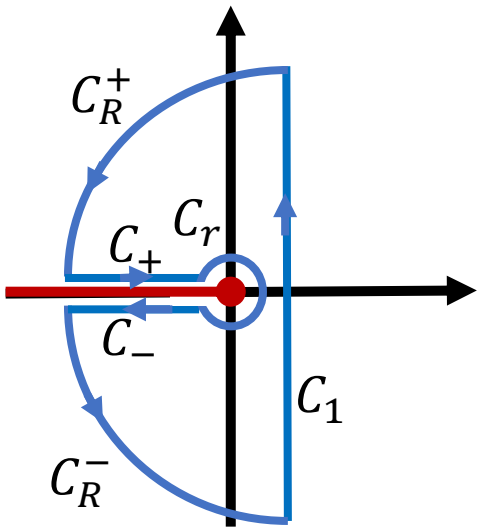


Figure 3.2: Contour transformation for the integral representation of the Γ function via the Inverse Laplace Transform. Integrand of Eq. (3.59) is analytic in the domain surrounded by the (blue) contour, therefore returning zero for the integral.

the negative real axis. Then the integral, $\int\limits_C dk \frac{e^{kx}}{k^{\nu+1}}$, over C, shown (blue) in Fig. (3.2), contains no singularities in the interior and it is therefore equal to zero according to the Cauchy theorem. (Here, $C = C_1 + C_R^+ + C_+ + C_r + C_- + C_R^-$, $R \to \infty$, $r \to 0$ and $C_1 =]\varepsilon - i\infty, \varepsilon + i\infty[$.) We can show (using the Jordan's lemma) that the integral along $C_R^{\pm}$ and C_r tend to zero at $R \to \infty$, $r \to 0$. Therefore, the Bromwich contour in Eq. (3.59) can be replaced by the integral over the $-C_+ - C_-$ contour, i.e. the contour consisting of first going along the negative part of the real axis from $-\infty$ to 0 a tiny bit under the cut and then returning back from 0 to $-\infty$ above the cut. We arrive at

$$\frac{1}{\Gamma(\nu+1)} = \frac{1}{2\pi i} \int\limits_{-C_- - C_+} dk \frac{e^k}{k^{\nu+1}}. \tag{3.60}$$

Eq. (3.60) just derived is quite useful for evaluating asymptotic expression of the function $f(x)$ at $x \to \infty$ with $\tilde{f}(k)$ which requires introduction of the branch cut (on the left from the Bromwich contour). Indeed, consider the Inverse Laplace Transform formula (3.49). The idea is that at $x \to +\infty$ the integral is dominated by the singularity furthest to the right in the complex k plane. (The idea obviously applies not only to the cuts but also to the poles of the integrand in Eq. (3.49).) Then, we can expand around the right-most singularity of

$\tilde{f}(k)$, i.e.

$$\tilde{f}(k) \approx \sum_{\nu} c_\nu (k-k_0)^{\lambda_\nu}, \tag{3.61}$$

where k_0 is the position of the right-most singularity of $\tilde{f}(k)$ and λ_ν may be non-integer. A loop integral around the branch point at k_0 results in an asymptotic series that can be obtained integrating (3.61) term by term

$$f(x) = \frac{1}{2\pi i}\int\limits_{C_{k_0}} dk \tilde{f}(k) e^{kx} \approx \frac{1}{2\pi i}\int\limits_{C_{k_0}} dk \left(\sum_{\nu} c_\nu (k-k_0)^{\lambda_\nu}\right) e^{kx} = e^{k_0 x}\sum_{\nu}\frac{c_\nu}{\Gamma(-\lambda_\nu) x^{\lambda_\nu+1}},$$

where C_{k_0} is the contour surrounding the k_0 singularity anti-clockwise around the cut and we have used Eq. (3.60) for the Γ function.

3.10 From Differential to Algebraic Equations with FT, FS and LT

The Fourier transform, the Fourier Series and the Laplace transform, introduced and discussed in the current Chapter of the notes, will be utilized extensively in the next Chapters of the notes, especially in the following one where we discuss differential equations.

To facilitate the transition consider the simplest possible Differential Equation, relating linearly derivative of the scalar function, $f(x)$ of $x \in \mathbb{R}$ to its derivative

$$\frac{d}{dx} f(x) + q f(x) = g(x), \tag{3.62}$$

where q is a positive constant, $q > 0$, and $g(x)$ is a known scalar function.

Assume that the FT of the function, $g(x)$, on the right hand side of Eq. (3.62) has a well-defined Fourier Transform (FT). Then applying the FT to the differential equation we arrive at a much simpler algebraic equation

$$ik\hat{f}(k) + q\hat{f}(k) = \hat{g}(k), \tag{3.63}$$

which is solved trivially,

$$\hat{f}(k) = \frac{\hat{g}(k)}{q+ik}. \tag{3.64}$$

Applying Inverse Fourier Transform to the result we derive

$$\begin{aligned} f(x) &= \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} dk \frac{\hat{g}(k)}{q+ik} e^{ikx} = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}\frac{dk}{q+ik}\int\limits_{-\infty}^{+\infty} dy g(y) e^{ik(x-y)} \\ &= \int\limits_{-\infty}^{x} dy g(y) e^{-q(x-y)}, \end{aligned} \tag{3.65}$$

which provides an explicit solution of the original differential equation stated in quadratures, i.e. as an integral over the known integrand. Note, that in transition from the first line in Eq. (3.65) to the second line we have exchanged the order of integration and evaluated the pole integral at $k = iq$, also observing that the integral returns non-zero only at $y < x$.

In summary, when $g(x)$ is well defined over the entire $x \in \mathbb{R}$ we solve the differential Eq. (3.62) in three straightforward steps: (a) applying FT and arriving at (much simpler) algebraic equation; (b) solving the algebraic equation; (c) applying the Inverse FT.

Let us now assume that $g(x), f(x) \neq 0$ only at $x \geq 0$. In this case we repeat the logic of the preceding derivation, however substituting the FT by the LT. We derive

$$\begin{aligned} f(x) &= \frac{\theta(x)}{2\pi i}\int\limits_{\varepsilon-i\infty}^{\varepsilon+i\infty} dk \frac{f(0^+)+\tilde{g}(k)}{q+k} e^{kx} \\ &= \frac{\theta(x)}{2\pi i}\int\limits_{\varepsilon-i\infty}^{\varepsilon+i\infty} \frac{dk}{q+k}\left(f(0^+)e^{kx} + \int\limits_0^{+\infty} dy g(y) e^{k(x-y)}\right) \\ &= \theta(x)\left(f(0^+)e^{-qx} + \int\limits_0^{x} dy g(y) e^{-q(x-y)}\right), \end{aligned} \tag{3.66}$$

where in transition from the first to the second line we have accounted for the fact that the LT version of the Eq. (3.63) acquires the additional $f(0^+)$ factor, and then, in transition from the second line to the third line we have exchanged the order of integration and evaluated the pole integrals at $k = -q$. We also observe that Eq. (3.65) transitions into Eq. (3.66) if we set $f(x), g(x)$ to be nonzero only at $x \geq 0$.

Finally, consider Eq. (3.62) in the case supporting 2π-periodicity of $f(x)$, i.e. requiring that $q = im$, where $m \in \mathbb{Z}$, and 2π-periodic $g(x)$. In this case we arrive, after application of the FT to Eq. (3.62) at the discrete version of the algebraic Eq. (3.63), where $\hat{f}(k)$ and $\hat{g}(k)$ are substituted by the Fourier coefficients, f_k and g_k, valid at $k \in \mathbb{Z}$. Then the Fourier Series version of the Eq. (3.65) becomes

$$\begin{aligned} f(x) &= \sum_{k=-\infty}^{+\infty} \frac{g_k}{i(k+m)} e^{ikx} = e^{-imx}\sum_{k'=-\infty}^{+\infty} \frac{g_{k'-m}}{ik'} e^{ik'x} = e^{-imx}\sum_{k'=-\infty}^{+\infty} g_{k'-m}\left(\frac{1}{ik'} + \int\limits_0^x dy e^{ik'y}\right) \\ &= \sum_{k=-\infty}^{+\infty} \frac{g_k}{i(k+m)} + \int\limits_0^x dy e^{im(y-x)} \sum_{k'=-\infty}^{+\infty} e^{ik'(y-x)} g_{k'} = f(0) + \int\limits_0^x dy e^{im(y-x)} g(y). \end{aligned} \tag{3.67}$$

Notice that in the 2π-periodic FS case, like in the semi-infinite LT case, final expressions, given by Eq. (3.67) and Eq. (3.66) respectively, require providing an additional condition

on $f(x)$, chosen to be imposed at $x = 0$. This is reflection of a more general fact that an n-th order ODE requires fixing n condition. The FT expression Eq. (3.65) does show this dependence (on $f(-\infty)$) only because for the Fourier integral to be well defined we effectively required that $f(-\infty) = 0$.

Part II

Differential Equations

Chapter 4

Ordinary Differential Equations.

A *differential equation* (DE) is an equation that relates an unknown function and its derivatives to other known functions or quantities. *Solving* a DE amounts to determining the unknown function. For a DE to be fully determined, it is necessary to define auxiliary information, typically available in the form of an initial or boundary condition.

Often several DE's may be coupled together in a system of DE's. Since this is equivalent to a DE of a vector-valued function, we will use the term "differential equation" to refer to both single equations and systems of equations and the term "function" to refer to both scalar- and vector-valued functions. We will distinguish between the singular and plural only when relevant.

The function to be determined may be a function of a single independent variable, (e.g. $f = f(t)$ or $f = f(x)$) in which case the differential equation is known as an *ordinary* differential equation, or it may be a function of two or more independent variables, (e.g. $f = f(x, y)$, or $f = f(t, x, y, z)$) in which case the differential equation is known as a *partial* differential equation.

The *order* of a differential equation is defined as the largest integer n for which the n^{th} derivative of the unknown function appears in the differential equation.

Most general differential equation is equivalent to the condition that a *nonlinear* function of an unknown function and its derivatives is equal to zero. An ODE is *linear* if the condition is linear in the function and its derivatives. We call the ODE linear, homogeneous if in addition the condition is both linear and homogeneous in the function and its derivatives. It follows for the homogeneous linear ODE that, if $f(t)$ is a solution, so is $cf(t)$, where c is a constant. A linear differential equation that fails the condition of homogeneity is called inhomogeneous. For example, an n^{th} order, inhomogeneous ordinary differential equation is one that can be written as $\alpha_n(t)f^{(n)}(t) + \cdots + \alpha_1(t)f'(t) + \alpha_0(t)g(t) = g(t)$,

where $\alpha_i(t), i = 0, \dots n$ and $g(t)$ are known functions. Typical methods for solving linear differential equations often rely on the fact that the linear combination of two or more solutions to the homogeneous DE is yet another solution, and hence the particular solution can be constructed from from a basis of general solutions. This cannot be done for nonlinear differential equations, and analytic solutions must often be tailor-made for each differential equation, with no single method applicable beyond a fairly narrow class of nonlinear DEs. Due to the difficulty in finding analytic solutions, we often rely on qualitative and/or approximate methods of analyzing nonlinear differential equations, e.g. through dimensional analysis, phase plane analysis, perturbation methods or linearization. In general, linear differential equations admit relatively simple dynamics, as compared to nonlinear differential equations.

An *ordinary differential equation* (ODE) is a differential equation of one or more functions of *one* independent variable, and of the derivatives of these functions. The term ordinary is used in contrast with the term *partial differential equation* (PDE) where the functions are with respect to *more than one* independent variables. PDEs will be discussed in the Chapter 5.

4.1 ODEs: Simple cases

For a warm up let us recall cases of simple ODEs which can be integrated directly.

4.1.1 Separable Differential Equations

A separable differential equation is a first order differential equation that can be written so that the derivative function appears on one side of the equation, and the other side contains the product or quotient of two functions, one of which is a function of the independent variable, and the other a function of the dependent variable.

$$\frac{dx}{dt} = \frac{f(t)}{g(x)} \quad \Rightarrow \quad g(x)dx = f(t)dt \quad \Rightarrow \quad \int g(x)dx = \int f(t)dt. \tag{4.1}$$

Example 4.1.1. Solve the differential equation $\dot{x}(t) = ax(t)t^2$.

Solution.

$$\frac{dx(t)}{dt} = ax(t)t^2 \quad \Rightarrow \quad \frac{dx}{x} = at^2dt \quad \Rightarrow \quad \int^x \frac{d\xi}{\xi} = \int^t a\tau^2 d\tau \quad \Rightarrow \quad \log(x) + c_1 = \frac{a}{3}t^3 + c_2$$
$$\Rightarrow \quad x(t) = c\, e^{at^3/3}.$$

4.1.2 Variation of Parameters

To solve the following linear, inhomogeneous ODE

$$\frac{dy}{dt} - p(t)y(t) = g(t), \quad y(t_0) = y_0, \tag{4.2}$$

let us substitute,

$$y(t) = c(t) \exp\left(\int_{t_0}^{t} d\tau p(\tau) \right), \tag{4.3}$$

where the second term on the right is selected based on solution of the homogeneous version of Eq. (4.2), i.e. $\frac{dy}{dt} = p(t)y(t)$, and one makes the first term, $c(t)$, which would be a constant in the homogeneous case, a function of t. This results in the following equation for the t-dependent $c(t)$

$$\frac{dc(t)}{dt} \exp\left(\int_{t_0}^{t} d\tau p(\tau) \right) = g(t).$$

Applying the method of separable differential equations (see Eq. (4.1)) and then recalling the substitution (4.3), one arrives at

$$y(t) = \exp\left(\int_{t_0}^{t} d\tau p(\tau) \right) \left(y_0 + \int_{t_0}^{t} d\tau g(\tau) \exp\left(- \int_{t_0}^{\tau} d\tau p(\tau) \right) \right).$$

The method just applied to the first order differential equation is called the *method of variation of parameters* because, $c(t)$, in the derivation above can be considered as a parameter which we vary, i.e. allow to depend on t.

Let us extend the idea of the parameter variation method to the case of the second order inhomogeneous differential equation of a general position

$$\frac{d^2y}{dt^2} - p(t)\frac{dy}{dt} - q(t)y(t) = g(t), \tag{4.4}$$

and try to find its general solution. Recall that the general solution of a linear inhomogeneous equation is a sum of the general solutions of the respective homogeneous equation and of a particular solution of the inhomogeneous equation.

Consider, first, solution of the homogeneous equation, i.e. solution of Eq. (4.4) with zero right hand side, $g(t) = 0$. This is a homogeneous differential equation of the second order which should therefore have two independent solutions (we call them linearly independent solutions). Let us denote the two solutions, $y_1(t)$ and $y_2(t)$ and form the so-called Wronskian of the two

$$W(t) = y_1 y_2' - y_2 y_1'. \tag{4.5}$$

Next we compute the derivative of the Wronskian and use the fact that y_1 and y_2 are solutions of the homogeneous version of Eq. (4.4). We derive

$$\frac{d}{dt}W = y_1y_2'' + \cancel{y_1'y_2'} - \cancel{y_2'y_1'} - y_2y_1'' = py_1y_2' + \cancel{qy_1y_2} - \cancel{qy_2y_1} - py_2y_1' = pW.$$

Therefore, the Wronskian becomes

$$W(t) = \exp\left(\int_{t_0}^{t} d\tau p(\tau)\right), \tag{4.6}$$

where t_0 can be chosen arbitrarily. Moreover, given the relation (4.5), we can express one of the two independent solutions via another one and the Wronskian (which we now know explicitly).

Now the question becomes if we can find a particular solution (just one of many) of the inhomogeneous Eq. (4.4)? Let us follow the idea of the method of the variation of parameters and look for a particular solution of the Eq. (4.4) as a linear combination of $x_1(t)$ and $x_2(t)$ multiplied by unknown parameters $A(t)$ and $B(t)$:

$$y(t) = A(t)y_1(t) + B(t)y_2(t). \tag{4.7}$$

Substituting Eq. (4.7) into Eq. (4.4) we derive

$$A''y_1 + B''y_2 + 2A'y_1' + 2B'y_2' - p\left(A'y_1 + B'y_2\right) = g. \tag{4.8}$$

Recall that we are looking for a particular solution. Therefore, we can choose to relate the two unknown coefficients, $A(t)$ and $B(t)$, as we find fit, leaving only one of them as a degree of freedom. The form of Eq. (4.7) suggests to pick the relation so that the dependence on $p(t)$ in Eq. (4.8) disappears, i.e.

$$A'y_1 + B'y_2 = 0. \tag{4.9}$$

Then Eq. (4.8) becomes

$$A'y_1' + B'y_2' = g. \tag{4.10}$$

Notice that the order of derivatives in Eq. (4.10) is reduced in comparison with Eq.(4.8) we started with. Furthermore, expressing B' via A', y_1, y_2 according to Eq. (4.9) and substituting the result in Eq. (4.10) we arrive at

$$WA' + y_2g = 0 \;\Rightarrow\; A = -\int_{t_0}^{t} d\tau \frac{y_2(\tau)g(\tau)}{W(\tau)}. \tag{4.11}$$

Similarly expressing A' via B', y_1, y_2 according to Eq. (4.9) and substituting the result in Eq. (4.10) we derive the B-analog of Eq. (4.11)

$$B = \int_{t_0}^{t} d\tau \frac{y_1(\tau)g(\tau)}{W(\tau)}. \tag{4.12}$$

In summary, to construct solution of Eq. (4.5) we follow the steps

- Find the Wronskian, $W(t)$, given by Eq. (4.6).
- Find a homogeneous solution, $x_1(t)$, and express its linearly independent counterpart, $x_2(t)$, via $x_1(t)$ and the previously found Wronskian, $W(t)$.
- Compute the time dependent factors A and B according to Eq. (4.11,4.12), therefore presenting a particular solution of the original (inhomogeneous) equation according to Eq. (4.7).
- The resulting general solution is a sum of x_1 and x_2, each multiplied by a time independent coefficient, with the particular solution of the inhomogeneous equation (just found) also added to the sum.

The general scheme is illustrated on the following two examples.

Example 4.1.2. Find the general solution to $t^2x''(t) + tx'(t) - x(t) = t$ (where $t \neq 0$) given that $x(t) = t$ is a solution.

Solution. Set the leading coefficient to unity by dividing by t to get $x'' + t^{-1}x' - t^{-2}x = t^{-1}$ (where $t \neq 0$). Therefore $p(t) = -t^{-1}$. We compute the Wronskian

$$W(t) = \exp\left(\int_{t_0}^{t} p(\tau)d\tau\right) = \exp\left(\int_{t_0}^{t} -\tau^{-1}d\tau\right) = t^{-1}$$

The second linearly independent solution is found by

$$W(t) = y_1y_2' - y_2y_1' \quad \Rightarrow \quad \frac{1}{t} = ty_2' - y_21 \quad \Rightarrow \quad y_2(t) = -\frac{1}{2t}$$

Computing A and B

$$A(t) = -\int_{t_0}^{t} d\tau \frac{y_2(\tau)g(\tau)}{W(\tau)} = -\int_{t_0}^{t} d\tau \frac{-\frac{1}{2}t^{-1}t^{-1}}{t^{-1}} = \frac{1}{2}\log(t)$$

.

$$B(t) = \int_{t_0}^{t} d\tau \frac{y_1(\tau)g(\tau)}{W(\tau)} = \int_{t_0}^{t} d\tau \frac{t\ t^{-1}}{t^{-1}} = \frac{t^2}{2}$$

The general solution to the differential equation is

$$x(t) = c_1t + c_2t^{-1} + \frac{1}{2}t\log(t) - \frac{t}{4}$$

Example 4.1.3. Find the general solution to $r''(\theta) + r(\theta) = \tan(\theta)$ for $-\pi/2 < \theta < \pi/2$.

Solution. We compute the Wronskian

$$W(\theta) = \exp\left(\int_{\theta_0}^{\theta} 0 \; d\theta'\right) = 1$$

Let $r_1(\theta) = \cos(\theta)$ be the first linearly independent solution. The second linearly independent solution is found by

$$W(t) = r_1(\theta)r_2'(\theta) - r_2(\theta)r_1'(\theta) \quad \Rightarrow \quad 1 = \cos(\theta)r_2(\theta) + r_2'(\theta)\sin(\theta) \quad \Rightarrow \quad r_2(\theta) = \sin(\theta)$$

Computing A and B

$$A(\theta) = -\int_{\theta_0}^{\theta} d\theta' \frac{r_2(\theta')g(\theta')}{W(\theta')} = -\int_{\theta_0}^{\theta} d\theta' \frac{\sin(\theta)\tan(\theta)}{1} = \sin(\theta) - \log(\sec(\theta) + \tan(\theta))$$

.

$$B(\theta) = \int_{\theta_0}^{t} d\theta' \frac{r_1(\theta')g(\theta')}{W(\theta')} = \int_{\theta_0}^{\theta} d\theta' \frac{\cos(\theta)\ \tan(\theta)}{1} = -\cos(\theta)$$

The solution to the differential equation is

$$x(\theta) = c_1\cos(\theta) + c_2\sin(\theta) + \cos(\theta)\log(\sec(\theta) + \tan(\theta))$$

Exercise 4.1. (a) Find a general solution, $x(t)$ to the following ODE,

$$\frac{dx}{dt} - \lambda(t)x = \frac{f(t)}{x^2},$$

where $\lambda(t)$ and $f(t)$ are known functions of t. (b) Solve the following general second-order, constant-coefficient, linear ODE

$$\tau_0^2 \frac{d^2}{dt^2}y + \tau_1 \frac{d}{dt}y + y = g(t),$$

with the initial conditions $y(0) = y_0, \quad \frac{d}{dt}y\big|_{t=0} = v_0$.

4.2 Direct Methods for Solving Linear ODEs

We continue our exploration of linear by gradually increasing the complexity of the problems and by developing more technical methods.

4.2.1 Homogeneous ODEs with Constant Coefficients

Consider the n-th order homogeneous ODE with constant coefficients

$$\mathcal{L}x(t) = 0, \quad \text{where} \quad \mathcal{L} \equiv \sum_{m=0}^{n} a_{n-m}\frac{d^{n-m}}{dt^{n-m}}. \tag{4.13}$$

(Here and below we will start using bold-calligraphic notation, $\mathcal{L}$, for the differential operators.) Let us look for the general solution of Eq. (4.13) in the form of a linear combination of exponentials

$$x(t) = \sum_{k=1}^{n} c_k \exp(\lambda_k t), \tag{4.14}$$

where c_k are constants. Substituting Eq.(4.14) into Eq.(4.13), one arrives at the condition that the λ_k are roots of the characteristic polynomial:

$$\left(\sum_{m=0}^{n} a_{n-m} \left(\lambda_k\right)^{n-m} \right) = 0. \tag{4.15}$$

Eq. (4.14) holds if the λ_k are not degenerate (that is, if there are n distinct solutions). In the case of degeneracy we generalize Eq. (4.14) to a sum of exponentials (or the non-degenerate λ_k and of polynomials in t multiplied by the respective exponentials for the degenerate λ_k, where for the degrees of the polynomials are equal to the degree of the respective root degeneracy.

$$x(t) = \sum_{k=1}^{m} \left(\sum_{l=0}^{d_k} c_k^{(l)} t^l \right) \exp(\lambda_k t), \tag{4.16}$$

where d_k is the degree of the k-th root degeneracy.

4.2.2 Inhomogeneous ODEs

Consider an inhomogeneous version of a generic linear ODE

$$\mathcal{L}x(t) = f(t). \tag{4.17}$$

Recall that if the particular solution is $x_p(t)$, and if $x_0(t)$ is a generic solution of the homogeneous version of the equation, then a generic solution of Eq. (4.17) can be expressed as $x(t) = x_0(t) + x_p(t)$.

Let us illustrate the utility of this simple but powerful statement on an example:

Example 4.2.1. For (a) $\omega_0 \neq 3$ and for (b) $\omega_0 = 3$, solve

$$\mathcal{L}x := \ddot{x} + \omega_0^2 x = \cos(3t). \tag{4.18}$$

Solution. The general solution to the homogeneous equation, $\mathcal{L}x = 0$ is $x_0(t) = c_1 \cos(\omega_0 t) + c_2 \sin(\omega_0 t)$. For $\omega_0 \neq 3$, a particular solution to Eq. (4.18) is $x_p(t) = \cos(3t)/(\omega_0^2 - 9)$, which can be found by variation of parameters (Section 4.1.2). Therefore, for $\omega_0 \neq 3$, the solution to the inhomogeneous Eq. (4.18) is

$$x(t) = c_1 \cos(\omega_0 t) + c_2 \sin(\omega_0 t) + \frac{\cos(3t)}{\omega_0^2 - 9}.$$

When $\omega_0 = 3$, the natural frequency of the system coincides with the forcing frequency of the right hand side and the system resonates. We must look for a new particular solution because the particular solution we found above is already represented in the solution to the inhomogeneous problem. This particular solution can be found by variation of parameters (Section 4.1.2). Therefore, for $\omega_0 = 3$ the solution to the inhomogeneous Eq. (4.18) is

$$x(t) = c_1 \cos(\omega_0 t) + c_2 \sin(\omega_0 t) + \frac{1}{6}\, t\, \sin(3t).$$

4.3 Linear Dynamics via the Green Function

So far our analysis of ODEs was formal. Often, even though not always, we can associate ODE with dynamics of a system, thus the term "dynamical system". In this case ODE describes evolution of the system variable, x, in time t, i.e. it studies x as a function of t, $x(t)$.

The dynamic considerations are very rich and in this course we will only scratch a surface of interesting phenomena it covers. For example, in Section **??** we discuss the so-called "conservative" dynamics. ODE may also describe a "dissipative" system which relaxes to an "equilibrium". If a dissipative system is in equilibrium, its state does not change in time. If the system is perturbed away from a stable equilibrium, the perturbation is small the system relaxes back to the equilibrium. The relaxation may not be monotonic, and the system may show some oscillations. If the relaxational (dissipative with possible oscillations) dynamics is close to the equilibrium we model it by a linear ODE. There are also many interesting situations when linear ODEs explain oscillations which do not decay.

The method of Green function, or "response" functions, will be the working horse of our analysis for such dynamics, when the ODE is linear. The Green function method offers a powerful and intuitive approach which also extends (in the next chapter) to the case of PDEs.

Let us start with the following general consideration. Given a linear differential equation, $\mathcal{L}x(t) = f(t)$, the goal is to find an operator, $\mathcal{L}^{-1}$, such that "$x(t) = \mathcal{L}^{-1} f(t)$". Since $\mathcal{L}$ is a differential operator, it is reasonable to expect $\mathcal{L}^{-1}$ to be an integral operator, which can be expressed as

$$\mathcal{L}^{-1} f(t) = \int d\tau\, G(t,\tau) f(\tau),$$

where $G(t,\tau)$ is the so-called *Green Function* which is to be determined. Formal manipulations show that

$$f(t) = \mathcal{L}\mathcal{L}^{-1} f(t) = \mathcal{L} \int d\tau\, G(t,\tau) f(t) = \int d\tau\, \mathcal{L} G(t,\tau) f(t).$$

We have already seen this equation as one of the properties of the δ-function. That is

$$f(t) = \int d\tau\, \mathcal{L}G(t,\tau) f(t) \quad \Leftrightarrow \quad \mathcal{L}G(t,\tau) = \delta(t-\tau).$$

The Green function for a differential operator, $\mathcal{L}$, is the function $G(t,\tau)$ that solves the differential equation $\mathcal{L}G(t,\tau) = \delta(t-\tau)$ subject to the prescribed side conditions. The Green function describes the 'response' of the system at time t to a 'impulse' applied at time τ.

Notation. Technically, the Green function is a function of two variables, t and τ, where τ represents the time of an impulse and t represents the time that we observe the system's response to the impulse. Notice that if $\mathcal{L}$ is a differential operator with constant (time-independent) coefficients, then the response of the system to an impulse does not depend on t and τ independently, but instead it only depends on the difference $t-\tau$. In this situation, $G(t,\tau)$ reduces to the "homogeneous in time" or "time-invariant" $G(t-\tau)$.

We will proceed exploring the method by revisiting the simple constant coefficient case of the linear scalar-valued first-order ODE (4.2).

4.3.1 Evolution of a linear scalar

Consider the simplest example of the scalar relaxation

$$\frac{d}{dt}x + \gamma x = f(t), \tag{4.19}$$

where γ is constant and $f(t)$ known function of t. This model appears, for example, when we consider an over-damped driving of a polymer through a medium, where the equation describes the balance of forces where $f(t)$ is the driving force, γx is the elastic (returning) force for a polymer with one end positioned at the origin and another at the position x; and $\dot{x}$ represents friction of the polymer against the medium. The general solution of this equation (recall discussion of the integral operator above, and also notice the time-homogeneous form of the Green function) is

$$x(t) = \int_{-\infty}^{t} d\tau\, G(t-\tau) f(\tau), \tag{4.20}$$

where we have assumed that the evolution starts at $t = -\infty$ with $\lim_{t\to-\infty} x(t) = 0$; and $G(t,\tau)$ is the Green function which satisfies

$$\frac{d}{dt}G(t,\tau) + \gamma G(t,\tau) = \delta(t-\tau), \tag{4.21}$$

and $\delta(t)$ is the δ-function.

Notice that the evolutionary problem we discuss here is an *initial value problem* (also called a Cauchy problem). Indeed, if we would not assume that back in the past (at $t = -\infty$) x is fixed, the solution of Eq. (4.19) would be defined unambiguously. Indeed, suppose $x_s(t)$ is a particular solution of Eq. (4.19), then $x_s(t) = C\exp(-\gamma t)$, where C is a constant, describes a family of solutions of Eq. (4.19). The freedom, eliminated by fixing the initial condition, is associated with the so-called zero mode of the differential operator, $d/dt + \gamma$.

Another remark is about causality, which may also be referred to, in this context, as the "causality principle". It follows from Eq. (4.20) that defining the Green function, one also enforces that, $G(t-\tau) = 0$ at $t < \tau$. This formal observation is, of course, consistent with the obvious—solutions of Eq. (4.19) at a particular moment in time t can only depend on external driving sources $f(\tau)$ that occurred in the past, when $\tau \leq t$. The solution cannot depend on external driving forces that will occur in the future, when $\tau > t$.

Now back to solving Eq. (4.21). Since $\delta(t-\tau) = 0$ at $t > \tau$, one associates $G(t-\tau)$ with the zero mode of the aforementioned differential operator, $G(t-\tau) = A\exp(-\gamma(t-\tau))$, where A is a constant. On the other hand due to the causality principle, $G(t-\tau) = 0$ at $t < \tau$. Integrating Eq. (4.21) over time from $\tau - \epsilon < 0$, where $0 < \epsilon \ll 1$, to τ, we observe that $G(t-\tau)$ should have a discontinuity (jump) at $t = \tau$: $G(t-\tau) = A\exp(-\gamma(t-\tau))\theta(t-\tau)$, where θ is the Heaviside function. Substituting the expression in Eq. (4.21) and integrating the result (left and right hand sides of the resulting equality) over $\tau - \epsilon < t < \tau + \epsilon$, one finds that $A = 1$. Substituting the expression into Eq. (4.20) one arrives at the solution

$$x(t) = \int_{-\infty}^{t} d\tau \exp(-\gamma(t-\tau))f(\tau). \tag{4.22}$$

We observe that the system "forgets" the past at the rate γ per unit time.

Lets sketch out a few different ways to solve Eq. (4.21).

Method 1: Multiply by the appropriate integrating factor. For this problem, the integrating factor is $e^{\gamma t}$.

$$\begin{aligned}
\mathcal{L}G(t,\tau) &= \delta(t-\tau) \\
\frac{d}{dt}\left(e^{\gamma t}G(t,\tau)\right) &= e^{\gamma t}\delta(t-\tau) \\
e^{\gamma t}G(t,\tau) &= \int_{-\infty}^{t} e^{\gamma t'}\delta(t'-\tau)dt' = \int_{-\infty}^{\infty} \theta(t'-\tau)e^{\gamma t'}\delta(t'-\tau)dt' \\
G(t,\tau) &= \theta(t-\tau)e^{-\gamma(t-\tau)}.
\end{aligned}$$

□

Method 2: Take the Fourier transform of both sides, solve the subsequent algebraic equation for $\hat{x}(k)$, and then use a contour integral to compute the inverse Fourier transform of $\hat{x}(k)$.

$$\begin{aligned}
\mathcal{F}\left[\dot{G}(t,\tau)+\gamma G(t,\tau)\right] &= \mathcal{F}\left[\delta(t-\tau)\right] \\
ik\hat{G}(k,\tau)+\gamma\hat{G}(k,\tau) &= e^{-ik\tau} \\
\hat{G}(k,\tau) &= \frac{e^{-ik\tau}}{\gamma+ik} \\
G(t,\tau) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-ik\tau}}{\gamma+ik}e^{ikt}dk \\
G(t,\tau) &= \theta(t-\tau)e^{-\gamma(t-\tau)}
\end{aligned}$$

Where in the last line, we have computed the contour integral by closing the contour with an semi-circular arc of radius R and taken the limit $R\to\infty$. To ensure that the closing arc has a vanishingly small contribution to the integral, the contour is closed off in the upper half plane for $t>\tau$ and in the lower half plane for $t<\tau$. The integrand has a simple pole at $k=i\gamma$ with residue $\frac{1}{2\pi i}e^{-\gamma(t-\tau)}$. Because the pole is in the upper half plane, the integral is equal to $e^{-\gamma(t-\tau)}$ for $t>\tau$ and because there are no poles in the lower half plane, the integral is equal to 0 for $t<\tau$. □

Method 3: Construct the Green function based on a number of properties that it must satisfy. Given $\mathcal{L}G(t,\tau)=\delta(t-\tau)$, we see that $G(t,\tau)$ must satisfy:

(i.) $G(t,\tau)$ solves $\mathcal{L}G(t,\tau)=0$ whenever $t\neq\tau$.

(ii.) $G(t,\tau)$ must satisfy the initial condition.

(iii.) $G(t,\tau)$ must have a jump of size unity at $t=\tau$. This can be explained by integrating from $\tau-\epsilon<t<\tau+\epsilon$ and taking the limit $\epsilon\to 0^+$. The calculation is as follows:

$$\int_{\tau-\epsilon}^{\tau+\epsilon}\dot{G}(t'-\tau)dt'+\gamma\int_{\tau-\epsilon}^{\tau+\epsilon}G(t'-\tau)dt'+\int_{\tau-\epsilon}^{\tau+\epsilon}\delta(t'-\tau)dt' \quad \Rightarrow G(\tau+\epsilon)-G(\tau-\epsilon)+0=1$$

Generalization of property (iii): In general, the Green function of an n^{th} order differential operator has a jump in the $(n-1)^{\text{st}}$ derivative at $t=\tau$. All further derivatives are continuous. The size of the jump is equal to the magnitude of the leading coefficient.

Let's use these properties to construct the Green function.

Step 1. Find candidate solutions by solving $\left(\frac{d}{dt}+\gamma\right)G(t,\tau)=0$. We know there is only one (non-trivial) linearly independent solution because the ODE is linear and first order.

This solution is $Ae^{-\gamma t}$. There is also the trivial solution. To construct the Green function, we must to mix and match between the candidates $G(t,\tau) = 0$ and $G(t,\tau) = Ae^{-\gamma t}$.
Step 2. Apply the initial condition. Our initial condition is $\lim_{t\to-\infty} = 0$. The only candidate solution that satisfies the initial condition is the trivial solution. That is, $G(t,\tau) = 0$ for $t < \tau$. We are not yet prepared to say what happens for $t > \tau$.
Step 3. Apply the jump condition. Given that $G(t,\tau) = 0$ for $t < \tau$, we must now determine $G(t,\tau)$ for $t > \tau$. We realize that $Ae^{-\gamma t}$ is the only candidate that can produce a jump at $t = \tau$. Furthermore, to ensure the jump is size unity, we must set $A = e^{\gamma\tau}$.

In summary, $G(t,\tau) = \theta(t-\tau)e^{-\gamma(t-\tau)}$. □

Exercise 4.2. Solve Eq. (4.19) at $t > 0$, where $x(0) = 0$ and $f(t) = A\exp(-\alpha t)$. Analyze the dependence on α and γ, including $\alpha \to \gamma$.

Recall that Eq. (4.21) assumes that the Green function depends on the difference between t and τ, $t-\tau$, and not on the two variables separately. This assumption is justified for the case considered here, however it will not be correct for situations where the decay coefficient $\gamma(t)$ depends on t. In this general case one needs to consider the general expressions for the Green function discussed above, $G(t,\tau)$. In the case of the constant γ the Green function depends on the difference because of Eq. (4.21) symmetry: invariance with respect to the time translation (time homogeneity), i.e. the equation does not change under the time shift, $t \to t + t_0$.

4.3.2 Evolution of a vector

Let us now generalize and consider

$$\frac{d}{dt}\boldsymbol{x} + \hat{\boldsymbol{\Gamma}}\boldsymbol{x} = \boldsymbol{f}(t), \tag{4.23}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ and $\boldsymbol{f} = (f_1, \ldots, f_n)^\top$ are n-dimensional vector-valued functions of t and $\hat{\boldsymbol{\Gamma}}$ is $n \times n$ time-independent matrix. We consider the two possible cases for $\hat{\boldsymbol{\Gamma}}$: first, when $\hat{\boldsymbol{\Gamma}}$ is either diagonal or diagonalizable, and second, when it is not diagonalizable.

If $\hat{\boldsymbol{\Gamma}}$ is a diagonal matrix, the vector-valued differential equation decouples into n scalar valued differential equations $\dot{x}_i(t) + \gamma_i x_i(t) = f_i(t)$, where $\gamma_1, \ldots \gamma_n$ are the n diagonal entries of $\hat{\boldsymbol{\Gamma}}$. Each of the n scalar-valued differential equations can each be solved independently of each other as discussed in section 4.3.1.

If $\hat{\boldsymbol{\Gamma}}$ is a diagonalizable matrix (but not necessarily diagonal), then we find the eigen-set of $\hat{\boldsymbol{\Gamma}}$

$$\hat{\boldsymbol{\Gamma}}\boldsymbol{a}_i = \lambda_i \boldsymbol{a}_i, \tag{4.24}$$

and expand $\boldsymbol{x}$ and $\boldsymbol{f}$ over the $\{\boldsymbol{a}_i|i\}$ basis,

$$\boldsymbol{x} = \sum_i y_i \boldsymbol{a}_i, \quad \boldsymbol{f} = \sum_i \phi_i \boldsymbol{a}_i. \tag{4.25}$$

Substituting the expansions into Eq. (4.23) one arrives at the n scalar-valued differential equations

$$\frac{dy_i}{dt} + \lambda_i y_i = \phi_i, \tag{4.26}$$

therefore reducing the vector equation to the set of scalar equations of the already considered in section 4.3.1.

If $\hat{\boldsymbol{\Gamma}}$ is not diagonalizable, it can be decomposed into Jordan Canonical form. This occurs when two (or more) eigenvalues share an eigenvector. As before, one introduces the Green function $\hat{\boldsymbol{G}}(t,\tau)$, which satisfies

$$\left(\frac{d}{dt} + \hat{\boldsymbol{\Gamma}}\right) \hat{\boldsymbol{G}}(t,\tau) = \delta(t-\tau)\hat{\mathbf{1}}. \tag{4.27}$$

The explicit solution of Eq. (4.27) is

$$\hat{\boldsymbol{G}}(t,\tau) = \theta(t) \exp\left(-\hat{\boldsymbol{\Gamma}}(t-\tau)\right), \tag{4.28}$$

which allows us to state the solution of Eq. (4.23) in the following invariant form

$$\boldsymbol{x}(t) = \int_{-\infty}^{t} d\tau \hat{\boldsymbol{G}}(t-\tau)\boldsymbol{f}(\tau) = \int_{-\infty}^{t} d\tau \theta(t-\tau) \exp\left(-\hat{\boldsymbol{\Gamma}}(t-\tau)\right) \boldsymbol{f}(\tau). \tag{4.29}$$

Notice that matrix exponential, introduced in Eq. (4.28) and utilized in Eq. (4.29), is the formal expression which may be interpreted in terms of the Taylor series

$$\exp\left(-(t-\tau)\hat{\boldsymbol{\Gamma}}\right) = \sum_{n=0}^{\infty} \frac{(-(t-\tau))^n \hat{\boldsymbol{\Gamma}}^n}{n!}, \tag{4.30}$$

which is always convergent (for the matrix $\hat{\boldsymbol{G}}$ with finite elements).

To relate the invariant expression (4.29) to the eigen-value decomposition of Eqs. (4.25,4.26) one introduces the eigen-decomposition

$$\hat{\boldsymbol{\Gamma}} = \hat{\mathbf{A}}\hat{\mathbf{J}}\hat{\mathbf{A}}^{-1}, \tag{4.31}$$

where $\hat{\mathbf{J}}$ is the matrix of Jordan blocks formed from the eigenvalues of $\hat{\boldsymbol{\Gamma}}$, and the columns of $\hat{\mathbf{A}}$ are the respective eigenvalues of $\hat{\boldsymbol{\Gamma}}$. Note that $\hat{\boldsymbol{\Gamma}}^n = \hat{\mathbf{A}}\hat{\mathbf{J}}^n\hat{\mathbf{A}}^{-1}$.

To illustrate the peculiarity of the degenerate case consider

$$\hat{\boldsymbol{\Gamma}} = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}, \quad \text{which can be re-written as } \lambda\hat{\mathbf{I}} + \hat{\mathbf{N}} \text{ where } \quad \hat{\mathbf{N}} \equiv \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

which is the canonical form of the Jordan (2×2) matrix/block. Observe that $\hat{\mathbf{N}}^2 = \hat{\mathbf{0}}$, which indicates that $\hat{\mathbf{N}}$ is a (2×2) nilpotent matrix. The nilpotent property can be leveraged when taking the matrix exponential,

$$\exp\left(-(t-\tau)\hat{\mathbf{\Gamma}}\right) = e^{-\lambda(t-\tau)}\left(\hat{\mathbf{I}} - (t-\tau)\hat{\mathbf{N}}\right), \tag{4.32}$$

where we have accounted for the nilpotent property of $\hat{\mathbf{N}}$. Incorporating Eq. 4.32 into Eq. 4.29, the solution can therefore be expressed as

$$\boldsymbol{x}(t) = \int_{-\infty}^{t} d\tau\theta(t-\tau)e^{-\lambda(t-\tau)}\left(\hat{\mathbf{I}} - (t-\tau)\hat{\mathbf{N}}\right)\boldsymbol{f}(\tau). \tag{4.33}$$

Alternatively, we could write Eqs. (4.23) in components

$$\frac{dx_1}{dt} + \lambda x_1 + x_2 = f_1, \quad \frac{dx_2}{dt} + \lambda x_2 = f_2,$$

integrating the second equation, substituting result in the first equation, and then changing from x_1 to $\tilde{x}_1 = x_1 + tx_2$, one arrives at

$$\frac{\tilde{x}_1}{dt} + \lambda\tilde{x}_1 = f_1 + tf_2.$$

Note the emergence of a secular term, (a polynomial in t), on the right hand side, which is generic in the case of degeneracy which is then straightforward to integrate.

Exercise 4.3. Find the Green function of Eq. (4.23) for

$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}.$$

Note that vector-valued ODEs appear as the result of the "vectorization" of an n^{th} order scalar-valued ODE for $y(t)$. The vectorization occurs by setting $x_1 = y$, $x_2 = dy/dt$, $\cdots$, $x_n = d^{n-1}y/dt^{n-1}$. Then $d\boldsymbol{x}/dt$ is expressed via the components of $\boldsymbol{x}$ and the original equation, thus resulting in Eq.(4.23).

4.3.3 Higher Order Linear Dynamics

The Green function approach illustrated above can be applied to any inhomogeneous linear differential equation. Let us see how it works in the case of the second-order differential equation for a scalar. Consider

$$\frac{d^2}{dt^2}x + \omega^2 x = f(t). \tag{4.34}$$

To solve Eq. (4.34) note that its general solution can be expressed as a sum of its particular solution and solution of the homogeneous version of Eq. (4.34) with zero right hand side. Let us choose a particular solution of Eq. (4.34) in the form of convolution (4.20) of the source term, $f(t)$, with the Green function of Eq. (4.34)

$$\left(\frac{d^2}{dt^2}+\omega^2\right)G(t)=\delta(t). \tag{4.35}$$

As established above $G(t)=0$ at $t<0$. Integration of Eq. (4.35) from $-\epsilon$ to τ and checking the balance of the integrated terms reveals that $\dot{G}$ jumps at $t=0$, and the value of the jump is equal to unity. An additional integration over time around the singularity shows that $G(t)$ is smooth (and zero) at $t=0$. Therefore, in the case of a second order differential equation considered here: $G=0$ and $\dot{G}=1$ at $t=+0$. Given that $\delta(+0)=0$ these two values can be considered as the initial conditions at $t=+0$ for the homogeneous version (zero right hand side) of the Eq. (4.35), defining $G(t)$ at $t>+0$. Finally, we arrive at the following result

$$G(t)=\theta(t)\frac{\sin(\omega t)}{\omega}, \tag{4.36}$$

where θ is the Heaviside function.

Furthermore, Eq. (4.20) gives the solution to Eq. (4.34) over the infinite time horizon, however one can also use the Green function to solve the respective Cauchy propblem (initial value problem). Since Eq. (4.34) is the second order ODE, one just needs to fix two values associated with $x(t)$ evaluated at the initial, $t=0$, for example $x(0)$ and $\dot{x}(0)$. Then, taking into account that, $G(+0)=0$ and $\dot{G}(+0)=1$, one finds the following general solution of the Cauchy problem for Eq. (4.34)

$$x(t)=\dot{x}(0)G(t)+x(0)\dot{G}(t)+\int\limits_0^t dt_1 G(t-t_1)f(t_1). \tag{4.37}$$

Let us now generalize and consider

$$\mathcal{L}x=f(t),\quad \mathcal{L}\equiv\sum_{m=0}^{n}a_{n-m}\frac{d^{n-m}}{dt^{n-m}}, \tag{4.38}$$

where a_i are constants and $\mathcal{L}$ is the linear differential operator of the n-th order with constant coefficients, already discussed in Section 4.2. We build a particular solution of Eq. (4.38) as the convolution (4.20) of the source term, $\phi(t)$, with the Green function, $G(t)$, of Eq. (4.38)

$$\mathcal{L}G=\delta(t), \tag{4.39}$$

where $G(t)=0$ at $t<0$.

Observe that the solution to the respective homogeneous equation, $\mathcal{L}x = 0$, (the zero modes of the operator $\mathcal{L}$) can be generally presented as

$$x(t) = \sum_i b_i \exp(z_i t), \tag{4.40}$$

where b_i are arbitrary constants.

Let us now use the general representation (4.40) to construct the Green function solving Eq. (4.39). Recall that, considering first and second order differential equations in the preceding Sections, we have transitioned above from the inhomogeneous equations for the Green function to the homogeneous equation supplemented with the initial conditions. Direct extension of the "integration around zero" approach (doing it n times) reveals that initial conditions one needs to set at $t = +0$ in the general case of the n-th order differential equation are

$$a_{n-1}\frac{d^{n-1}}{dt^{n-1}}G(0^+) = 1, \quad \forall 0 \le m < n-1: \quad \frac{d^m}{dt^m}G(0^+) = 0. \tag{4.41}$$

Consider, formally, $\mathcal{L}$, as a polynomial in z, where z is the elementary differential operator, $z = d/dt$, i.e. $\mathcal{L}(z)$. Then, at $t > 0^+$ the Green function satisfies the homogeneous equation, $\mathcal{L}(d/dt)G = 0$. Solution of the homogeneous equation can generally be presented as

$$t > 0^+: \quad G(t) = \sum_i b_i \exp(z_i t), \tag{4.42}$$

where b_i are arbitrary constants which are defined unambiguously from the system of algebraic equations for the coefficients one derives substituting Eq. (4.42) in Eq. (4.41).

Example 4.3.1. Let $\gamma, \nu \in \mathbb{R}$ with $\gamma > 0$. Find the Green function for the differential operator and use it to solve the ODE

$$\frac{d^2}{dt^2}x + 2\gamma\frac{d}{dt}x + \nu^2 x = f \quad \text{subject to:} \quad \lim_{x\to-\infty} x(t) = 0, \quad \lim_{t\to-\infty} \dot{x}(t) = 0. \tag{4.43}$$

Consider the case $\nu < \gamma$ and $\nu > \gamma$

Notation. As before, $G(t,\tau)$ is a function of both t and τ, where the variable τ represents the time that an impulse is applied. For any fixed τ, $G(t,\tau)$ is the response of the impulse a function of t. It's Fourier transform, $\hat{G}(\omega,\tau)$, is the decomposition of $G(t,\tau)$ into its oscillatory modes.

Solution. Longer method: Solve $\mathcal{L}G = \delta(t-\tau)$ by taking Fourier Transforms: To find $G(t,\tau)$, take the Fourier transform of $LG(t,\tau) = \delta(t-\tau)$ and solve for $\hat{G}(\omega,\tau)$.

$$\widehat{\mathcal{L}G}(\omega;\tau) = \hat{\delta}(\omega;\tau) \;\Rightarrow\; \left(-\omega^2 - 2i\gamma\omega + \nu^2\right)G = e^{-i\omega\tau} \;\Rightarrow\; \hat{G}(\omega;\tau) = \frac{-e^{-i\omega\tau}}{(\omega-\omega_+)(\omega-\omega_-)},$$

where $\omega_\pm = -i\gamma \pm \sqrt{\nu^2 - \gamma^2}$. The inverse Fourier transform of $\hat{G}$ is computed by a contour integral where the contour must be closed off by a semi-circular arc of radius R under the limit $R \to \infty$. To ensure that the semi-circular arc has vanishing contribution to the integral, we must close off the contour in the upper-half plane if $t < \tau$ and in the lower-half plane if $t > \tau$. The integrand has poles with associated residues at:

- If $\nu > \gamma$:
  Simple poles at $\omega = \omega_\pm = -i\gamma \pm \sqrt{\nu^2 - \gamma^2}$ (with both real and imaginary components)
  $Res(f, \omega_-) = (\omega_- - \omega_+)^{-1} \exp(-i\omega_- \tau) = -(\nu^2 - \gamma^2)^{-1/2} \exp(-\gamma\tau) \exp\left(-i\sqrt{\nu^2 - \gamma^2}\tau\right)$
  $Res(f, \omega_+) = (\omega_+ - \omega_-)^{-1} \exp(-i\omega_+ \tau) = +(\nu^2 - \gamma^2)^{-1/2} \exp(-\gamma\tau) \exp\left(-i\sqrt{\nu^2 - \gamma^2}\tau\right)$
- If $\nu < \gamma$:
  Simple poles at $\omega = \omega_\pm = -i\gamma \pm i\sqrt{\gamma^2 - \nu^2}$ (purely imaginary)
  $\mathrm{Res}(f, \omega_-) = (\omega_- - \omega_+)^{-1} \exp(-i\omega_- \tau) = -(\gamma^2 - \nu^2)^{-1/2} \exp(-\gamma\tau) \exp\left(-\sqrt{\gamma^2 - \nu^2}\tau\right)$
  $\mathrm{Res}(f, \omega_+) = (\omega_+ - \omega_-)^{-1} \exp(-i\omega_+ \tau) = +(\gamma^2 - \nu^2)^{-1/2} \exp(-\gamma\tau) \exp\left(-\sqrt{\gamma^2 - \nu^2}\tau\right)$

The Green function is given by

$$G(t, \tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega t} dt = \frac{1}{2\pi} \Big(2\pi \,\mathrm{Res}(f, \omega_-) + 2\pi \,\mathrm{Res}(f, \omega_+) \Big)$$

Because there are no singularities in the upper half plane, $G(t, \tau) = 0$ if $t < \tau$. Physically, this means that the system has no response to an impulse that will happen in the future (also called causality). Simplifying the algebra (expressing the complex exponentials as sines or hyperbolic sines where appropriate), we get

$$G(t - \tau) = \frac{\theta(t - \tau)\ e^{-\gamma(t-\tau)}}{\sqrt{|\nu^2 - \gamma^2|}} \begin{cases} \sin\Big(\sqrt{\nu^2 - \gamma^2}\ (t - \tau)\Big), & \gamma < \nu \\ \sinh\Big(\sqrt{\gamma^2 - \nu^2}\ (t - \tau)\Big), & \gamma > \nu \end{cases}$$

Finally, the solution to the ODE is given by

$$x(t) = \begin{cases} \displaystyle\int_{-\infty}^{t} \frac{e^{-\gamma(t-\tau)}}{\sqrt{|\nu^2 - \gamma^2|}} \sin\Big(\sqrt{\nu^2 - \gamma^2}\ (t - \tau)\Big) f(\tau) d\tau & \text{if } \gamma < \nu \\ \displaystyle\int_{-\infty}^{t} \frac{e^{-\gamma(t-\tau)}}{\sqrt{|\nu^2 - \gamma^2|}} \sinh\Big(\sqrt{\gamma^2 - \nu^2}\ (t - \tau)\Big) f(\tau) d\tau & \text{if } \gamma > \nu \end{cases}$$

□

Solution. Shorter Method. Construct the Green function based on the properties it must satisfy. We solve it for the case $\nu > \gamma$. The case $\nu < \gamma$ follows analogously. Given $\mathcal{L}G(t, \tau) = \delta(t - \tau)$, we see that $G(t, \tau)$ must satisfy:

(i.) $G(t, \tau)$ solves $\mathcal{L}G(t, \tau) = 0$ whenever $t \neq \tau$.

(ii.) $G(t,\tau)$ must satisfy the initial conditions.

(iii.) $G(t,\tau)$ must be continuous everywhere, including $t=\tau$, and the derivative of $G(t,\tau)$ must have a jump of magnitude unity at $t=\tau$.

Generalization of property (iii): In general, the Green function of an n^{th} order differential operator has a jump in the $(n-1)^{\text{st}}$ derivative at $t=\tau$. All further derivatives are continuous. The size of the jump is equal to the magnitude of the leading coefficient.

Let's use these properties to construct the Green function.

Step 1. Find candidate solutions by solving $\left(\frac{d^2}{dt^2}+2\gamma\frac{d}{dt}+\nu^2\right)G(t,\tau)=0$. We know there are two linearly independent solution because the ODE is linear and second order. The solutions are $A_1e^{-\gamma t-\sqrt{\nu^2-\gamma^2}t}$ and $A_2e^{-\gamma t+\sqrt{\nu^2-\gamma^2}t}$ where A_1 and A_2 are functions of τ. The solution in this form correct, but not very helpful. We take linear combinations of the two solutions and express them as $c_1e^{-\gamma(t-\tau)}\sin(\sqrt{\nu^2+\gamma^2}(t-\tau))$ and $c_2e^{-\gamma(t-\tau)}\cos(\sqrt{\nu^2+\gamma^2}(t-\tau))$. To construct the Green function, we must mix and match the two linearly independent candidate solutions. Tentatively, we can write

$$G(t,\tau)=\begin{cases}c_1e^{-\gamma(t-\tau)}\sin\left(\sqrt{\nu^2-\gamma^2}\ (t-\tau)\right)+c_2e^{-\gamma(t-\tau)}\cos\left(\sqrt{\nu^2-\gamma^2}\ (t-\tau)\right) & \text{if} \quad t<\tau\\ c_3e^{-\gamma(t-\tau)}\sin\left(\sqrt{\nu^2-\gamma^2}\ (t-\tau)\right)+c_4e^{-\gamma(t-\tau)}\cos\left(\sqrt{\nu^2-\gamma^2}\ (t-\tau)\right) & \text{if} \quad t>\tau\end{cases}$$

Step 2. Apply the initial conditions. Our initial conditions are $\lim_{t\to-\infty}G(t,\tau)=0$ and $\lim_{t\to-\infty}\dot{G}(t,\tau)=0$. The only candidate solution that satisfies the initial condition is the trivial solution. That is, $G(t,\tau)=0$ for $t<\tau$. We are not yet prepared to say what happens for $t>\tau$. We can improve our tentative Green function by

$$G(t,\tau)=\begin{cases}0 & \text{if} \quad t<\tau\\ c_3e^{-\gamma(t-\tau)}\sin\left(\sqrt{\nu^2-\gamma^2}\ (t-\tau)\right)+c_4e^{-\gamma(t-\tau)}\cos\left(\sqrt{\nu^2-\gamma^2}\ (t-\tau)\right) & \text{if} \quad t>\tau\end{cases}$$

Step 3. Apply the continuity and jump conditions. Given that $G(t,\tau)=0$ for $t<\tau$, we must now determine $G(t,\tau)$ for $t>\tau$. The continuity condition requires that $c_4=0$. We compute $\frac{\partial}{\partial t}G(t,\tau)=-c_3\gamma e^{-\gamma(t-\tau)}+c_3\sqrt{\nu^2-\gamma^2}e^{-\gamma(t-\tau)}\cos(\sqrt{\nu^2-\gamma^2}(t-\tau))$, and find that $\lim_{t\to\tau^+}=c_3\sqrt{\nu^2-\gamma^2}$. To ensure the jump is size unity, we must set $c_3=\sqrt{\nu^2-\gamma^2}$.

In summary, the Green function is given by

$$G(t-\tau)=\frac{\theta(t-\tau)e^{-\gamma(t-\tau)}}{\sqrt{\nu^2-\gamma^2}}\sin\left(\sqrt{\nu^2-\gamma^2}\right), \qquad \text{if} \quad \nu<\gamma,$$

and the solution to the ODE is

$$x(t)=\int_{-\infty}^{t}\frac{e^{-\gamma(t-\tau)}}{\sqrt{\nu^2-\gamma^2}}\sin\left(\sqrt{\nu^2-\gamma^2}\ (t-\tau)\right)f(\tau)d\tau, \qquad \text{if} \quad \gamma<\nu.$$

A similar calculation can be used to find the Green function and the solution for $\nu<\gamma$. □

Exercise 4.4. Follow the logic of Example 4.3.1 and suggest two methods of finding the Green function ((a) based on Fourier transform, and (b) based on properties of the Green function) for solving $\left(\frac{d^2}{dt^2}+\nu^2\right)^2 x(t) = f(t)$, where $\left(\frac{d^2}{dt^2}+\nu^2\right)^2 := \left(\frac{d^2}{dt^2}+\nu^2\right)\left(\frac{d^2}{dt^2}+\nu^2\right)$, at $t>0$, assuming that x is real-valued and $x(0^-) = \frac{d}{dt}x(0^-) = \frac{d^2}{dt^2}x(0^-) = \frac{d^3}{dt^3}x(0^-) = 0$.

4.3.4 Laplace Transform and Laplace Method

So far we have solved linear ODEs by using the Green function approach and constructing the Green function as a solution of the homogeneous equation with additionally prescribed initial conditions (one less than the order of the differential equation). In this section we discuss an alternative way of solving the problem via, first, application of the Laplace transform introduced in Section 3.9 for solving linear ODEs with constant coefficients, and then discussing the so-called Laplace method for solving linear ODEs with coefficients dependent linearly on the (time/space) variable. Connection between the two is not only via the name of Laplace, who has contributed developing both, but also due to the fact that the Laplace method can be considered as utilizing a generalization of the Laplace transform.

The Laplace transform is natural for solving dynamic problems with causal structure. Let us see how it works for finding the Green function defined by Eqs. (4.38,4.39). We apply the Laplace transform to Eq. (4.39), integrating it over time with the $\exp(-kt)$ Laplace weight from a small positive value, ϵ, to ∞. In this case, the integral of the right hand side is zero. Each term on the left hand side can be transformed through a sequence of integrations by parts to a product of a monomial in k with $\tilde{G}(k)$, the Laplace transform of $G(t)$. We also check all boundary terms which appear at $t=\epsilon$ and $t=\infty$. Assuming that $G(\infty)=0$ (which is always the case for stable systems), all contributions at $t=+\infty$ are equal to zero. All $t=\epsilon$ boundary terms, but one, are equal to zero, because $\forall 0 \le m < n-1, \quad d^m G(\epsilon)/dt^m = 0$. The only nonzero boundary contribution originates from $d^{n-1}G(\epsilon)/dt^{n-1} = 1$. Overall, one arrives at the following equation

$$L(k)\tilde{G}(k) = 1, \quad L(k) := \sum_{m=0}^{n} a_{m-k}(-k)^{n-m}. \tag{4.44}$$

Therefore, we just found that $G(k)$ has poles (in the complex plain of k) associated with zeros of the $L(k)$ polynomial. To find $G(t)$ one applies to $\tilde{G}(k)$ the inverse Laplace transform

$$G(t) = \int_{c-i\infty}^{c+\infty} \frac{dk}{2\pi i} \exp(kt)\tilde{G}(k). \tag{4.45}$$

The Laplace method allows us to solve ODEs where the coefficients are linear in t.

Remark. Notice, again, that the Laplace's method for differential equations is not to be confused with Laplace Transforms or Laplace's method for approximating integrals. They are not the same, even though related.

Consider an ODE that can be written as

$$\sum_{m=0}^{N} (a_m + b_m t) \frac{d^m y}{dt^m} = 0, \tag{4.46}$$

We look for solutions in the form of the integral

$$y(t) = \int_C dk\, Z(k) e^{kt}, \tag{4.47}$$

where $Z(k)$ is a function of the complex variable k and C is a contour in the complex plane of k that will depend on $Z(k)$ but that will not depend on t.

Remark. Notice, that the substitution (4.47) is similar to the inverse Laplace transform (4.45), however with an important difference that the contour C in the former does not necessarily coincides with the contour used in the latter. We remind that the contour used in the basic formula of the inverse Laplace transform, i.e. in Eq. (4.45), goes up, on the right to the imaginary axis.

The derivatives of y are computed from Eq. (4.47),

$$\frac{d^m y}{dt^m} = \int_C dk\, Z(k) k^m e^{kt}$$

and are substituted into the left hand side of equation 4.46 which gives

$$\int_C dk\, Z(k) \Big(\underbrace{(a_0 + a_1 k + \cdots + a_n k^n)}_{P(k)} + \underbrace{(b_0 + b_1 k + \cdots + b_n k^n)}_{Q(k)} t \Big) = 0$$

where we have introduced the notation $P(k)$ and $Q(k)$ for convenience. We integrate by parts to get

$$\begin{aligned} 0 &= \int_C dk\, Z(k) \Big(P(k) + Q(k) t \Big) e^{kt} \\ &= \Big[Z(k) Q(k) e^{kt} \Big]_{k_1}^{k_2} + \int_C dk\, \left(Z(k) P(k) - \frac{d}{dk} \Big(Z(k) Q(k) \Big) e^{kt} \right) \end{aligned} \tag{4.48}$$

where k_1 and k_2 represent the two endpoints of C. If we can pick $Z(k)$ and a contour C such that Eq. (4.48) holds, then we can use them to express the solution to the ODE (4.46) by Eq. (4.47). In summary, we must find $Z(k)$ and the contour C such that

$$\frac{d}{dk} \Big(Q(k) Z(k) \Big) - P(k) Z(k) = 0 \quad \text{and} \quad \Big[Z(t) Q(t) e^{kt} \Big]_{k_1}^{k_2} = 0.$$

The differential equation above, for $Z(k)$, can be solved either by finding an integrating factor, or by separation of variables, as follows:

$$d\,[QZ] = PZdk \quad \Rightarrow \quad \frac{d\,[QZ]}{QZ} = \frac{P}{Q}dk \quad \Rightarrow \quad \ln(QZ) = \int \frac{P}{Q}dk + \text{const}$$

$Z(k)$, is given by

$$Z(k) = \frac{c}{Q(k)} \exp\left(\int dk\, \frac{P(k)}{Q(k)}\right) \tag{4.49}$$

Once $Z(k)$ is determined, we must find a contour with endpoints k_1 and k_2 such that $Q(k_1)Z(k_1)e^{k_1 t} = Q(k_2)Z(k_2)e^{k_2 t}$.

Example 4.3.2. Use the Laplace's method to find the general solution to the boundary value problem:

$$x\frac{d^3}{dx^3}u + 2u = 0, \quad u(0) = 1, \quad u(\infty) = 0.$$

Solution. For this problem, we compute

$$P(k) = 2, \quad Q(k) = k^3, \quad Z(k) = -\frac{c\,e^{-1/k^2}}{k^3}, \quad Q(k)Z(k)e^{kt} = e^{kx-1/k^2} \tag{4.50}$$

We see that $e^{kx-1/k^2} \to 0$ as $k \to -\infty$ and that $e^{kx-1/k^2} = 0$ as $k \to 0$. Therefore, we take the contour of integration to be along the negative real axis. The solution is

$$u(x) = -2c \int_{-\infty}^{0} \frac{e^{kx-1/k^2}}{k^3}dk$$

which can be expressed as

$$u(x) = \int_0^{\infty} e^{-x/\sqrt{z}-z}dz$$

by the change of variables $z = t^{-2}$. The constant $c = 1/2$ was chosen to satisfy the boundary condition $u(0) = 1$

Exercise 4.5. Consider the sum

$$S(x) = \sum_{n=0}^{\infty} \frac{x^n}{(n!)^2}.$$

Find a second order, linear differential equation which, when supplied with proper initial conditions at $x = 0$, results in $S(x)$ as a solution. Solve the initial value problem by the Laplace method, therefore, representing $S(x)$ as an integral.

Example 4.3.3. (a) Use Laplace's method to find a general solution to the Hermite equation

$$\frac{d^2y}{dt^2} - 2t\frac{dy}{dt} + 2ny = 0. \tag{4.51}$$

(b) Simplify your result for the case where n is a non-negative integer.

Solution. (a) In this case we derive

$$P(k) = k^2 + 2n, \quad Q(k) = -2k, \quad Z(k) = -\frac{c\, e^{-k^2/4}}{2k^{n+1}}, \quad Q(k)Z(k)e^{kt} = \frac{e^{kt-k^2/4}}{k^n} \tag{4.52}$$

thus resulting in the following explicit solution of Eq. (4.51), defined up to a multiplicative constant, is

$$y(t) = \int_C dk\, \frac{e^{kt-k^2/4}}{k^{n+1}}. \tag{4.53}$$

Let us make the change in variables $k \to z$ according to $k = 2(t - z)$ which gives

$$y(t) = e^{t^2} \int_{C'} \frac{e^{-z^2} dz}{(z-t)^{n+1}}, \tag{4.54}$$

where C' is a suitable contour in the complex plane of z, which is yet undefined, that is we have a freedom in choosing the contour (as above we had a choice with contour C in the complex space of k).

(b) When n is a non-negative integer, the integrand in Eq. (4.54) has a simple pole, and thus choosing the contour to go around the pole both satisfies the requirement on the boundary terms and allows us to evaluate the integral by residue calculus. Applying Cauchy's formula to the resulting contour integral, one therefore arrives at the expression for the so-called Hermite polynomials

$$y(t) = H_n(t) = (-1)^n e^{t^2} \frac{d^n}{dt^n} e^{-t^2}, \tag{4.55}$$

where re-scaling (which is a degree of freedom in linear differential equations) is selected according to the normalization constraint introduced in the following exercise. □

Hermite polynomials will come back later in the context of the Sturm-Liouville problem in Section 4.5.3.

Example 4.3.4. Consider another example particular case of Eqs.(4.46) that can be solved by the Laplace method,

$$\frac{d^2}{dt^2} y - ty = 0. \tag{4.56}$$

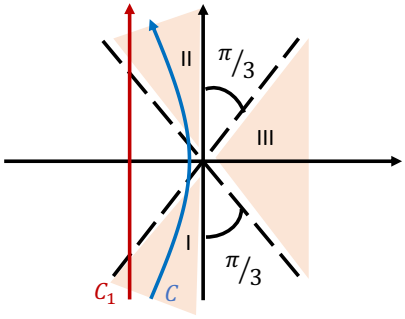


Figure 4.1: Layout of contours in the complex plane of k needed for saddle-point estimations of the Airy function described in Eq. (4.58).

Solution. Following the general Laplace method described above we derive

$$P(k) = k^2, \quad Q(k) = -1, \quad Z(k) = -\exp(-k^2/3). \tag{4.57}$$

According to Eq. (4.47) the general solution of Eq. (4.57) can be represented as

$$y(t) = \text{const} \int\limits_C dk \exp\left(kt - k^3/3\right), \tag{4.58}$$

where we choose an infinite integration path shown in Fig. (4.1) such that values of the integrand at the two (infinite) end points coincide (and equal to zero). Indeed, this choice guarantees that the infinite end points of the contour lie in the regions where $\mathrm{Re}(k^2) > 0$ (shaded regions I, II, III in Fig. (4.1)). Moreover, by choosing that the contour starts in the region I and ends in the region II (blue contour C in Fig. (4.1)) we guarantee that the Airy function given by Eq. (4.56) remains finite at $t \to +\infty$. Notice that the contour can be shifted arbitrarily under condition that the end points remain in the sectors I and II. In particular one can shift the contour to coincide with the imaginary axis (in the complex k plane shown in Fig. (4.1), then Eq. (4.58) becomes (up to a constant) the so-called Airy function

$$Ai(t) = \frac{1}{\pi} \int\limits_0^\infty dz \cos\left(\frac{z^3}{3} + zt\right) = \frac{1}{2\pi} \mathrm{Re}\left(\int\limits_{-\infty}^{\infty} dz \exp\left(i\frac{z^3}{3} + itz\right)\right). \tag{4.59}$$

Asymptotic expression for the Airy function at $t > 0$, $t \gg 1$, can be derived utilizing the saddle-point method described in Section 2.4. At $k = \pm\sqrt{t}$, the integrand in Eq. (4.58) has an extremum along the direction of its "steepest descent" from the saddle point along the imaginary axis. Since the contour end-points should stay in the sectors I and II, we shift the contour to the left from the imaginary axis while keeping it parallel to the imaginary

axis. (See C_1 shown in red in Fig. (4.1) which crosses the real axis at $k = -\sqrt{t}$.) The integral is dominated by the saddle-point at $k = -\sqrt{t}$, thus resulting (after substitution $k = \sqrt{t} + iz$, changing integration variable from k to z, making expansion over z, keeping quadratic term in z, ignoring higher order terms, and evaluating a Gaussian integral) in the following asymptotic estimation for the Airy function

$$t > 0,\ t \gg 1: \quad Ai(t) \approx \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp\left(-\frac{2}{3}t^{3/2} - \sqrt{t}z^2\right) dz = \frac{\exp(-2t^{3/2}/3)}{t^{1/4}\sqrt{4\pi}}. \tag{4.60}$$

(Notice that one can also provide an alternative argument and exclude contribution of the second, potentially dominating, saddle-point $k = \sqrt{t}$ simply by observing that Gaussian integral evaluated along the steepest descent path from this saddle-point gives zero contribution after evaluating the real part of the result, as required by Eq. (4.59).)

4.4 Linear Static Problems

We will now turn to problems which normally appear in the static case. In many natural and engineered systems, a dynamic system that reaches equilibrium may have spatial characteristics that are non-trivial and worthy of analysis. Here we discuss a number of linear spatially one-dimensional problems that are relevant to applications.

4.4.1 One-Dimensional Poisson Equation

Poisson's equation describes, in the case of electrostatics, the potential field caused by a given charge distribution.

Let us discuss the function $u(x)$ whose distribution over a finite spatial interval is described by the following set of equations

$$\frac{d^2}{dx^2}u(x) = f(x), \quad \forall x \in (a,b) \quad \text{with} \quad u(a) = u(b) = 0. \tag{4.61}$$

We introduce the Green function which satisfies

$$\forall a < x, y < b: \quad \frac{d^2}{dx^2}G(x;y) = \delta(x-y), \quad G(a;y) = G(b;y) = 0. \tag{4.62}$$

Notice that the Green function now depends on both x and y.

According to Eq. (4.62), $\frac{d^2}{dx^2}G(x;y) = 0$ if $x \neq y$. That is, $G(x,y)$ is a linear function of x for all $x \neq y$. Then enforcing the boundary conditions one derives

$$x > y: \quad G(x;y) = B(x-b), \tag{4.63}$$

$$y > x: \quad G(x;y) = A(x-a). \tag{4.64}$$

Furthermore, given that the differential equation in (4.62) is the second order, $G(x, y)$ should be continuous at $x = y$ and the jump of its first derivative at $x = y$ should be equal to unity. Summarizing, one finds

$$G(x; y) = \frac{1}{b-a} \begin{cases} (y-b)(x-a), & x < y \\ (y-a)(x-b), & x > y. \end{cases} \tag{4.65}$$

The solution the Eq. (4.61) is given by the convolution operator

$$u(x) = \int_a^b dy G(x; y) f(y). \tag{4.66}$$

Example 4.4.1. Find the Green function for the equation, $\mathcal{L}u(x) = f(x)$, where the operator $\mathcal{L} = -\frac{d}{dx}\left(x\frac{d}{dx}\right)$, and the boundary conditions on $u(x)$ are $u(1) = 0$ and $u'(2) = 0$.

Solution. We construct the Green function by finding a function that satisfies the necessary properties. *Property (i.): The Green function satisfies* $\mathcal{L}G(x; y) = 0$ *for* $x \neq y$. Two linearly independent solutions to the homogeneous equation are $u(x) = \text{const}$ and $u(x) = \log(x)$. Therefore for any y in $1 < y < 2$, we can write

$$G(x; y) = \begin{cases} c_1 + c_2 \log(x), & x < y \\ c_3 + c_4 \log(x), & x > y \end{cases}$$

Property (ii.): The Green function satisfies the boundary conditions. Enforcing the boundary conditions $u(1) = 0$ and $u'(2) = 0$, gives $c_1 = c_4 = 0$,

$$G(x; y) = \begin{cases} c_2 \log(x) & x < y \\ c_3 & x > y \end{cases}$$

Property (iii.): $G(x; y)$ is continuous and $G_x(x; y)$ has a jump of magnitude $-1/y$ at $x = y$. To ensure continuity at y, we must set $c_2 \log(y) = c_3$. This ensures continuity, but does not yet give the appropriate jump condition. Computing the derivative of G as $x \to y^-$ and as $x \to y^+$ gives $\lim_{x\to y^-} G_x = c_2/y$ and $\lim_{x\to y^+} G_x = 0$. That is, we must set $c_2 = 1$ to ensure that the derivative has a jump with magnitude $-1/y$ at $x = y$.

$$G(x; y) = \begin{cases} \log(x), & x < y \\ \log(y), & x > y \end{cases}$$

□

Remark. Explanation for Property (iii.): To determine the magnitude of the jump at $x = y$, integrate the ODE on the interval $y - \epsilon < x < y + \epsilon$ and take the limit $\epsilon \to 0$.

$$-\frac{d}{dx}\left(x\frac{d}{dx}\right)u(x) = \delta(x-y)$$

$$\lim_{\epsilon\to 0}\int_{y-\epsilon}^{y+\epsilon} -\frac{d}{dx}\left(x\frac{d}{dx}\right)u(x)dx = \lim_{\epsilon\to 0}\int_{y-\epsilon}^{y+\epsilon}\delta(x-y)dx$$

$$\lim_{\epsilon\to 0}\left[-x\frac{d}{dx}u(x)\right]_{y-\epsilon}^{y+\epsilon} = 1$$

$$\left.\frac{d}{dx}u(x)\right|_{x=y} = -\frac{1}{y}$$

Exercise 4.6. Find the Green function for the equation, $\mathcal{L}u(x) = f(x)$, where the operator $\mathcal{L} = -\frac{d^2}{dx^2} - \kappa^2$, and the boundary conditions on $u(x)$ are

(a) $u(0) = u'(1) = 0$;

(b) $u(x)$ is periodic with the period 2π.

4.5 Sturm–Liouville (spectral) theory

We enter the study of differential operators which map a function to another function, and it is therefore imperative to first discuss the Hilbert space where the functions of reside.

4.5.1 Hilbert Space and its completeness

Let us first review some basic properties of a Hilbert space, in particular, condition on its completeness. (These will be discussed at greater length in the companion Math 527 course of the AM core.) A linear (vector) space is called a Hilbert space, $\mathcal{H}$, if

1. For any two elements, f and g there exists a a scalar product (f, g) which satisfies the following properties:

   (a) linear with respect to the second argument,

   $$(f, \alpha g_1 + \beta g_2) = \alpha(f, g_1) + \beta(f, g_2),$$

   for any $f, g_{1.2} \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{C}$.

   (b) self-conjugation (Hermitian)

   $$(f, g) = (g, f)^*;$$

(c) non-negativity of the norm, $\|f\|^2 := (f,f) > 0$, where $(f,f) = 0$ means $f = 0$.

2. $\mathcal{H}$ has a countable basis, B, i.e. a countable number of elements, $B := \{f_n,\ n = 1, \cdots, \infty\}$ such that any element $g \in \mathcal{H}$ can be represented in the form of a linear combination f_n. that is, for any $g \in \mathcal{H}$, there exist coefficients c_n such that $g = \sum c_n f_n$.

Remark. The Hilbert space defined above for complex-valued functions can also be considered over real-valued functions. In the following we will use the two interchangeably.

Any basis B can be turned into an ortho-normal basis with respect to a given scalar product, i.e. $x = \sum\limits_{n=1}^{\infty} (x, f_n) f_n, \quad \|x\|^2 = \sum\limits_{n=1}^{\infty} |(x, f_n)|^2$. (For example, the Gram-Schmidt process is a standard ortho-normalization procedure.)

One primary example of a Hilbert space is the $L^2(\Omega)$ space of complex-valued functions $f(x)$ defined in the space $\Omega \in \mathbb{R}^n$ such that $\int\limits_{\Omega} dx |f(x)|^2 < \infty$ (one may say, casually, that the square modulus of the function is integrable). In this case the scalar product is defined as

$$(f,g) := \int\limits_{\Omega} dx f^*(x) g(x).$$

Properties 1a-c from the definition of Hilbert space above are satisfied by construction and property 2 can be proven (it is a standard proof in the course of mathematical analysis).

Consider a fixed infinite ortho-normal sequency of functions

$$\{f_n,\ n = 1, \cdots, \infty,\ (f_n, f_m) = \delta_{nm}\}.$$

The sequence is a basis in $L^2(\Omega)$ iff the following relation of completeness holds

$$\sum_{n=1}^{\infty} f_n^*(x) f_n(y) = \delta(x - y). \tag{4.67}$$

As custom for the δ function (an other generalized functions), Eq. (4.67) should be understood as equality of integrals of the two sides of Eq. (4.67) integrated with a function from $L^2(\Omega)$.

4.5.2 Hermitian and non-Hermitian Differential Operators

Consider a function from the Hilbert space $L^2(a,b)$ over the reals, i.e. function of a single variable, $x \in \mathbb{R}$, over a bounded domain, $a \le x \le b$ with an integrable square modulus and a linear differential operator $\hat{L}$ acting on the function.

A differential operator is called Hermitian (self-conjugated) if for any two functions (from a certain class of interest, e.g. from $L^2(a,b)$) the following relation holds:

$$(f,\hat{L}g) := \int_a^b dx\, f(x)\hat{L}g(x) = \int_a^b dx\, g(x)\hat{L}f(x) = (g,\hat{L}f). \tag{4.68}$$

It is clear from how the condition (4.68) was stated that it depends on both the class of functions and on the operator $\hat{L}$. For example, considering functions f and g with zero boundary conditions or functions which are periodic and which derivative is periodic too, will result in the statement that the operator

$$\hat{L} = \frac{d^2}{dx^2} + U(x), \tag{4.69}$$

where $U(x)$ is a function mapping from $\mathbb{R}$ to $\mathbb{R}$, is Hermitian.

The natural generalization of the Shrödinger operator 4.69 is the Sturm-Liouville operator

$$\hat{L} = \frac{d^2}{dx^2} + Q\frac{d}{dx} + U(x). \tag{4.70}$$

The Sturm-Liouville operator is not Hermitian, i.e. Eq. (4.68) does not hold in this case. However, it is straightforward to check that at the zero boundary conditions or periodic boundary conditions imposed on the functions, $f(x)$ and $g(x)$, and their derivatives, the following generalization of Eq. (4.68) holds

$$\int_a^b dx\rho(x)f(x)\hat{L}g(x) = \int_a^b dx\rho(x)g(x)\hat{L}f(x), \tag{4.71}$$

$$\text{where} \quad \frac{d}{dx}\rho = Q\rho \Rightarrow \rho = \exp\left(\int dxQ\right). \tag{4.72}$$

Consider now the eigen-functions f_n of the operator $\hat{L}$, which satisfy

$$\hat{L}f_n = \lambda_n f_n, \tag{4.73}$$

where λ_n is the spectral parameter (eigenvalue) of the eigen-function, f_n, of the Sturm-Liouville operator (4.70), indexed by n. (We assume that, $\forall n \neq m: \quad \lambda_n \neq \lambda_m$.)

Notice that the value of λ_n is not specified in Eq. (4.73) and finding the values of λ_n for which there exists a non-trivial solution, satisfying respective boundary conditions (describing the class of functions considered) is an instrumental part of the Sturm-Liouville problem.

Observe that the conditions (4.71,4.72) translates into

$$\int dx\rho f_n\hat{L}f_m = \lambda_m \int dx\rho f_n f_m = \lambda_n \int dx\rho f_n f_m,$$

that becomes the following eigen-function orthogonality condition

$$\int dx\rho f_n f_m = 0 \quad (\forall n \neq m)\,. \tag{4.74}$$

As a corollary of this statement one also finds that in the Hermitian case the distinct eigen-functions are orthogonal to each other with unitary weight, $\rho = 1$.

Let us check Eq. (4.74) on the example, $\hat{L}_0 = d^2/dx^2$, where $Q(x) = U(x) = 0$, over the functions which are 2π-periodic. $\cos(nx)$ and $\sin(nx)$, where $n = 0, 1, \cdots$ are distinct eigen-functions with the eigen-values, $\lambda_n = -n^2$. Then, forall $m \neq n$,

$$\int_0^{2\pi} dx \cos(nx)\cos(mx) = \int_0^{2\pi} dx \cos(nx)\sin(mx) = \int_0^{2\pi} dx \sin(nx)\sin(mx) = 0. \tag{4.75}$$

Note that the example just discussed has a degeneracy: $\cos(nx)$ and $\sin(nx)$ are two distinct real eigen-functions corresponding to the same eigen-value. Therefore, any combination of the two is also an eigen-function corresponding to the same eigen-value. If we would choose any other pair of the degenerate eigen-functions, say $\cos(nx)$ and $\sin(nx) + \cos(nx)$, the two would not be orthogonal to each other. Therefore, what we see on this example is that the eigen-functions corresponding to the same-eigenvalue should be specially selected to be orthogonal to each other.

We say that the set of eigen-functions, $\{f_n(x)|n \in \mathbb{N}\}$, of $\hat{L}$ is complete over a given class (of functions) if any function from the class can be expanded into the series over the eigen-functions from the set

$$f = \sum_n c_n f_n. \tag{4.76}$$

Relating this eigen-functions' property to completeness of the Hilbert space basis, one observes that eigen-vectors of a self-adjoint (Hermitian) operator over $L^2(\Omega)$ form an orthonormal basis of $L^2(\Omega)$.

Multiplying both sides of Eq. (4.76) by ρf_n, integrating over the domain, and applying (4.74) to the right one derives

$$c_n = \frac{\int dx\rho f_n f}{\int dx\rho (f_n)^2}. \tag{4.77}$$

Note that for the example $\hat{L}_0$, Eq. (4.76) is a Fourier Series expansion of a periodic function.

Returning to the general case and substituting Eq. (4.77) back into (4.76), one arrives at

$$f(x) = \int dy \left(\rho(y) \sum_n \frac{f_n(x) f_n(y)}{\int dx\rho(x)(f_n(x))^2} \right) f(y). \tag{4.78}$$

If the set of functions $\{f_n(x)|n\}$ is complete relation (4.78) should be valid for any function f from the considered class. Consistently with this statement one observes that the part of the integrand in Eq. (4.78) is just the $\delta(x)$, which is the special function which maps convolution of the function to itself, i.e.

$$\sum_n \frac{f_n(x)f_n(y)}{\int dx\rho(x)(f_n(x))^2} = \frac{1}{\rho(y)}\delta(x-y). \tag{4.79}$$

Therefore, one concludes that Eq. (4.79) is equivalent to the statement of the set of functions $\{f_n(x)|n\}$ completeness.

Example 4.5.1. Check validity of Eq. (4.79), and thus completeness of the respective set of eigen-functions, for our enabling example of $\hat{L}_0 = d^2/dx^2$ over the functions which are 2π-periodic.

4.5.3 Hermite Polynomials.

Let us now depart from our enabling example and consider the case of $Q(x) = -2x$ and $U(x) = 0$, i.e.

$$\hat{L}_2 = \frac{d^2}{dx^2} - 2x\frac{d}{dx}, \quad \rho(x) = \exp\left(-x^2\right), \tag{4.80}$$

over the class of functions mapping from $\mathbb{R}$ to $\mathbb{R}$, which also decay sufficiently fast at $x \to \pm\infty$. That is we are discussing now

$$\hat{L}_2 f_n = \lambda_n f_n. \tag{4.81}$$

Changing from $f_n(x)$ to $\Psi_n(x) = f_n(x)\sqrt{\rho}$ one thus arrives at the following equation for Ψ_n:

$$e^{-x^2/2}\hat{L}_2 f_n(x) = e^{-x^2/2}\hat{L}_2\left(e^{x^2/2}\Psi_n(x)\right) = \frac{d^2}{dx^2}\Psi_n + (1-x^2)\Psi_n = \lambda_n\Psi_n. \tag{4.82}$$

Observe that when $\lambda_n = -2n$, Eq. (4.81) coincides with the Hermite Eq. (4.51).

Let us look for solution of Eq. (4.81) in the form of the Taylor series around $x = 0$

$$f_n(x) = \sum_{k=0}^{\infty} a_k x^k. \tag{4.83}$$

Substituting the series into the Hermite equation and then equating terms for the same powers of x one arrives at the following regression for the expansion coefficients:

$$\forall k = 0, 1, \cdots : \quad a_{k+2} = \frac{2k+\lambda_n}{(k+2)(k+1)}a_k. \tag{4.84}$$

This results in the following two linearly independent solutions (even and odd, respectively, with respect to the $x \to -x$ transformation) of Eq. (4.81) represented in the form of a series

$$f_n^{(e)}(x) = a_0 \left(1 + \frac{\lambda_n}{2!}x^2 + \frac{\lambda_n(4+\lambda_n)}{4!}x^4 + \cdots\right), \tag{4.85}$$

$$f_n^{(o)}(x) = a_1 \left(x + \frac{(2+\lambda_n)}{3!}x^3 + \frac{(2+\lambda_n)(6+\lambda_n)}{5!}x^5 + \cdots\right), \tag{4.86}$$

where the two first coefficients in the series (4.83) are kept as the parameters. Observe that the series (4.85) and (4.86) terminate if $\lambda_n = -4n$ and $\lambda_n = -4n-2$, respectively, where $n = 0, 1, \cdots$, then $f_n^{(e)}$ are polynomials – in fact the Hermite polynomials. We combine the two cases in one and use the standard, $H_n(x)$, notations for the Hermite polynomials of the n-th order, which satisfies Eq. (4.82). Per statement of the Exercise 4.7, Hermite polynomials are normalized and orthogonal (weighted with ρ) to each other.

Exercise 4.7. (a) Prove that

$$\int_{-\infty}^{+\infty} dt\, e^{-t^2} H_n(t) H_m(t) = 2^n n! \sqrt{\pi}\delta_{nm}, \tag{4.87}$$

where δ_{nm} is unity when $n = m$ and it is zero otherwise (Kronecker symbol).
(b) Verify that the set of functions

$$\left\{ \Psi_n(x) = \frac{1}{\pi^{1/4}\sqrt{2^n n!}} \exp(-x^2/2) H_n(x) \;\middle|\; n = 0, 1, \cdots \right\}, \tag{4.88}$$

satisfies

$$\sum_{n=0}^{\infty} \Psi_n(x)\Psi_n(y) = \delta(x-y). \tag{4.89}$$

Hint: The following identity may be useful

$$\frac{d^n}{dx^n}\exp(-x^2) = \sqrt{\pi}\int_{-\infty}^{+\infty} \frac{dq}{2\pi}(iq)^n \exp\left(-q^2/4 + iqx\right).$$

A corollary of the Exercise 4.7 is the statement of "completeness": the set of functions (4.88) forms an orthogonal basis of the Hilbert space of functions, $f(x) \in L^2$, i.e. satisfying $\int_{-\infty}^{\infty} |f(x)|^2 dx < \infty$. (A bit more formally, an orthogonal basis for the L^2 functions is a complete orthogonal set. For an orthogonal set, completeness is equivalent to the fact that the 0 function is the only function $f \in L^2$ which is orthogonal to all functions in the set.)

Note that the last equality in Eq. (4.82) is the spectral version of the the Schrödinger PDE (in the imaginary time) in the quadratic potential

$$\partial_x^2 \Psi(t;x) + (1-x^2)\Psi(t;x) = -\partial_t \Psi(t;x), \tag{4.90}$$

discussed next.

4.5.4 Case study: Schrödinger Equation in $1d$ *

Schrödinger equation *

$$\frac{d^2\Psi(x)}{dx^2} + (E - U(x))\Psi(x) = 0, \tag{4.91}$$

described the so-called (complex-valued) wave function describing de-location of a quantum particle in $x \in \mathbb{R}$ with energy E in the potential $U(x)$. We are seeking for solutions with $|\Psi(x)| \to 0$ at $x \to \infty$ and our goal here is do describe the spectrum (allowed values of E) and respective eigen-functions.

As a simple, but instructive, example consider the case of a quantum particle in a rectangular potential, i.e. $U(x) = U_0$ at $x \notin [0, a]$ and zero otherwise. General solution of Eq. (4.91) becomes

$$\underline{U_0 > E > 0:}$$
$$\Psi_E(x) = \begin{cases} c_L \exp(x\sqrt{U_0 - E}), & x < 0 \\ a_+ \exp(ix\sqrt{E}) + a_- \exp(-ix\sqrt{E}), & x \in [0, a] \\ c_R \exp(-x\sqrt{U_0 - E}), & x > a \end{cases}, \tag{4.92}$$
$$\underline{U_0 < E:}$$
$$\Psi_E(x) = \begin{cases} c_{L+} \exp(ix\sqrt{E - U_0}) + c_{L-} \exp(-ix\sqrt{E - U_0}), & x < 0 \\ a_+ \exp(ix\sqrt{E}) + a_- \exp(-ix\sqrt{E}), & x \in [0, a] \\ c_{R+} \exp(ix\sqrt{E - U_0}) + c_{R-} \exp(-ix\sqrt{E - U_0}), & x > a \end{cases}, \tag{4.93}$$

where we account for the fact that E cannot be negative (ODE simply does not allow such solutions) and in the $U_0 > E > 0$ regime we select one solution (of the two linearly independent solutions) which does not grow with $x \to \pm\infty$.

The solutions in the three different intervals should be "glued" together - or stating it less casually Ψ and $d\Psi/dx$ should be continuous at all $x \in \mathbb{R}$. These conditions applied to Eq. (4.92) or Eq. (4.93) result in an algebraic "consistency" conditions for E. We expect to get a continuous spectrum at $E > U_0$ and discrete at $U_0 > E > 0$.

Example 4.5.2. Complete calculations above for the case of $U_0 > E > 0$ and find the allowed values of the discrete spectrum. What is the condition for appearance of at least one discrete level?

Consider another example.

*This auxiliary Subsection can be dropped at the first reading. Material form the Subsection will not contribute midterm and final exams.

Example 4.5.3. Find eigen-functions and energy of stationary states of the Schrödinger equation for an oscillator:

$$\frac{d^2\Psi(x)}{dx^2} + (E - x^2)\Psi(x) = 0, \tag{4.94}$$

where $x \in \mathbb{R}$ and $\Psi : \mathbb{R} \to \mathbb{C}^2$.

Solution. As we saw already in the preceding section analysis of Eq. (4.94) is reduced to studying the Hermite equation, with its spectral version desribed by Eq. (4.81). However, we will follow another route here. Let us introduce the so-called "creation" and "annihilation" operators

$$\hat{a} = \frac{i}{\sqrt{2}}\left(\frac{d}{dx} + x\right), \quad \hat{a}^\dagger = \frac{i}{\sqrt{2}}\left(\frac{d}{dx} - x\right), \tag{4.95}$$

and then rewrite the Schrödinger Eq. (4.94) as

$$\hat{H}\Psi(x) = \hat{a}^\dagger\hat{a}\Psi(x) = \left(2E - \frac{1}{2}\right)\Psi(x). \tag{4.96}$$

It is straightforward to check that the operator $\hat{H}$ is positive definite for all functions from L^2:

$$\int dx\Psi^\dagger(x)\hat{H}\Psi(x) = \int dx\Psi^\dagger(x)\hat{a}\hat{a}^\dagger\Psi(x) = \int dx|\hat{a}\Psi(x)|^2 \geq 0,$$

where the equality is achieved only if

$$\hat{a}\Psi_0(x) = \frac{i}{\sqrt{2}}\left(\frac{d}{dx} + x\right)\Psi_0(x) = 0,$$

thus resulting in $\Psi_0(x) = A\exp(-x^2/2)$ and $E_0 = 1/4$. We have just found the eigen-function and eigen-value correspondent to the lowest possible energy, so-called ground state. To find all other eigen-function, correspondent to the so-called "excited" states, consider the so-called commutation relations

$$\hat{a}\hat{a}^\dagger\Psi(x) = \hat{a}^\dagger\hat{a}\Psi(x) + \Psi(x), \tag{4.97}$$

$$\begin{aligned}\hat{a}^\dagger\hat{a}\left(\hat{a}^\dagger\right)^n\Psi(x) &= \left(\hat{a}^\dagger\right)^2\hat{a}\left(\hat{a}^\dagger\right)^{n-1}\Psi(x) + \left(\hat{a}^\dagger\right)^n\Psi(x) \\ &= n\left(\hat{a}^\dagger\right)^n\Psi(x) + \left(\hat{a}^\dagger\right)^{n+1}\hat{a}\Psi(x).\end{aligned} \tag{4.98}$$

Introduce $\Psi_n(x) := (\hat{a}^\dagger)^n\Psi_0(x)$. Since $\hat{a}\Psi_0(x) = 0$, the commutation relations (4.98) shows immediately that

$$\left(2E - \frac{1}{2}\right)\Psi_n(x) = \hat{H}\Psi_n(x) = \hat{a}^\dagger\hat{a}\Psi_n(x) = n\Psi_n(x).$$

We observe that eigen-functions $\Psi_n(x)$ of the states with energies, $2E_n = n + 1/2$ are expressed via the Hermite polynomials, $H_n(x)$, introduced in Eq. (4.55),

$$\begin{aligned}\Psi_n(x) &= A_n \left(\frac{i}{\sqrt{2}}\left(\frac{d}{dx} - x\right)\right)^n \exp\left(-\frac{x^2}{2}\right) \\ &= A_n \frac{i^n}{2^{n/2}} \exp\left(\frac{x^2}{2}\right) \frac{d^n}{dx^n} \exp\left(-x^2\right),\end{aligned}$$

where we have used the identity, $(\frac{d}{dx} - x)\exp(x^2/2) = \exp(x^2/2)\frac{d}{dx}$. From the condition of the Hermite polynomials orthogonality (4.87) one derives, $A_n = (n!\sqrt{\pi})^{-1/2}$.

4.6 Phase Space Dynamics for Conservative and Perturbed Systems

4.6.1 Integrals of Motion

Consider equation describing conservative dynamics of a particle of unit mass in the potential (conservative means there is no dissipation of energy)

$$\dot{x} = v, \quad \dot{v} = -\partial_x U(x). \tag{4.99}$$

The energy of the particle is

$$E = \frac{\dot{x}^2}{2} + U(x), \tag{4.100}$$

which consists of the kinetic energy (the first term), and the potential energy (the second term). It is straightforward to check that the energy is constant, that is $dE/dt = 0$. Therefore,

$$\dot{x} = \pm 2\sqrt{E - U(x)}, \tag{4.101}$$

where $\pm$ on the right hand side is chosen according to the initial condition chosen for $\dot{x}(0)$ (there may be multiple solutions, corresponding to the same energy). Eq. (4.101) is separable, and it can thus be integrated resulting in the following classic implicit expression for the particle coordinate as a function of time

$$\int_{x_0} \frac{dx}{\sqrt{E - U(x)}} = \pm t, \tag{4.102}$$

which depends on the particle's initial position, x_0, and its energy, E which is conserved.

In the example above, E is an *integral of motion* or equivalently a *first integral*, which is defined as a quantity that is conserved along solutions to the differential equation. In this case E was constant along the trajectories $x(t)$.

The idea of an integral of motion or first integral extends to conservative systems described by a system of ODEs. (Here and in the next section we follow [2, 1].) For example, consider the situation where a quantity H, called Hamiltonian, is a twice-differentiable function of $2n$ variables, $p_1, \cdots, p_n$ (momenta) and $q_1, \cdots, q_n$ (coordinates). Corresponding system of the dynamic equations is called the Hamilton's canonical equations,

$$\dot{p}_i = -\frac{\partial H}{\partial q_i}, \quad \dot{q}_i = \frac{\partial H}{\partial p_i} \left(\forall i = 1, \cdots, N\right). \tag{4.103}$$

Computing the rate of change of the Hamiltonian in time

$$\frac{dH}{dt} = \sum_{i=1}^{N} \left(\frac{\partial H}{\partial p_i}\dot{p}_i + \frac{\partial H}{\partial q_i}\dot{q}_i \right) = \sum_{i=1}^{N} \left(-\dot{q}_i\dot{p}_i + \dot{p}_i\dot{q}_i\right), \tag{4.104}$$

we observe that H is constant, that is, H is an integral of motion.

This dynamical system with a single degree of freedom ("single" particle), described by (4.99), is an example of a canonical Hamilton system. Energy (4.100) of the Hamiltonian system, considered as a function of x and v, is the Hamiltonian. x and v correspond to (scalar) q and p respectively. We continue exploring a single particle, $N = 1$, Hamiltonian system in the next Subsection – Section 4.6.2.

We will also discuss Hamiltonian systems, as derived from the variational principle, in the optimization part of the course (early second semester). We reiterate that reader interested in a broader and comprehensive mathematical introduction into the subject of the Hamiltonian dynamics is advised to consult with [2].

4.6.2 Phase Portrait

Following [2] and Section 1.3 of [1] we now turn to discussing the famous example of a conservative (Hamiltonian) system with one degree of freedom (4.99).

We have established above that the energy (Hamiltonian) is conserved, and it is thus instructive to study isolines, or level curves, of the energy drawn in the two-dimensional (x, v) space, $\{\{x, v\} \mid \frac{v^2}{2} + U(x) = E\}$. To draw a level curve of energy we simply fix E and evaluate how $\{x, v\}$ evolves with t according to Eqs. (4.99).

Let us build some intuition for level curves using an analogy. Suppose that the potential curve is the same shape as a length of wire (literally the same shape, i.e. if the potential curve is a parabola the wire is shaped like a parabola). We will say that this wire is perfectly rigid, and frictionless. Now imagine that there is a ball or bead which slides along the wire, subject to gravity. One may start the bead at any position on the wire, with any initial velocity (left or right). The path that the bead traces out in position-velocity space is (qualitatively) a level curve of the corresponding potential function.

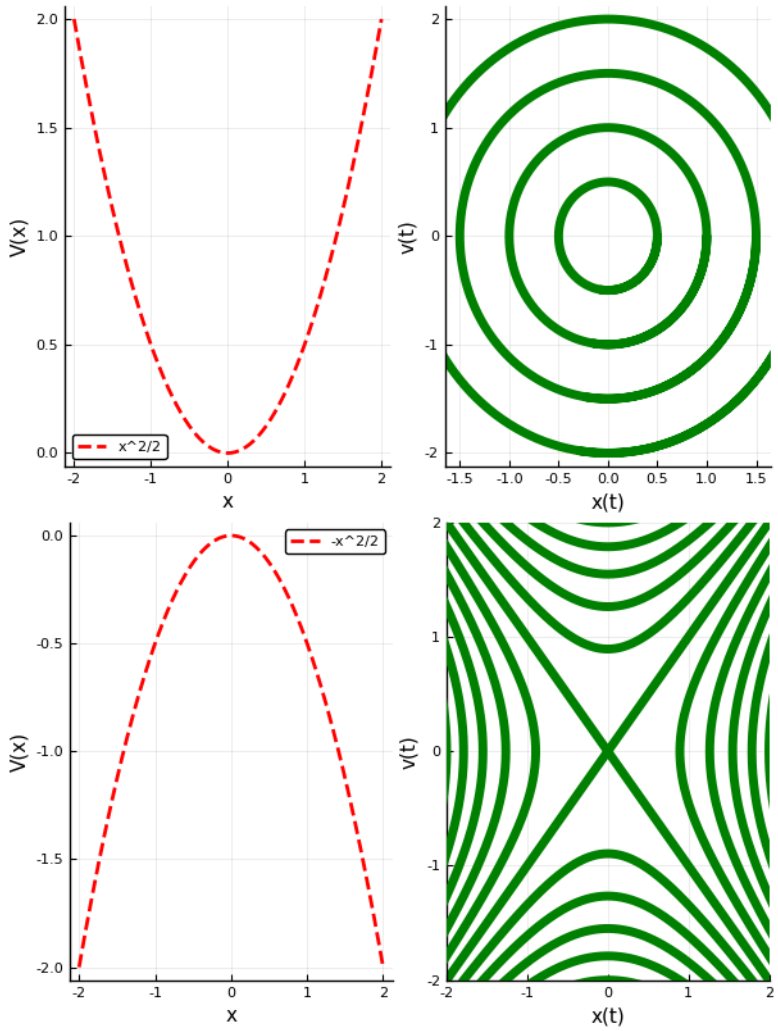


Figure 4.2: Phase portrait, i.e. (x, v) level-curves of the conservative system Eq. (4.99) with the potential, $U(x) = kx^2/2$ with $k > 0$ (top) and $k < 0$ (bottom).

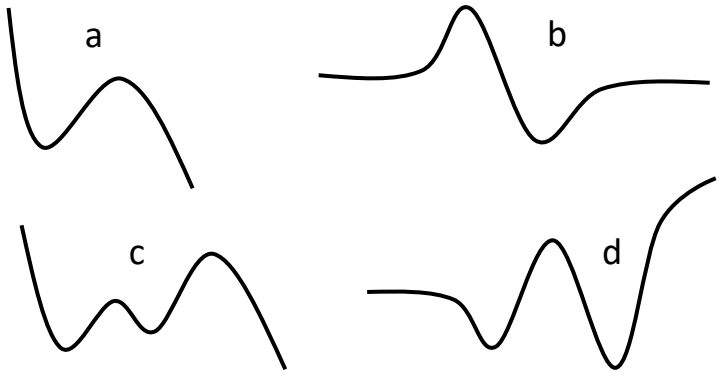


Figure 4.3: What is appearance of the level curves (phase portrait) of the energy for each of these potentials?

Consider the quadratic potential, $U(x) = \frac{1}{2}kx^2$. The two cases of positive and negative k are illustrated in Fig. (4.2), see the snippet *Portrait.ipynb*. We observe that with the exception of the equilibrium position $(x, v) = (0, 0)$, the level curves of the energy are smooth. Generalizing, we find that the exceptional points are critical, or stationary, points of the Hamiltonian, which are points where the derivatives of the Hamiltonian with respect to the canonical variables, q and p, are zero. Note that each level curve, which we draw observing how a particle slides in a potential well, $U(x)$, also has a direction (not shown in Fig. (4.2)).

Consider the case where $k > 0$, and fix the value of the energy E. Due to Eq. (4.100), the coordinate of the particle, x, should lie within the set where the potential energy is less than the energy, $\{x \mid U(x) \leq E\}$. We observe that $E \geq 0$, and that equality corresponds to the particle sitting still at the minimum of the potential, which is called a critical point, or fixed point. Furthermore, the larger the kinetic energy, the smaller the potential energy. Any position where the particle changes its velocity from positive to negative or vice-versa is called a turning point. For any $E > 0$, there are two turning points, $x_\pm = \pm 2E/k$. Testing different values of $E > 0$, we sketch different level curves, resulting in different ellipsoids centered around 0. This is the canonical example of a oscillator. The motion of the particle is periodic, and its period, T, can be computed evaluating Eq. (4.102) between the turning points

$$T := \int_{x_-}^{x_+} \frac{dx}{\sqrt{E - U(x)}} = \int_{-\sqrt{2E/k}}^{\sqrt{2E/k}} \frac{dx}{\sqrt{E - kx^2/2}} = 2\pi. \tag{4.105}$$

For this case, the period is a constant, 2π, and we note that it is independent of k.

In the $k < 0$ case where all the values of energy (positive and negative) are accessible, $x = v = E = 0$ is the critical point again. When $E > 0$ there are no turning points (points where direction of the velocity changes). When $E > 0$ the particle may turn only once or not at all. If $x(0) \neq 0$ and regardless of the sign of E, $x(t)$ increases with t to become unbounded at $t \to \infty$. As seen in Fig. (4.2)b, in this case the (x, v) phase space splits into four quadrants, separated by the $v = \pm\sqrt{kx}$ separatrices. The level curves of the energy are hyperbolas centered around $x = v = 0$.

A qualitative study of the dynamics in more complex potentials $U(x)$ can be conducted by sketching the level curves in a similar way.

Example 4.6.1. Sketch level curves of the energy for the Kepler potential, $U(x) := -\frac{1}{x} + \frac{C}{x^2}$, and for the potentials shown in Fig. (4.3).

4.6.3 Small Perturbation of a Conservative System

Let us analyze the following simple but very instructive example of a system which deviates very slightly from the quadratic potential with $k = 1$:

$$\dot{x} = v + \varepsilon f(x, v), \quad \dot{v} = -x + \varepsilon g(x, v), \tag{4.106}$$

in the regime where $\varepsilon \ll 1$ and $x^2 + v^2 \leq R^2$.

For $\varepsilon = 0$, and assuming that $x(0) = x_0$, one derives

$$x(t) = x_0 \cos(t), \quad v(t) = -x_0 \sin(t).$$

We calculate the energy and find that $E = (x^2 + v^2)/2$ (within the limit) is conserved and so the system cycles with the period given by $T = 2\pi$.

The general case where $0 < \varepsilon \ll 1$ is not conservative. Let us examine how the energy changes with time. One derives

$$\frac{d}{dt}E = x\dot{x} + v\dot{v} = \varepsilon\,(xf + vg) = \varepsilon\left(x^{(0)}f + v^{(0)}g\right) + O(\varepsilon^2). \tag{4.107}$$

Integrating over a period, one arrives at the following expression for the gain (or loss) of energy

$$\Delta E = \varepsilon \int_0^{2\pi} dt\left(x^{(0)}f + v^{(0)}g\right) + O(\varepsilon^2) = \varepsilon \oint (-f dv + g dx) + O(\varepsilon^2), \tag{4.108}$$

where the integral is taken over the level curve, which is also iso-energy cycle, of the unperturbed ($\varepsilon = 0$) system in the (x, v) space. Obviously ΔE depends on x_0.

For the case of increasing energy, $\Delta E > 0$, we see an unwinding spiral in the (x, v) plane. For the case of decreasing energy, $\Delta E < 0$, the spiral contracts to a stationary point.

There are also systems where the sign of ΔE depends on x_0. Consider for example the van der Pol oscillator

$$\ddot{x} = -x + \varepsilon\dot{x}(1 - x^2). \tag{4.109}$$

As in Eq. (4.108), we integrate $\frac{d}{dt}E$ over a period, which in this case gives

$$\begin{aligned}\Delta E &= \varepsilon \int_0^{2\pi} \dot{x}^2(1 - x^2)dt + O(\varepsilon^2) = \varepsilon x_0^2 \int_0^{2\pi} \sin^2 t \left(1 - x_0^2 \cos^2 t\right) dt + O(\varepsilon^2) \\ &= \pi \left(x_0^2 - \frac{x_0^4}{4}\right) \varepsilon + O(\varepsilon^2).\end{aligned} \tag{4.110}$$

The $O(\varepsilon)$ part of this expression is zero when $x_0 = 2$, positive when $x_0 < 2$ and negative when $x_0 > 2$. Therefore, if we start with $x_0 < 2$ the system will be gaining energy, and the maximum value of $x(t)$ within a period will approach the value 2. On the contrary, if $x_0 > 2$ the system will be lose energy, and the maximum value of $x(t)$ over a period will decrease approaching the same value 2. This type of behavior, established for ΔE including only $O(\varepsilon)$ contributions (and thus ignoring all contributions of higher order in small ε) is characterized as the stable limit cycle, which can be characterized by

$$\Delta E(x_0) = 0 \quad \text{and} \quad \frac{d}{dx_0}\Delta E(x_0) < 0.$$

In summary, the van der Pol oscillator is an example of behavior where the perturbation is singular, meaning that is categorically different from the unperturbed case. Indeed, in the unperturbed case the particle oscillates cycling an orbit which depends on the initial condition, while in the perturbed case the particle ends up moving along the same limit cycle.

Exercise 4.8. Recall properties of stable / unstable limit cycles:

Limit Cycle is Stable at $x = x_0$ if	$\Delta E(x_0) = 0$	and	$\frac{d}{dx_0}\Delta E(x_0) < 0$
Limit Cycle is unstable at $x = x_0$ if	$\Delta E(x_0) = 0$	and	$\frac{d}{dx_0}\Delta E(x_0) > 0$

Suggest an example of perturbations, f and g, in Eq. (4.106) which leads to (a) an unstable limit cycle at $x_0 = 2$, and (b) one stable limit cycle at $x_0 = 2$ and one unstable limit cycle at $x_0 = 4$. Illustrate your suggested perturbations by building a computational snippet.

Consider another ODE example

$$\dot{I} = \varepsilon \left(a + b\cos(\theta/\omega)\right), \quad \dot{\theta} = \omega, \tag{4.111}$$

where $\omega, \varepsilon, a, b$ are constants, and ε-term in the first Eq. (4.111) is a perturbation. When ε is zero, I is an integral of motion, (meaning that it is constant along solutions of the ODE), and we think of θ as an angle in the phase space increasing linearly with the frequency ω. Note that the unperturbed system is equivalent to the one described by Eq. (4.106).

Example 4.6.2.

(a) Show that one can transform the unperturbed (i.e. $\varepsilon = 0$) version of the system described by Eq. (4.106) to the unperturbed version of the system described by Eq. (4.111) via the following transformation (change of variables)

$$v = \sqrt{I/2}\cos(\theta/\omega), \quad x = \sqrt{I/2}\sin(\theta/\omega). \tag{4.112}$$

(b) Restate Eq. (4.111) in the (x, v) variables.

The transformation discussed in the Example 4.6.2 is the so-called canonical transformation that preserves the Hamiltonian structure of the equations. In this case the Hamiltonian, which is generally a function of θ and I, depends only on I, $H = I\omega$, and one can indeed rewrite the unperturbed version of Eq. (4.111) as

$$\dot{\theta} = \frac{\partial H}{\partial I} = \omega, \quad \dot{I} = -\frac{\partial H}{\partial \theta} = 0, \tag{4.113}$$

therefore interpreting θ and I as the new coordinate and the new momentum respectively.

Averaging perturbed Eq. (4.111) over one $(2\pi\omega)$ angle revolution, as done in Section 4.6.3, one arrives at the following expression for the change in I over the 2π-period (of time)

$$\Delta I = 2\pi\varepsilon a. \tag{4.114}$$

Taking many, $2\pi n\omega$, revolutions, replacing $2\pi n$ by t, and ΔI by J, where the latter is the action averaged over time t, one arrives at the following equation

$$\dot{J} = \varepsilon a, \tag{4.115}$$

which has the solution, $J(t) = J_0 + \varepsilon a t$.

In fact Eqs. (4.111) can also be solved exactly

$$I(t) = \varepsilon a t + \varepsilon b \sin t, \tag{4.116}$$

and one can check the consistency, that is indeed solution of the averaged Eq. (4.115) does not deviate (with time) from the exact solution of Eq. (4.111)

$$\omega \neq 0: \quad |J(t) - I(t)| \leq O(1)\varepsilon. \tag{4.117}$$

In a general n-dimensional case one considers the following system of bare (unperturbed) differential equations

$$\dot{\boldsymbol{I}} = 0, \quad \dot{\boldsymbol{\theta}} = \boldsymbol{\omega}(\boldsymbol{I}), \quad \boldsymbol{I} := (I_1, \cdots, I_n), \quad \boldsymbol{\theta} := (\theta_1, \cdots, \theta_n), \tag{4.118}$$

where thus each component of $\boldsymbol{I}$ is an integral of motion of the unperturbed system of equations. Perturbed version of Eq. (4.118) becomes

$$\dot{\boldsymbol{I}} = \varepsilon g(\boldsymbol{I}, \boldsymbol{\theta}, \varepsilon), \quad \dot{\boldsymbol{\theta}} = \boldsymbol{\omega}(\boldsymbol{I}) + \varepsilon \boldsymbol{f}(\boldsymbol{I}, \boldsymbol{\theta}, \varepsilon), \tag{4.119}$$

where f and g are $2\pi\omega$-periodic functions of each of the components of $\boldsymbol{\theta}$. Since $\boldsymbol{I}$ changes slowly, due to smallness of ε, the perturbed system can be substituted by a much simpler averaged system for the slow (adiabatic) variables, $\boldsymbol{J}(t) = \boldsymbol{I}(t) + \boldsymbol{O}(\varepsilon)$:

$$\dot{\boldsymbol{J}} = \varepsilon \boldsymbol{G}(\boldsymbol{J}), \quad \boldsymbol{G}(\boldsymbol{J}) := \frac{\oint g(\boldsymbol{I}, \boldsymbol{\theta}, 0) d\boldsymbol{\theta}}{\oint d\boldsymbol{\theta}}, \tag{4.120}$$

where as in Section 4.6.3 $\oint$ stands for averaging over the period (one rotation) in the phase-space. Notice that the procedure of averaging over the periodic motion may brake at higher dimensions, $n > 1$, if the system has resonances, i.e. if $\sum_i N_i \omega_i = 0$, where N_i are integers.

If the perturbed system is Hamiltonian $\boldsymbol{\theta}$ plays the role of generalized coordinates and $\boldsymbol{I}$ of generalized momenta, then Eqs. (4.119) become

$$\dot{\boldsymbol{I}} = -\frac{\partial H}{\partial \boldsymbol{\theta}}, \quad \dot{\boldsymbol{\theta}} = \frac{\partial H}{\partial \boldsymbol{I}}. \tag{4.121}$$

In this case averaging over $\boldsymbol{\theta}$ the rhs of the first equation in Eq. (4.121) results in $\dot{\boldsymbol{J}} = 0$. This means that the slow variables, $J_1, \cdots, J_n$, also called adiabatic invariants, do not change with time. Notice that the main difficulty of applying this rather powerful approach consists in finding proper variables which remain integrals of motion of the unperturbed system.

Chapter 5

Partial Differential Equations.

A partial differential equation (PDE) is a differential equation that contains one or more unknown multivariate functions and their partial derivatives. We begin our discussion by introducing first-order ODEs, and how to resolve them to a system of ODEs by the method of characteristics. We then utilize ideas from the method of characteristics to classify (hyperbolic, elliptic and parabolic) linear, second-order PDEs in two dimensions (Section 5.2). We will discuss how to generalize and solve elliptic PDE, normally associated with static problems, in Section 5.3. Hyperbolic PDEs, discussed in section 5.4 are normally associated with waves. Here, we take a more general approach originating from intuition associated with waves as the phenomena (then wave solving a hyperbolic PDE is a particular example of a sound wave). We will discuss diffusion (also heat) equation as the main example of a generalized (to higher dimension) parabolic PDE in Section 5.5.

5.1 First-Order PDE: Method of Characteristics

The method of characteristics reduces PDE to multiple ODEs. The method applies mainly to first-order PDEs (meaning PDEs which contain only first-order derivatives) which are moreover linear in the first-order derivatives.

To motivate this technique we will first consider a function u of two independent variables (x, y), $u(x, y)$. Suppose that $u(x, y)$ solves

$$a(x, y, u)\frac{\partial u}{\partial x} + b(x, y, u)\frac{\partial u}{\partial y} = c(x, y, u). \tag{5.1}$$

Now consider some arbitrary differentiable parametric curve $(x(t), y(t))$ and consider the total derivative $\frac{du}{dt}$. By the chain rule we have

$$\frac{du}{dt} = \frac{dx}{dt}\frac{\partial u}{\partial x} + \frac{dy}{dt}\frac{\partial u}{\partial y}. \tag{5.2}$$

Observe, that since the parametric curve was arbitrary we may chose to define it by

$$\begin{aligned}\frac{dx}{dt} &= a(x,y,u),\\ \frac{dy}{dt} &= b(x,y,u),\\ \frac{du}{dt} &= c(x,y,u).\end{aligned} \tag{5.3}$$

Substituting this into (5.2) gives us precisely (5.1). Thus we have a family of characteristic curves from which we can construct our solution to (5.1). (It is a family of curves since we never gave an initial condition for the system (5.3).)

Let $u(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ be a function of a d-dimensional coordinate, $\boldsymbol{x} := (x_1, \ldots, x_d)$. Introduce the gradient vector, $\boldsymbol{\nabla}_{\boldsymbol{x}} u := (\partial_{x_i} u; i = 1, \ldots, d)$, and consider the following linear in $\boldsymbol{\nabla}_{\boldsymbol{x}} u$ equation

$$(\boldsymbol{V} \cdot \boldsymbol{\nabla}_{\boldsymbol{x}} u) := \sum_{i=1}^{d} V_i \partial_{x_i} u = f, \tag{5.4}$$

where the velocity, $\boldsymbol{V}(\boldsymbol{x}) \in \mathbb{R}^d$ and forcing, $f(\boldsymbol{x}) \in \mathbb{R}$ are given functions of $\boldsymbol{x}$.

First, consider the homogeneous version of Eq. (5.4)

$$(\boldsymbol{V} \cdot \boldsymbol{\nabla}_{\boldsymbol{x}} u) = 0. \tag{5.5}$$

Introduce an auxiliary parameter (or dimension) $t \in \boldsymbol{R}$, call it time, and then introduce the *characteristic equations*

$$\frac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{V}(\boldsymbol{x}(t)), \tag{5.6}$$

describing the evolution of the characteristic trajectory $\boldsymbol{x}(t)$ in time according to the function $\boldsymbol{V}$. A first integral is a function for which $\frac{d}{dt} F(\boldsymbol{x}(t)) = 0$. Observe that any first integral of Eqs. (5.6) is a solution to Eq. (5.5), and that any function of the first integrals of Eqs. (5.6), $g(F_1, \ldots, F_k)$, is also a solution to Eq. (5.5).

Indeed, a direct substitution of $u = g$ in Eq. (5.5) leads to the following sequence of equalities

$$(\boldsymbol{V} \cdot \boldsymbol{\nabla}_{\boldsymbol{x}} g) = \sum_{i=1}^{k} \frac{\partial g}{\partial F_i} \sum_{j=1}^{d} \frac{\partial F_i}{\partial x_j} V_j = \sum_{i=1}^{k} \frac{\partial g}{\partial F_i} \frac{d}{dt} F_i = 0. \tag{5.7}$$

The system of equations (5.6) has $d-1$ first integrals independent of t (directly). Then a general solution to Eq. (5.7) is

$$u(\boldsymbol{x}(t)) = g\left(F_1(\boldsymbol{x}(t)), \ldots, F_{d-1}(\boldsymbol{x}(t))\right), \tag{5.8}$$

where g is assumed to be sufficiently smooth (at least twice differential over the first integrals).

Eq. (5.5) has a nice geometrical/flow interpretation. If we think of $\boldsymbol{V}$, which is the d dimensional vector of the coefficients of $\nabla_x g$, as a "velocity", then Eq. (5.5) means that derivative of u over $\boldsymbol{x}$ projected to the vector $\boldsymbol{V}$ is equal to zero. Therefore, the solution to and ODE by the method of characteristics is reduced to reconstructing integral curves from vectors $\boldsymbol{V}(\boldsymbol{x})$, defined at every point $\boldsymbol{x}$ of the space, which are tangent to the curves. Then, the solution $u(\boldsymbol{x})$ is constant along the curves. If in the vicinity of each point $\boldsymbol{x}$ of the space, one changes variables, $\boldsymbol{x} \to (t, F_1, \ldots, F_{d-1})$, where t is considered as a parameter along an integral curve and, if the transformation is well defined (i.e. Jacobian of the transformation is not zero), then Eq. (5.5) becomes $du/dt = 0$ along the characteristic.

Let us illustrate how to find a characteristic on the example of the following homogeneous PDE

$$\partial_x u(x, y) + y\partial_y u(x, y) = 0.$$

The characteristic equations are $dx/dt = 1,\ dy/dt = y$, with the general solution $x(t) = t + c_1,\ y = c_2 \exp(t)$. The only first integral of the characteristic equation is $F(x, y) = y \exp(-x)$, therefore $u = g(F(x, y))$, where g is an arbitrary function is a general solution. It is useful to visualize the flow along the characteristics in the (x, y) space.

Example 5.1.1. Find the characteristics of the following PDEs and use them to find the the general solutions to the PDEs. Verify your solutions by direct substitution.

(a) $\partial_x u - y^2 \partial_y u = 0$,

(b) $x\partial_x u - y\partial_y u = 0$,

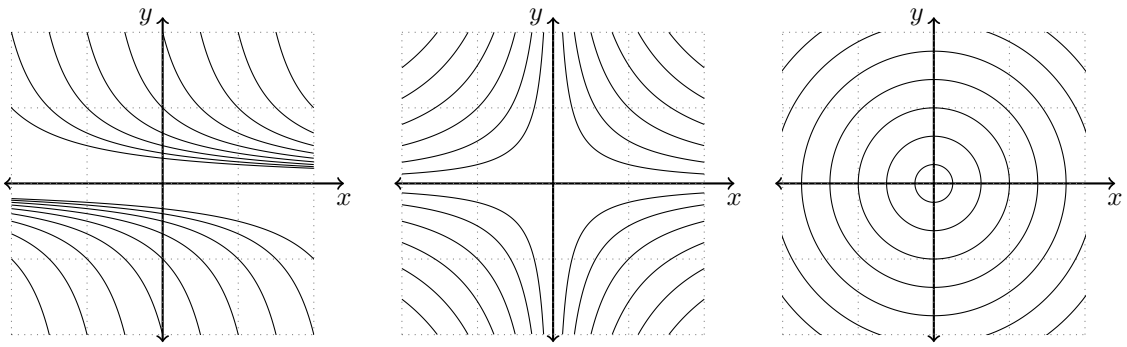
(c) $y\partial_x u - x\partial_y u = 0$.

Visualize the characteristics in (x, y)-plane.

Solution.

(a) The goal is to find curves parameterized by t expressing the left hand side as a total derivative giving $\frac{d}{dt}u(x(t), y(t)) = 0$. By the chain rule, this is equivalent to $\partial_x u\, \dot{x}(t) + \partial_y u\, \dot{y}(t) = 0$, which is equivalent to our PDE if we set $\dot{x}(t) = 1$ and $\dot{y}(t) = -y^2$. These are the characteristic equations. Their solutions are $x(t) = t + c_1$ and $y(t) = (t + c_2)^{-1}$. Eliminating t gives $c = y^{-1} - x =: F(x, y)$ as the only first integral. General solutions to the PDE are in the form $u(x, y) = g(y^{-1} - x)$ where $g : \mathbb{R} \to \mathbb{R}$ can be any function.

(b) The goal is to find curves parameterized by t expressing the left hand side as a total derivative giving $\frac{d}{dt}u(x(t), y(t)) = 0$. By the chain rule, this is equivalent to $\partial_x u\, \dot{x}(t) + \partial_y u\, \dot{y}(t) = 0$, which is equivalent to our PDE if we set $\dot{x}(t) = x$ and $\dot{y}(t) = -y$. These



(a) $1/y - x = \text{const}$ (b) $xy = \text{const}$ (c) $x^2 + y^2 = \text{const}$

Figure 5.1: Characteristic curves for PDEs in example 5.1.1

are the characteristic equations. Their solutions are $x(t) = c_1 e^t$ and $y(t) = c_2 e^{-t}$. Eliminating t gives $c = xy =: F(x, y)$ as the only first integral. General solutions to the PDE are in the form $u(x, y) = g(xy)$ where $g : \mathbb{R} \to \mathbb{R}$ can be any function.

(c) The goal is to find curves parameterized by t expressing the left hand side as a total derivative giving $\frac{d}{dt}u(x(t), y(t)) = 0$. By the chain rule, this is equivalent to $\partial_x u\, \dot{x}(t) + \partial_y u\, \dot{y}(t) = 0$, which is equivalent to our PDE if we set $\dot{x}(t) = y$ and $\dot{y}(t) = -x$. These are the characteristic equations. Their solutions are $x(t) = c\cos(t)$ and $y(t) = c\sin(t)$. Eliminating t gives $c = x^2 + y^2 =: F(x, y)$ as the only first integral. General solutions to the PDE are in the form $u(x, y) = g(x^2 + y^2)$ where $g : \mathbb{R} \to \mathbb{R}$ can be any function.

Consider the following initial value (boundary) Cauchy problem: solve Eq. (5.5) subject to the boundary condition

$$u(\boldsymbol{x})|_{\boldsymbol{x}_0 \in S} = \vartheta(\boldsymbol{x}_0), \tag{5.9}$$

where S is a surface (boundary) of the dimension $d - 1$. This Cauchy problem has a well-defined solution in at least some vicinity of S if S is not tangent to a characteristic of Eq. (5.5). Consistently with what was described above solution to Eq. (5.6) with the initial/boundary condition Eq. (5.9) can be thought as the change of variables.

Example 5.1.2. Let us illustrate the solution to the Cauchy problem on the example

$$\partial_x u = y\partial_y u, \qquad u(0, y) = \cos(y).$$

Solution. The characteristic equations, $\dot{x} = 1,\ \dot{y} = -y$, have solutions $x(t) = t - t_1,\ y(t) = \exp(t_2 - t)$, and one first integral

$$F(x, y) = y\exp(x) = \text{constant},$$

therefore

$$u(x, y) = g(y \exp(x)),$$

where g is an arbitrary function, is a general solution. Boundary/initial conditions are given at the straight line, $x = 0$, which is not tangent to any of the characteristic, $y = \exp(-x+x_1)$. Therefore, substituting the general solution in the boundary condition one finds a particular form of the function g for the specific Cauchy problem:

$$u(0, y) = g(y) = \cos(y).$$

This results in the desired solution: $u(x, y) = \cos(y \exp(x))$. □

Exercise 5.1. (a) Solve

$$y\partial_x u - x\partial_y u = 0,$$

for initial condition, $u(0, y) = y^2$. (b) Explain why the same problem with the initial condition $u(0, y) = y$ is ill-posed. (c) Determine whether the same problem with the initial condition $u(1, y) = y^2$ is ill-posed.

Example 5.1.3. Let $(\boldsymbol{q}, \boldsymbol{p}) = (q_1, \ldots, q_n, p_1, \ldots, p_n)$ be a set of canonical coordinates for a Hamiltonian system with Hamiltonian $H(\boldsymbol{q}, \boldsymbol{p})$, and let $f = f(t, \boldsymbol{q}, \boldsymbol{p})$ be any function of t (time), $\boldsymbol{q}$ and $\boldsymbol{p}$. Liouville's theorem states that

$$\partial_t f + \{f, H\} = 0, \quad \text{where } \{f, H\} := \sum_{i=1}^{n} \left(\frac{\partial f}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial f}{\partial p_i}\frac{\partial H}{\partial q_i} \right),$$

where $\{f, H\}$ is the so-called Poisson bracket of f and H. Find the characteristics of the Liouville's PDE.

Solution. We wish to find a $\boldsymbol{V}(t, \boldsymbol{q}, \boldsymbol{p})$ such that the left hand side of the PDE can be expressed in the form $\boldsymbol{V} \cdot \boldsymbol{\nabla} f$. This is satisfied for $\boldsymbol{V} = (1, \partial_{\boldsymbol{p}} H, -\partial_{\boldsymbol{q}} H)$. We interpret $\boldsymbol{V}$ as the vector field $\boldsymbol{V} = (\frac{dt}{ds}, \frac{dq}{ds}, \frac{dp}{ds})$. Introducing $\boldsymbol{V}$ we also define a family of curves, $(t(s), \boldsymbol{q}(s), \boldsymbol{p}(s))$, called the characteristic curves. A little algebra allows us to simplify the curves to

$$\frac{d}{dt} q_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q_i}, \quad \forall i = 1, \cdots, n.$$

Interpretation: In the next chapter, we will see that the Hamilton's equations, $d\boldsymbol{q}/dt = \partial_{\boldsymbol{p}} H$, and $d\boldsymbol{p}/dt = -\partial_{\boldsymbol{q}} H$, describe the evolution of the state of the system in phase space. Since the characteristic curves are precisely the solutions to the Hamilton's equations, which reduces Liouville's PDE to $\frac{df}{dt} = 0$, we infer that for a Hamiltonian system any function of the system's state variables $(\boldsymbol{q}, \boldsymbol{p})$ does not change as the system evolves in time. □

Now let us get back to the inhomogeneous Eq. (5.4). As is standard for linear equations, the general solution to an inhomogeneous equation is constructed as the superposition of the a particular solution and general solution to the respective homogeneous equation. To find the former we transition to characteristics, then Eq. (5.4) becomes

$$(\boldsymbol{V}\cdot\boldsymbol{\nabla}_{\boldsymbol{x}})\,u=(\dot{\boldsymbol{x}}\cdot\boldsymbol{\nabla}_{\boldsymbol{x}})\,u=\frac{d}{dt}u=f(\boldsymbol{x}(t)), \tag{5.10}$$

which can be integrated along the characteristic thus resulting in a desired particular solution to Eq. (5.4)

$$u_{inh}=\int\limits_{t_0}^{t}f(\boldsymbol{x}(s))ds\quad\text{where }\boldsymbol{x}(s)\text{ satisfies }(\boldsymbol{V}\cdot\boldsymbol{\nabla}_{\boldsymbol{x}})u=(\dot{\boldsymbol{x}}\cdot\boldsymbol{\nabla}_{\boldsymbol{x}})u. \tag{5.11}$$

Notice that this solution is not constant along characteristics.

Example 5.1.4. Solve the Cauchy problem for the following inhomogeneous equation

$$\partial_x u-y\partial_y u=y,\ u(0,y)=\sin(y).$$

The method of characteristics can also be generalized to quasi-liner first-order ODEs, (first-order ODEs (5.4) where $\boldsymbol{V}$ and f depend not only on the vector of coordinate, $\boldsymbol{x}$, but also on the function $u(\boldsymbol{x})$). In this case the characteristic equations become

$$\frac{d\boldsymbol{x}}{dt}=\boldsymbol{V}(\boldsymbol{x},u),\quad\frac{du}{dt}=f(\boldsymbol{x},u). \tag{5.12}$$

The general solution to a quasi-linear ODE is given by $g(F_1,F_2,\ldots,F_n)=0$, where g is an arbitrary function of n first integrals of Eq. (5.12).

Consider the example of the Hopf's equation in $d=1$

$$\partial_t u+u\partial_x u=0, \tag{5.13}$$

which, when $u(t;x)$ refers to the velocity of a particle at location x and time t, describes the one dimensional flow of non-interacting particles. The characteristic equations and initial conditions are

$$\dot{x}=u,\ \dot{u}=0,\ x(t=0)=x_0,\ u(t=0)=u_0(x_0).$$

Direct integration produces, $x=u_0(x_0)t+x_0$ giving the following implicit equation for u

$$u=u_0(x-ut). \tag{5.14}$$

Under the specific conditions, $u_0(x)=c(1-\tanh x)$, this results in the following (still implicit) equation, $u=c(1-\tanh(x-ut))$. Computing partial derivative, one derives

$\partial_x u = -c/(\cosh^2(x-ut)-ct)$, which shows that it diverges in finite time at $t_* = 1/c$ and $x = ut$. The phenomenon is called wave breaking, and has the physical interpretation of fast particles catching slower ones and aggregating, leading to sharpening of the velocity profile and eventual breakdown. This singularity is formal, meaning that the physical model is no longer applicable when the singularity occurs. Introducing a small $\kappa\partial_x^2 u$ term to the right hand side of Eq. (5.13) regularizes the non-physical breakdown, and explains creation of shock. The regularized second-order PDE is called Burgers' equation.

5.2 Classification of linear second-order PDEs:

Consider the most general linear second-order PDE over two independent variables:

$$a_{11}\partial_x^2 u + 2a_{12}\partial_x\partial_y u + a_{22}\partial_y^2 u + b_1\partial_x u + b_2\partial_y u + cu + f = 0, \tag{5.15}$$

where all the coefficients may depend on the two independent variables x and y.

The method of characteristics, (which applies to first-order PDEs, for example, when $a_{11} = a_{12} = a_{21} = c = 0$ in Eq. (5.15)), can inform the analysis of second-order PDEs. Therefore, let us momentarily return to the first-order PDE,

$$b_1\partial_x u + b_2\partial_y u + f = 0, \tag{5.16}$$

and interpret its solution as the variable transformation from the (x, y) pair of variables to the new pair of variables, $(\eta(x,y), \xi(x,y))$, assuming that the Jacobian of the transformation is neither zero no infinite anywhere within the domain of (x, y) of interest.

$$J = \det\begin{pmatrix} \partial_x\eta & \partial_y\eta \\ \partial_x\xi & \partial_y\xi \end{pmatrix} \neq 0, \infty. \tag{5.17}$$

Substituting $u = w(\eta(x,y), \xi(x,y))$ into the sum of the first derivative terms in Eq. (5.16) one derives

$$\begin{aligned} b_1\partial_x u + b_2\partial_y u &= b_1\left(\partial_x\eta\partial_\eta w + \partial_x\xi\partial_\xi w\right) + b_2\left(\partial_y\eta\partial_\eta w + \partial_y\xi\partial_\xi w\right) \\ &= \left(b_1\partial_x\eta + b_2\partial_y\eta\right)\partial_\eta w + \left(b_1\partial_x\xi + b_2\partial_y\xi\right)\partial_\xi w. \end{aligned} \tag{5.18}$$

Requiring that the second term in Eq. (5.18) is zero one observes that it is satisfied for all x, y if $\xi(y(x))$, i.e. it does not depend on x explicitly but only via $y(x)$ if the latter satisfies the characteristic equation, $b_1 dy/dx + b_2 = 0$.

Let us now try the same logic, but now focusing on the sum of the second-order terms in Eq. (5.15). We derive

$$a_{11}\partial_x^2 u + 2a_{12}\partial_x\partial_y u + a_{22}\partial_y^2 u = \left(A\partial_\xi^2 + 2B\partial_\xi\partial_\eta + C\partial_\eta^2\right) w, \tag{5.19}$$

where

$$A := a_{11}(\partial_x\xi)^2 + 2a_{12}(\partial_x\xi)(\partial_y\xi) + a_{22}(\partial_y\xi)^2$$
$$B := a_{11}(\partial_x\xi)(\partial_x\eta) + a_{12}(\partial_x\xi\partial_y\eta + \partial_y\xi\partial_x\eta) + a_{22}(\partial_y\xi)(\partial_y\eta)$$
$$C := a_{11}(\partial_x\eta)^2 + 2a_{12}(\partial_x\eta)(\partial_y\eta) + a_{22}(\partial_y\eta)^2.$$

Let us now attempt, by analogy with the case of the first-order PDE, to force first and last term on the rhs of Eq. (5.19) to zero, i.e. $A = C = 0$. This is achieved if we require that $\xi(y_+(x))$ and $\eta(y_-(x))$, where

$$\frac{dy_\pm}{dx} = \frac{a_{12} \pm \sqrt{D}}{a_{11}}, \quad \text{where} \quad D := a_{12}^2 - a_{11}a_{22}. \tag{5.20}$$

and D is called the *discriminant*. Eqs. (5.20) have in a general case distinct (first) integrals $\psi_\pm(x, y) = \text{const}$. Then, we can choose the new variables as $\xi = \psi_+(x, y)$ and $\eta = \psi_-(x, y)$

If $D > 0$ Eq. (5.15) is called a *hyperbolic* PDE. In this case, the characteristics are real, and any real pair (x, y) is mapped to the real pair (η, ξ). Eq. (5.15) gets the following canonical form

$$\partial_\xi\partial_\eta u + \tilde{b}_1\partial_\xi u + \tilde{b}_2\partial_\eta u + \tilde{c}u + \tilde{f} = 0. \tag{5.21}$$

Notice that another (second) canonical form for the hyperbolic equation is derived if we transition further from (ξ, η) to $(\alpha, \beta) := ((\eta + \xi)/2, (\xi - \eta)/2)$. Then Eq. (5.21) becomes

$$\partial_\alpha^2 u - \partial_\beta^2 u + \tilde{b}_1^{(2)}\partial_\alpha u + \tilde{b}_2^{(2)}\partial_\beta u + \tilde{c}^{(2)}u + \tilde{f}^{(2)} = 0. \tag{5.22}$$

If $D < 0$ Eq. (5.15) is called an *elliptic* PDE. In this case, Eqs. (5.21) are complex conjugate of each other and their first integrals are complex conjugate as well. To make the map from old to new variables real, we choose in this case, $\alpha = \text{Re}(\psi_+(x, y)) = (\psi_+(x, y) + \psi_-(x, y))/2$, $\beta = \text{Im}(\psi_+(x, y)) = (\psi_+(x, y) - \psi_-(x, y))/(2i)$. This change of variables results in the following canonical form for the elliptic second-order PDE:

$$\partial_\alpha^2 u + \partial_\beta^2 u + b_1^{(e)}\partial_\alpha u + b_2^{(e)}u_\beta + c^{(e)}u + f^{(e)} = 0. \tag{5.23}$$

$D = 0$ is the degenerate case, $\psi_+(x, y) = \psi_-(x, y)$, and the resulting equation is a *parabolic* PDE. Then we can choose $\beta = \psi_+(x, y)$ and $\alpha = \varphi(x, y)$, where φ is an arbitrary independent (of $\psi_+(x, y)$) function of x, y. In this case Eq. (5.15) gets the following canonical parabolic form

$$\partial_\alpha^2 u + b_1^{(p)}\partial_\alpha u + b_2^{(p)}\partial_\beta u + c^{(p)}u + f^{(p)} = 0. \tag{5.24}$$

Example 5.2.1. Define the type of equation and then perform change of variables reducing it to the respective canonical form

(a) $\partial_x^2 u + \partial_x \partial_y u - 2\partial_y^2 u - 3\partial_x u - 15\partial_y u + 27x = 0$,

(b) $\partial_x^2 u + 2\partial_x \partial_y u + 5\partial_y^2 u - 32u = 0$,

(c) $\partial_x^2 u - 2\partial_x \partial_y u + \partial_y^2 u + \partial_x u + \partial_y u - u = 0$.

Solution.

(a) The coefficients of the second order terms are $a_{11} = 1$, $a_{12} = 1/2$ and $a_{22} = -2$. The discriminant is $D = a_{12}^2 - a_{11}a_{22} = 9/4$. The equation is hyperbolic (everywhere) because the discriminant is positive (everywhere). There are two families of characteristics, defined by $dy/dx = (a_{12} \pm D)/a_{11}$ giving $dy/dx = 2$ and $dy/dx = -1$, respectively. The general solutions of the two equations are

$$y = 2x + \xi, \quad y = -x + \eta,$$

where ξ and η are arbitrary constants. Expressing, ξ and η via x and y one derives

$$\xi = y - 2x, \quad \eta = y + x.$$

Transitioning to the new variables one derives

$$\partial_\xi \partial_\eta u + \partial_\xi u + 2\partial_\eta u + 3(\eta - \xi) = 0.$$

(b) The coefficients of the second order terms are $a_{11} = 1$, $a_{12} = 1$ and $a_{22} = 5$. The discriminant is $D = a_{12}^2 - a_{11}a_{22} = -4$. The equation is elliptic (everywhere) because the discriminant is negative (everywhere). There are two families of the characteristics, defined by $dy/dx = (a_{12} \pm D)/a_{11}$ giving $dy/dx = 1 + 2i$ and $dy/dx = 1 - 2i$, respectively. The general solutions of the two equations are

$$y = (1 - 2i)x + \tilde{\xi}, \quad y = (1 - 2i)x + \tilde{\eta}.$$

Setting $\xi := (\tilde{\xi} + \tilde{\eta})/2 = y - x$ and $\eta = (\tilde{\xi} - \tilde{\eta})/2i = 2x$ as the new variables, we arrive a the following canonical form

$$\partial_\xi^2 u + \partial_\eta^2 u - 8u = 0.$$

(c) The coefficients of the second order terms are $a_{11} = 1$, $a_{12} = -1$ and $a_{22} = 1$. The discriminant is $D = a_{12}^2 - a_{11}a_{22} = 0$. The equation is parabolic (everywhere). We have one characteristic, $dy/dx = (a_{12} \pm D)/a_{11} = -1$, giving $y = -x + \xi$ therefore one of the new variables is the integral of the characteristic equations, $\xi = x + y$. We can

take any independent function of x and y as the second new variable. Let us pick, $\eta = x$. Then the equation becomes,

$$\partial_\eta^2 u + 2\partial_\xi u + \partial_\eta u - u = 0.$$

(Notice that the condition of functional independence consists in the requirement that Jacobian of the transformation is nonzero. Any other choice of second (independent) variable will result in another canonical form.) □

5.3 Elliptic PDEs: Method of Green Function

Elliptic PDEs often originate from the description of static phenomena in two or more dimensions.

Let us, first clarify the higher dimensional generalization aspect. We generalize Eq. (5.23) to

$$\sum_{i,j=1}^{d} a_{ij}\partial_{x_i}\partial_{x_j} u(\boldsymbol{x}) + \text{lower order terms} = 0, \tag{5.25}$$

where we assume that it is not possible to eliminate at least one second derivative term from the condition of the respective Cauchy problem. Notice that in $d > 2$ Eq. (5.25) cannot be reduced to a canonical form (introduced, in the previous Section, in $d = 2$).

Our primary focus will be on the generalization where $a_{ij} = \delta_{ij}$ in Eq. (5.25) in $d \geq 2$, and also on solving inhomogeneous equations, where a nontrivial (nonzero) solution is driven by a nonzero source. It is natural to find solutions to these equations using Green functions.

We have discussed in Section 4.4.1 how to solve static linear one dimensional case of the Poisson equation using the Green functions. Here we generalize and consider the Poisson equation in $d \geq 2$,

$$\boldsymbol{\nabla}_{\boldsymbol{r}}^2 u = \phi(\boldsymbol{r}), \tag{5.26}$$

where $\boldsymbol{\nabla}_{\boldsymbol{r}}^2 := \Delta_{\boldsymbol{r}}$ is the Laplace operator. In $d = 2$, $\boldsymbol{r} = (x, y) \in \mathbb{R}^2$ and $\Delta_{\boldsymbol{r}} = \partial_x^2 + \partial_y^2$. In $d = 3$, $\boldsymbol{r} = (x, y, z) \in \mathbb{R}^3$ and $\Delta_{\boldsymbol{r}} = \partial_x^2 + \partial_y^2 + \partial_z^2$. The Poisson Eq. (5.26) has many applications, for example, its solution $u(\boldsymbol{r})$ describes the electrostatic potential of the charge distributed in $\mathbb{R}^d$ with the density $\rho(\boldsymbol{r})$, for this example, $\phi(\boldsymbol{r}) = -4\pi\rho(\boldsymbol{r})$.

The Poisson's equation, defined in all of $\mathbb{R}^d$, can be solved by the method of Green functions. Recall, that the Green function is the solution to the inhomogenous equation with a point source on the right hand side,

$$\boldsymbol{\nabla}_{\boldsymbol{r}}^2 G = \delta(\boldsymbol{r}). \tag{5.27}$$

Then the solution to Eq. (5.26) becomes

$$u(\boldsymbol{r}) = \int d\boldsymbol{r}' G(\boldsymbol{r}-\boldsymbol{r}')\phi(\boldsymbol{r}'). \tag{5.28}$$

The solution to Eq. (5.27) can be found by applying the Fourier transform, resulting in the following algebraic equation, $k^2\hat{G}(\boldsymbol{k}) = -1$, where $k = |\boldsymbol{k}|$. Solving (trivially) for $\hat{G}$ and applying the Inverse Fourier transform, one derives for $d = 3$,

$$\begin{aligned} G(\boldsymbol{r}) &= -\int \frac{d^3\boldsymbol{k}}{(2\pi)^3}\frac{\exp(i(\boldsymbol{k}\cdot\boldsymbol{r}))}{k^2} = -\int \frac{d^2k_\perp}{(2\pi)^3}\int_{-\infty}^{\infty} dk_\parallel \frac{\exp(ik_\parallel r)}{k_\parallel^2+k_\perp^2} \\ &= -\int \frac{d^2k_\perp}{(2\pi)^3}\frac{\pi}{k_\perp}\exp(-k_\perp r) = -\frac{1}{4\pi r}, \end{aligned} \tag{5.29}$$

where for each $\boldsymbol{r}$, we change from the Cartesian to cylindrical representation associated with $\boldsymbol{r}$, i.e. $\boldsymbol{k} = (k_\parallel, \boldsymbol{k}_\perp)$, the one-dimensional $k_\parallel = (\boldsymbol{k}\cdot\boldsymbol{r})/r$ is along $\boldsymbol{r}$ and the two dimensional vector, $\boldsymbol{k}_\perp$, stands for the remaining two components of $\boldsymbol{k}$ orthogonal to $\boldsymbol{r}$. Substituting Eq. (5.29) into Eq. (5.28) one derives

$$u(\boldsymbol{r}) = -\int d^3\boldsymbol{r}' \frac{\phi(\boldsymbol{r})}{4\pi|\boldsymbol{r}-\boldsymbol{r}'|} = \int d^3\boldsymbol{r}' \frac{\rho(\boldsymbol{r})}{|\boldsymbol{r}-\boldsymbol{r}'|}, \tag{5.30}$$

which is thus expression for the electrostatic potential of a given distribution of the charge density in the space.

Note, that for $d = 2$, we usually write $\phi(\boldsymbol{r}) = -2\pi\rho(\boldsymbol{r})$. In this case the Green function is found to be $G(\boldsymbol{r}-\boldsymbol{r}') = \ln(|\boldsymbol{r}-\boldsymbol{r}'|)$.

The homogeneous case of $\phi = 0$ is often called the Laplace equation. We will distinguish the two cases calling them the (inhomogeneous) Laplace equation and the homogeneous Laplace equation respectively.

We will also discuss in the following the Debye equation

$$\left(\boldsymbol{\nabla}_{\boldsymbol{r}}^2 - \kappa^2\right) u = \phi(\boldsymbol{r}), \tag{5.31}$$

which describes distribution of charge $\rho(\boldsymbol{r})$ in plasma for $\phi(r) = -4\pi\rho(\boldsymbol{r})$.

Functions that satisfy the homogeneous Laplace equation are called harmonic functions. Notice that there exists no nonzero harmonic function defined on the whole of $\mathbb{R}^2$ that approach 0 as $|\boldsymbol{r}| \to \infty$. This can be seen by applying Fourier transform to the homogeneous Laplace equation. One derives $k^2\hat{u}(\boldsymbol{k}) = 0$, which results in $\hat{u}(\boldsymbol{k}) \sim \delta(\boldsymbol{k})$, and then (applying Inverse Fourier transform), $u(\boldsymbol{r}) = \text{const}$. Finally, requiring that $u \to 0$ at $r \to \infty$ one observes that the constant is zero. This argument extends to any dimension, and it also applies to the Debye equation; there exists no solution to the (homogeneous) Debye equation defined in the entire space that decays to zero at $r \to \infty$.

Consequently, nonzero harmonic functions must be defined in a bounded domain. For many physical applications, the homogeneous Laplace equation is supplemented with some form of boundary conditions. For example, u, or the normal component of its gradient to the boundary, $\boldsymbol{\nabla} u \cdot \boldsymbol{n}$, may be fixed at the boundary.

Example 5.3.1. Find the Green function for the Laplace equation in the region outside of the sphere of radius R and zero boundary condition on the sphere, i.e. solve

$$\boldsymbol{\nabla}_{\boldsymbol{r}}^2 G(\boldsymbol{r};\boldsymbol{r}') = \delta(\boldsymbol{r}-\boldsymbol{r}'), \tag{5.32}$$

for $\boldsymbol{r}$ such that $R \le r, r'$, with the boundary condition $G(\boldsymbol{r};\boldsymbol{r}') = 0$, for $R = r \le r'$.

Solution. The Green function can be constructed by recognizing that $|\boldsymbol{r}-\boldsymbol{r}''|^{-1}$ solves $\nabla_{\boldsymbol{r}}^2 G(\boldsymbol{r};\boldsymbol{r}') = 0$ for all $\boldsymbol{r} \neq \boldsymbol{r}''$. The trick is to find for each $\boldsymbol{r}' \in \mathcal{D}$ a fictitious image point $\boldsymbol{r}'' \notin \mathcal{D}$ such that $G(\boldsymbol{r};\boldsymbol{r}') = 0$ whenever $|\boldsymbol{r}| = R$. The problem distills down to finding the correct strength and position of the image point to enforce the boundary condition at every point on the boundary. Using symmetry, it is clear that $\boldsymbol{r}'$ and $\boldsymbol{r}''$ must be collinear. Therefore, we can write $\boldsymbol{r}'' = \alpha \boldsymbol{r}'$

$$\text{Find } A,\ \alpha \text{ such that} \quad G(\boldsymbol{r};\boldsymbol{r}') := -\frac{1}{4\pi|\boldsymbol{r}-\boldsymbol{r}'|} + \frac{A}{4\pi|\boldsymbol{r}-\alpha\boldsymbol{r}'|} = 0 \quad \text{whenever } |\boldsymbol{r}| = r = R,$$

For any $\boldsymbol{r}$ on the boundary and $\boldsymbol{r}' \in \mathcal{D}$, $|\boldsymbol{r}-\boldsymbol{r}'| = \sqrt{R^2 + |\boldsymbol{r}'|^2 - 2|\boldsymbol{r}'|R\cos(\theta)}$. Similarly, $|\boldsymbol{r}-\boldsymbol{r}''| = \sqrt{R^2 + \alpha^2|\boldsymbol{r}'|^2 - 2\alpha|\boldsymbol{r}'|R\cos(\theta)}$ (See blue and orange triangles respectively in the Fig. 5.2). To enforce the boundary condition, their contributions must cancel. That is,

$$-\frac{1}{4\pi\sqrt{|\boldsymbol{r}'|^2 + R^2 - 2|\boldsymbol{r}'|R\cos(\theta)}} + \frac{A}{4\pi\sqrt{\alpha^2|\boldsymbol{r}'|^2 + R^2 - 2\alpha|\boldsymbol{r}'|R\cos(\theta)}} = 0.$$

We are looking for values of A and α that are independent of θ. The algebra is a bit tedious, but we find that $A = R/|\boldsymbol{r}'|$ and $\alpha = (R/|\boldsymbol{r}'|)^2$. Hence, the Green function is

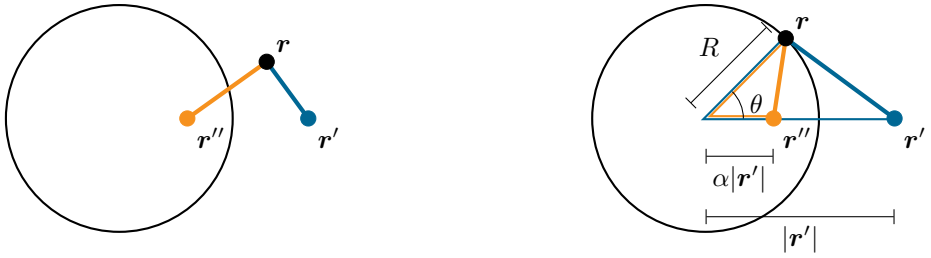
$$G(\boldsymbol{r};\boldsymbol{r}') := -\frac{1}{4\pi|\boldsymbol{r}-\boldsymbol{r}'|} + \frac{R/|\boldsymbol{r}'|}{4\pi|\boldsymbol{r}-\boldsymbol{r}''|},$$

where $\boldsymbol{r}'' = (R/|\boldsymbol{r}'|)^2\boldsymbol{r}'$. □

Exercise 5.2. Find general solutions to the inhomogeneous Debye equation

$$\left(\boldsymbol{\nabla}_{\boldsymbol{r}}^2 - \kappa^2\right) f = -4\pi\rho(\boldsymbol{r}),$$

where the charge density, $\rho(\boldsymbol{r})$ depends only on the distance from the origin (zero), i.e. $\rho(r = |\boldsymbol{r}|)$. (Hint: Consider finding the Green function first, acting by analogy with how we found the Green function of the Laplace equation above.)



(a) For each $\boldsymbol{r}' \in \mathcal{D}$, identify the associated image $\boldsymbol{r}'' \notin \mathcal{D}$ to find response at any $\boldsymbol{r}$

(b) Contributions from $\boldsymbol{r}'$ and $\boldsymbol{r}''$ must cancel to enforce the boundary condition

Figure 5.2: Method of images applied to the exterior of a sphere in the example 5.3.1.

5.4 Waves in a Homogeneous Media: Hyperbolic PDE *

Although hyperbolic PDEs are normally associated with waves *, we begin our discussion by developing intuition which generalizes to a broader class of an integro-differential equations beyond hyperbolic PDEs. In other words, we act here in reverse to what may be considered the standard mathematical process; we begin by describing properties of solutions associated with waves, and then walk back to the equations which are describing such waves.

Consider the propagation of waves in homogeneous media, for example: electro-magnetic waves, sound waves, spin-waves, surface-waves, electro-mechanical waves (in power systems), and so on. In spite of such a variety of phenomena, they all admit one rather universal description. The wave process at a general position in d-dimensional space $\boldsymbol{r}$ and time t is represented as the following integral over the wave vector $\boldsymbol{k}$

$$u(t;\boldsymbol{r}) = \int \frac{d\boldsymbol{k}}{(2\pi)^k} \exp\left(i(\boldsymbol{k}\cdot\boldsymbol{r})\right) \psi_{\boldsymbol{k}}(t)\hat{u}(\boldsymbol{k}), \quad \psi_{\boldsymbol{k}}(t) \equiv \exp\left(-i\omega(\boldsymbol{k})t\right), \tag{5.33}$$

where $\omega(\boldsymbol{k})$ and $\hat{u}(\boldsymbol{k})$ are the dispersion law and wave amplitude dependent on the wave vector $\boldsymbol{k}$. (Notice the similarities and the differences with the Fourier integral.) In Eq. (5.33) $\psi_{\boldsymbol{k}}(t)$ is a solution to the following first-order (in time) linear ODE

$$\left(\frac{d}{dt} + i\omega(\boldsymbol{k})\right) \psi_{\boldsymbol{k}} = 0, \tag{5.34}$$

or alternatively of the following second-order linear ODE

$$\left(\frac{d^2}{dt^2} + (\omega(\boldsymbol{k}))^2\right) \psi_{\boldsymbol{k}} = 0. \tag{5.35}$$

These are called the wave equations in Fourier representation. The linearity of the equations is principal and is due to the fact that generally nonlinear dynamics is linearized. Waves

*This is an Auxiliary Section which can be dropped at the first reading. Material from the Section will not contribute midterm and final exams.

may also interact with each other. The interaction of waves can only come from accounting for nonlinearities in the original equations. In this analysis, we focus primarily on the linear regime.

Dispersion Laws

Consider the case where $\omega_k = c|\boldsymbol{k}|$, where c is a constant having dimensionality and sense of velocity. In this case, the inverse Fourier transform version of Eq. (5.35) becomes

$$\left(\frac{d^2}{dt^2} - c^2\boldsymbol{\nabla}_{\boldsymbol{r}}^2\right)\psi(t;\boldsymbol{r}) = 0. \tag{5.36}$$

Note that the two differential operators in Eq. (5.36), one in time and another in space, have opposite signs. Therefore, we naturally arrive at the case which generalizes the hyperbolic PDE (5.22). It is a generalization because $\boldsymbol{r}$ is not one-dimensional but d-dimensional, $d \geq 1$.

Eq. (5.26) with c constant, explains a variety of important physical situations: as mentioned already, it describes propagation of sound in a homogeneous gas, liquid or crystal media. In this case ψ describes the shift of an element of the matter from its equilibrium position and c is the speed of sound in the material [b].

Another example is given by the electro-magnetic waves, described by the Maxwell equations on the electric, $\boldsymbol{E}$, and magnetic, $\boldsymbol{B}$, fields,

$$\partial_t\boldsymbol{E} = c\boldsymbol{\nabla}_{\boldsymbol{r}} \times \boldsymbol{B}, \quad \partial_t\boldsymbol{B} = -c\boldsymbol{\nabla}_{\boldsymbol{r}} \times \boldsymbol{E}, \tag{5.37}$$

supplemented by the divergence-free conditions,

$$(\boldsymbol{\nabla}_{\boldsymbol{r}} \cdot E) = (\boldsymbol{\nabla}_{\boldsymbol{r}} \cdot B) = 0, \tag{5.38}$$

where $\times$ is the vector product in $d = 3$[c], and c is the speed of light in the media. Differentiating the first equation in the pair of Eqs. (5.37) over time, substituting the resulting $\partial_t\boldsymbol{\nabla}_{\boldsymbol{r}} \times \boldsymbol{B}$ by $-c(\boldsymbol{\nabla}_{\boldsymbol{r}} \times (\boldsymbol{\nabla}_{\boldsymbol{r}} \times \boldsymbol{E}))$, consistently with the second equation in the pair, and taking into account that for the divergence-free, $\boldsymbol{E}$, $(\boldsymbol{\nabla}_{\boldsymbol{r}} \times (\boldsymbol{\nabla}_{\boldsymbol{r}} \times \boldsymbol{E})) = \boldsymbol{\nabla}_{\boldsymbol{r}}^2\boldsymbol{E}$, one arrives at Eq. (5.36) for all components of the electric field, i.e. with ψ replaced by $\boldsymbol{E}$.

[b] Note, that there is a unique speed of sound in gas or liquid, while $3d$ crystal supports three different waves (with different three different c) each associated with a distinct polarization. For example, in an isotropic crystals there are longitudinal and transversal waves propagating along and, respectively, perpendicular to the media shift.

[c] $(\boldsymbol{\nabla}_{\boldsymbol{r}} \times \boldsymbol{B})_i = \varepsilon_{ijk}\nabla_j B_k$, where $i, j, = 1, \cdot 3$ and ε_{ijk} is the absolutely skew-symmetric tensor in $d = 3$

The dispersion law in the case of sound and light waves is linear, $\omega(\boldsymbol{k}) = \pm c|\boldsymbol{k}|$, however there are other more complex examples. For example, surface waves propagating over the surface of water (with air), are characterized by the following dispersion law

$$\omega(\boldsymbol{k}) = \sqrt{gk + (\sigma/\rho)k^3}, \tag{5.39}$$

where g, σ and ρ are gravity coefficient, surface tension coefficient and density of the fluid, respectively. Eq. (5.39) is so complex because it accounts for both capillary and gravitational effects. Gravitational waves dominate at small q (large distances), where Eq. (5.39) transforms to $\omega(\boldsymbol{q}) = \sqrt{gq}$, while the capillary waves dominate in the opposite limit of large q (small distances), where one gets asymptotically $\omega = (\sigma/\rho)^{1/2}q^{3/2}$.

Recall that Eq. (5.34) or Eq. (5.35) are stated in the Fourier k-representation. Transitioning to the respective r-representation in the case of a nonlinear dispersion relation, for example associated with Eq. (5.39), will NOT result in a PDE. We arrive in this general case at an integro-differential equation, reflecting the fact that the nonlinear dispersion relation, even though local in the k-space becomes nonlocal in r-space.

In general, propagation of waves in the homogeneous media is characterized by the dispersion law dependent only of the absolute value, $k = |\boldsymbol{k}|$ of the wave vector, $\boldsymbol{k}$. $\omega(k)/k$ and $d\omega(k)/dk$, both having dimensionality of speed, are called, respectively, phase velocity and group velocity.

Example 5.4.1. Solve the Cauchy (initial value) problem for amplitude of spin-waves which satisfy the following PDE

$$\partial_t^2\psi = -(\Omega - b\boldsymbol{\nabla}_{\boldsymbol{r}}^2)^2\psi, \tag{5.40}$$

in $d = 3$, where $\psi(t = 0; \boldsymbol{r}) = \exp(-r^2)$ and $d\psi/dt(t = 0; \boldsymbol{r}) = 0$.

Solution. Note, first, that applying the Fourier transform over $\boldsymbol{r}$ to Eq. (5.40) one arrives at Eq. (5.34), where

$$\omega(k) = \Omega + bk^2, \tag{5.41}$$

is the respective (spin wave) dispersion law. The Fourier transform of the initial condition over $\boldsymbol{k}$ is, $\hat{\psi}(t = 0; \boldsymbol{k}) = \pi^{3/2}\exp(-k^2/4)$. Since $d\psi/dt(t = 0; \boldsymbol{r}) = 0$, the Fourier transform of the initial condition is zero as well, that is, $d\hat{\psi}/dt(t = 0; \boldsymbol{k}) = 0$. Then, the solution to Eqs. (5.34,5.41) becomes $\hat{\psi}(t; \boldsymbol{k}) = \pi^{3/2}\exp(-k^2/4)\cos((\Omega + bk^2)t)$. Evaluating the inverse

Fourier transform one derives

$$\begin{aligned}
\psi(t;\boldsymbol{r}) &= \pi^{3/2}\int \frac{d^3k}{(2\pi)^3}e^{-q^2/4}\cos((\Omega+bk^2)t)\exp(i(\boldsymbol{k}\cdot\boldsymbol{r}))\\
&= \int_0^\infty \frac{kdk}{2\pi^{1/2}r}e^{-k^2/4}\cos((\Omega+bk^2)t)\sin(kr)\\
&= -\int_0^\infty \frac{dk}{2\pi^{1/2}r}e^{-k^2/4}\cos((\Omega+bk^2)t)\frac{d}{dr}\cos(kr)\\
&= -\mathrm{Re}\left(\frac{\exp(i\Omega t)}{4\pi^{1/2}r}\frac{d}{dr}\int_{-\infty}^\infty dk\exp\left(-\frac{1-4ibt}{4}k^2+ikr\right)\right)\\
&= \mathrm{Re}\left(\frac{\exp\left(i\Omega t-\frac{r^2}{1-4ibt}\right)}{(1-4ibt)^{3/2}}\right).
\end{aligned}$$

Example 5.4.2. Solve the Cauchy (initial value) problem for the wave Eq. (5.36) in $d=3$, where $\psi(t=0;\boldsymbol{r})=\exp(-r^2)$ and $d\psi/dt(t=0;\boldsymbol{r})=0$.

Stimulated Waves: Radiation

So far we have discussed the free propagation of waves. Consider the inhomogeneous equation generalizing Eq. (5.35) that arises from a source term $\chi(t;\boldsymbol{r})$ on the right hand side:

$$\left(\frac{d^2}{dt^2}+(\omega(-i\boldsymbol{\nabla}_{\boldsymbol{r}}))^2\right)\psi(t;\boldsymbol{r})=\chi(t;\boldsymbol{r}). \tag{5.42}$$

where we have used $-i\boldsymbol{\nabla}_{\boldsymbol{r}}\exp(i\boldsymbol{kr})=k\exp(i\boldsymbol{kr})$. You may assume that the dispersion law, $\omega(k)$ is continuous value of its argument (absolute value of the wave vector) so that the operator $\omega(-i\boldsymbol{\nabla}_{\boldsymbol{r}}))^2$ is well defined in the sense of the function's Taylor series.

The Green function for the PDE is defined as the solution to

$$\left(\frac{d^2}{dt^2}+(\omega(-i\boldsymbol{\nabla}_{\boldsymbol{r}}))^2\right)G(t;\boldsymbol{r})=\delta(t)\delta(\boldsymbol{r}). \tag{5.43}$$

The solution to the inhomogeneous PDE, Eq. (5.42), can be expressed as the convolution of the source term $\chi(t_1;\boldsymbol{r}_1)$ with the Green function, $G(t;\boldsymbol{r})$

$$\psi(t;\boldsymbol{r})=\int dt_1 d\boldsymbol{r}_1 G(t-t_1;\boldsymbol{r}-\boldsymbol{r}_1)\chi(t_1;\boldsymbol{r}_1), \tag{5.44}$$

The solution to Eq. (5.42) is expressed as sum of the forced solution (5.44) and a zero mode of the respective free equation, i.e. Eq. (5.42) with zero right hand side.

To solve Eq. (5.43) for the Green function, or equivalently equation for its Fourier transform

$$\left(\frac{d^2}{dt^2} + (\omega(\boldsymbol{k}))^2\right)\hat{G}(t;\boldsymbol{k}) = \delta(t). \tag{5.45}$$

Recall that the inhomogeneous ODE. (5.45) was already discussed earlier in the course. Indeed Eq. (4.36) solves Eq. (5.45). Then recalling that ω depends on $\boldsymbol{k}$ and applying the inverse Fourier transform over $\boldsymbol{k}$ to Eq. (4.36) one arrives at

$$G(t;\boldsymbol{r}) = \theta(t)\int \frac{d^3k}{(2\pi)^3}\frac{\sin(\omega(k)t)}{\omega(k)}\exp\left(i(\boldsymbol{k}\cdot\boldsymbol{r})\right). \tag{5.46}$$

Example 5.4.3. Show that the general expression (5.46) in the case of the linear dispersion law, $\omega(\boldsymbol{k}) = ck$, becomes

$$G(t;\boldsymbol{r}) = \frac{\theta(t)}{4\pi cr}(\delta(r-ct) - \delta(r+ct)), \tag{5.47}$$

where $r = |\boldsymbol{r}|$.

Solution. The linear dispersion law means that $w(\boldsymbol{k}) = ck$, where $k = |\boldsymbol{k}|$. Then

$$G(t;\boldsymbol{r}) = \theta(t)\int_{\mathbb{R}^3}\frac{\sin ckt}{ck}\exp(i\boldsymbol{r}\cdot\boldsymbol{k})\frac{d^3\boldsymbol{k}}{(2\pi)^3}.$$

Compute the integral by rotating the coordinate system so that $\boldsymbol{r}$ is pointed in the z-direction (i.e. $\boldsymbol{r} = (0,0,r)^\top$) and then switch to spherical coordinates (i.e. $(k_1,k_2,k_3) \mapsto (k\cos\theta\sin\phi, k\sin\theta\sin\phi, k\cos\phi)$). The scalar product $\boldsymbol{r}\cdot\boldsymbol{k}$ then evaluates to $0+0+rk\cos\phi$.

$$\begin{aligned}
G(t;\boldsymbol{r}) &= \frac{\theta(t)}{(2\pi)^3}\int_0^{2\pi}\int_0^{\pi}\int_0^{\infty}\frac{\sin ckt}{ck}\exp(irk\cos\phi)k^2\sin\phi\,dk\,d\phi\,d\theta \\
&= \frac{\theta(t)}{(2\pi)^2}\int_0^{\infty}k^2\frac{e^{ickt}-e^{-ickt}}{2ick}\cdot\frac{e^{ikr}-e^{-ikr}}{ikr}dk \quad \text{via the sub. } u = -\cos\theta \\
&= \frac{\theta(t)}{(2\pi)^2cr}\int_0^{\infty}\frac{e^{ik(r-ct)}+e^{-ik(r-ct)}}{2} - \frac{e^{ik(r+ct)}+e^{-ik(r+ct)}}{2}dk \\
&= \frac{\theta(t)}{4\pi cr}(\delta(r-ct)-\delta(r+ct)),
\end{aligned}$$

which is equivalent to the expression we were given. □

Substituting Eq. (5.47) into Eq. (5.44) one derives the following expression for linear dispersion (light or sound) radiation from a source

$$\psi(t;\boldsymbol{r}) = \frac{1}{4\pi c^2}\int\frac{d\boldsymbol{r}_1}{R}\chi\left(t-\frac{R}{c};\boldsymbol{r}_1\right). \tag{5.48}$$

The solution suggests that action of the source is delayed by R/c correspondent to propagation of light (or sound) from the source to the observation point.

Example 5.4.4. Solve the radiation Eq. (5.42) in the case of the linear dispersion law for the case of a point harmonic source, $\chi(t;\boldsymbol{r}) = \cos(\omega t)\delta(\boldsymbol{r})$.

Solution.

$$\psi(t;\boldsymbol{r}) = \int dt_1 d\boldsymbol{r}_1 G(t-t_1;\boldsymbol{r}-\boldsymbol{r}_1)\chi(t_1;\boldsymbol{r}_1) = \frac{1}{4\pi c^2}\int \frac{d\boldsymbol{r}_1}{R}\cos(\omega(t-R/c))\delta(\boldsymbol{r}_1)$$
$$= \frac{1}{4\pi R c^2}\cos(\omega(t-R/c)). \quad \square$$

5.5 Diffusion Equation

The most common example of a multi-dimensional generalization of the parabolic equation Eq. (5.24) is the homogeneous diffusion equation

$$\partial_t u = \kappa \nabla^2_{\boldsymbol{r}} u, \tag{5.49}$$

where κ is the diffusion coefficient. The equation appears in a number of applications, for example, this equation can be used to describe the evolution of the density of number of particles, or the spatial variation of temperature. The same equation describes properties of the basic stochastic process (Brownian motion).

Consider the Cauchy problem with $u(t;\boldsymbol{r})$ given at $t = 0$. The Fourier transform over $\boldsymbol{r} \in \mathbb{R}^d$ is

$$\hat{u}(t;\boldsymbol{k}) = \int dy_1 \dots dy_d \exp(i\boldsymbol{k}\cdot\boldsymbol{x})\, u(t;\boldsymbol{y}). \tag{5.50}$$

Integrating Eq. (5.49) with the Fourier weight one arrives at

$$\partial_t \hat{u}(t;\boldsymbol{k}) = -k^2 \hat{u}(t;\boldsymbol{k}) \tag{5.51}$$

Integrating the equation over time, $\hat{u}(t;\boldsymbol{q}) = \exp(-q^2 t)\hat{u}(0;\boldsymbol{k})$, and evaluating the inverse Fourier transform over $\boldsymbol{q}$ of the result one arrives at

$$u(t;\boldsymbol{x}) = \int \frac{dy_1, \dots dy_d}{(4\pi t)^{d/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{y})^2}{4t}\right) u(0;\boldsymbol{y}). \tag{5.52}$$

If the initial field, $u(0;x)$, is localized around some x, say around $x = 0$, that is if $u(0;x)$ decays with $|x|$ increase sufficiently fast, then one may find a universal asymptotic of $u(t;x)$ at long times, $t \gg l^2$, where l is the length scale on which $u(0;x)$ is localized. At these sufficiently large times dominant contribution to the integral in Eq. (5.52) is acquired from the $|y| \sim l$ vicinity of the origin, and therefore in the leading order one can ignore y-dependence of the diffusive kernel in the integrand of Eq. (5.52), i.e.

$$u(t;\boldsymbol{x}) \approx \frac{A}{(4\pi t)^{d/2}} \exp\left(-\frac{x^2}{4t}\right), \quad A = \int u(0;\boldsymbol{y}) dy_1 \dots dy_d. \tag{5.53}$$

Notice that the approximation (5.53) corresponds to the substitution of $u(0,\boldsymbol{y}) \to A\delta(\boldsymbol{y})$ in Eq. (5.52). Another interpretation of Eq. (5.53) corresponds to expanding, $\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{y})^2}{4t}\right)$, in the Taylor series in y, and then ignoring all but the leading order term, $O(y^0)$, in the expansion. If $A = 0$ one needs to account for the $O(y^1)$ term, and drop the rest. In this case the analog of Eq. (5.53) becomes

$$u(t;\boldsymbol{x}) \approx \frac{(\boldsymbol{B}\cdot\boldsymbol{x})}{(4\pi t)^{d/2+1}} \exp\left(-\frac{x^2}{4t}\right), \quad B = 2\pi \int \boldsymbol{y} u(0;\boldsymbol{y}) dy_1 \dots dy_d. \tag{5.54}$$

Exercise 5.3. Find asymptotic behavior of a one-dimensional diffusion equation at sufficiently long times for the following initial conditions

(a) $u(0;x) = x \exp\left(-\dfrac{x^2}{2l^2}\right)$

(b) $u(0;x) = \exp\left(-\dfrac{|x|}{l}\right)$

(c) $u(0;x) = x \exp\left(-\dfrac{|x|}{l}\right)$

(d) $u(0;x) = \dfrac{1}{x^2+l^2}$

(e) $u(0;x) = \dfrac{x}{(x^2+l^2)^2}$

Hint: Think about expanding the diffusion kernel in the integrand of Eq.(5.52) in a series over y.

Our next step is to find the Green function of the heat equation, i.e. to solve

$$\partial_t G - \kappa \nabla^2_{\boldsymbol{r}} G = \delta(t)\delta(\boldsymbol{x}), \tag{5.55}$$

In fact, we have solved this problem already as Eq. (5.52) describes it with $u(0;\boldsymbol{y}) = G(+0;\boldsymbol{x}) = \delta(\boldsymbol{x})$ set as the initial condition. The result is

$$G(t;\boldsymbol{x}) = \frac{1}{(4\pi t)^{d/2}} \exp\left(-\frac{x^2}{4t}\right). \tag{5.56}$$

As always, the Green function can be used to solve the inhomogeneous diffusion equation

$$\partial_t u - \kappa \nabla^2_{\boldsymbol{x}} u = \phi(t;\boldsymbol{x}) \tag{5.57}$$

which solution is expressed via the Green function as follows

$$u(t;\boldsymbol{x}) = \int_{-\infty}^{t} dt' \int d\boldsymbol{y} G(t';\boldsymbol{y}) \phi(t-t';\boldsymbol{x}-\boldsymbol{y}), \tag{5.58}$$

where we assume that $u(\infty;\boldsymbol{x}) = 0$.

Example 5.5.1. Solve Eq. (5.57) for $\phi(t;\boldsymbol{x}) = \theta(t)\exp\left(-x^2/(2l^2)\right)$ in the $d=4$-dimensional space.

5.6 Boundary Value Problems: Fourier Method

Consider the boundary value problem associated with sound waves:

$$\partial_t^2 u(t;x) - c^2\partial_x^2 u(t;x) = 0, \tag{5.59}$$

$$0 \le x \le L, \quad u(t,0) = u(t,L) = 0, \quad u(0,x) = \varphi(x), \quad \partial_t u(0,x) = \psi(x). \tag{5.60}$$

This problem can be solved by the Fourier Method (also called the method of variable separation), which is split in two steps.

First, we look for a particular solution which satisfy only boundary conditions over one of the coordinates, x. We look for $u(t,x)$ in the separable form $u(t,x) = X(x)T(t)$. Substituting this ansatz in Eq. (5.59) one arrives at

$$\frac{X''(x)}{X(x)} = \frac{T''(t)}{T(t)} = -\lambda, \tag{5.61}$$

where λ is an arbitrary constant. General solution to the equation for X is

$$X = A\cos(\sqrt{\lambda}x) + B\sin(\sqrt{\lambda}x).$$

Require that $X(x)$ satisfies the same boundary conditions as in Eq. (5.60). This is possible only if $A = 0$ and $L\sqrt{\lambda} = n\pi, \quad n = 1, 2, \ldots$. From here we derive solution labeled by integer n and respective spatial form of the solution

$$\lambda_n = \left(\frac{n\pi}{L}\right)^2, \quad X_n(x) = \sin\left(\frac{n\pi x}{L}\right).$$

We are now ready to get back to Eq. (5.61) and resolve equation for $T(t)$:

$$T_n(t) = A_n\cos\left(\frac{n\pi ct}{L}\right) + B_n\sin\left(\frac{n\pi ct}{L}\right),$$

where A_n, B_n are arbitrary constants. $X_n(x)$ form a complete basis and therefore a general solution can be written as a linear combination of the basis solutions:

$$u(t,x) = \sum_{n=1}^{\infty} X_n(x)T_n(t).$$

On the second step we fix A_n and B_n resolving the initial portion of the conditions (5.60):

$$\varphi(x) = \sum_{n=1}^{\infty} A_n X_n(x), \quad \psi(x) = \sum_{n=1}^{\infty} \lambda_n B_n X_n(x). \tag{5.62}$$

Notice that the eigen-functions, $X_n(x)$, are ortho-normal

$$\int_0^L dx X_n(x) X_m(x) = \frac{L}{2}\delta_{nm}.$$

Multiplying both Eqs. (5.62) on $X_m(x)$, integrating them from 0 to L, and accounting for the ortho-normality of the eigen-functions, one derives

$$A_m = \frac{2}{L}\int_0^L dx\varphi(x)X_m(x), \quad B_m = \frac{2}{\lambda_m L}\int_0^L dx\psi(x)X_m(x). \tag{5.63}$$

Example 5.6.1. The equation describing deviation of a string from the straight line, $u(t;x)$, is $\partial_t^2 u - c^2\partial_x^2 u = 0$, where x is the position along the line, t, is the time, and, c, is a constant (speed of sound). Assume that the string has at $t = 0$ a parabolic shape, $u(0;x) = 4hx(L-x)/L^2$, with both ends, at $x = 0$ and $x = L$, respectively, attached to the straight line. Let us also assume that the speed of the string is equal to zero at $t = 0$, i.e. $\forall x \in [0, L]$, $\partial_t u(0;x) = 0$. Find dependence of the string deviation, $u(t;x)$, on time, t, at a position, $x \in [0, L]$, along the straight line.

Let us now analyze the following parabolic boundary value problem over $x \in [0, L]$:

$$\partial_t u = a^2\partial_x^2 u, \quad u(t,0) = u(t,L) = 0, \quad u(0,x) = \begin{cases} x, & x < L/2 \\ L-x, & x > L/2. \end{cases} \tag{5.64}$$

Here we follow the same Fourier method approach. In fact the spectral part of the solution here is identical to the one just described above in the hyperbolic case, while the temporal components are obviously different. One derives, $T_n' = -\lambda_n T_n$, which has a decaying solution

$$T_n = A_n \exp\left(-\left(\frac{n\pi}{L}\right)^2 a^2 t\right).$$

Expansion of the initial conditions in the Fourier series is equivalent to conducted above, therefore resulting in

$$u(t,x) = \frac{4L}{\pi^2}\sum_{n=0}^{\infty}\frac{(-1)^n}{(2n+1)^2}\exp\left(-\left(\frac{(2n+1)\pi}{L}\right)^2 a^2 t\right)\sin\left(\frac{2n+1}{L}\pi x\right).$$

Notice that the solution is symmetric with respect to the middle of the interval, $u(t,x) = u(t,L-x)$, as this symmetry is inherited from the initial conditions.

Exercise 5.4. Solve the following boundary value problem

$$\partial_t u = a^2\partial_x^2 u - \beta u, \quad u(t,0) = u(t,L) = 0, \quad u(0,x) = \sin\left(\frac{2\pi x}{L}\right).$$

5.7 Case study: Burgers' Equation *

Burgers' equation * is a generalization of the Hopf's equation, Eq. (5.13), discussed when illustrating the method of characteristics. Recall that the Hopf's equation results in a wave breaking which leads to a non-physical multi-valued solution. Modification of the Hopf's equation by adding dissipation/diffusion results in the Burgers' equation:

$$\partial_t u + u\partial_x u = \partial_x^2 u. \tag{5.65}$$

Like practically every other nonlinear PDE, Burgers' equation seems rather hopeless to resolve at first glance. However, Burgers' equation is in fact special. It allows the Cole-Hopf's transformation, from $u(t;x)$ to $\Psi(t;x)$

$$u(t;x) = -2\frac{\partial_x \Psi(t;x)}{\Psi(t;x)}, \tag{5.66}$$

reducing Burgers' equation to the diffusion equation

$$\partial_t \Psi = \partial_x^2 \Psi. \tag{5.67}$$

The solution to the Cauchy problem associated with Eq. (5.67) can be expressed as an integral convolving the initial profile $\Psi(0;x)$, with the Green function of the diffusion equation described in Eq. (5.56)

$$\Psi(t;x) = \int \frac{dy}{\sqrt{4\pi t}} \exp\left(-\frac{(x-y)^2}{4t}\right) \Psi(0;y). \tag{5.68}$$

This latter expression can be used to find *some* exact solutions to Burgers' equation. Consider, for example, $\Psi(0;x) = \cosh(ax)$. Substitution into Eq. (5.68) and conducting integration over y, one arrives at $\Psi(t;x) = \cosh(ax)\exp(a^2 t)$, which results, according to Eq. (5.66), in stationary (time independent, i.e. standing) "shock" solution to Burgers' equation, $u(t;x) = -2a\tanh(ax)$. Notice that the following more general solution to Burgers' equation corresponds to a shock moving with the constant speed u_0

$$u(t;x) = u_0 - 2a\tanh(a(x - x_0 - u_0 t)).$$

Example 5.7.1. Solve the diffusion equation Eq. (5.67) with the initial conditions $\Psi(0,x) = \cosh(ax) + B\cosh(bx)$. Reconstruct respective $u(t;x)$ solving the Burgers Eq. (5.65). Analyze the result in the regime $b > a$ and $B \gg 1$ and also verify, by building a computational snippet, that the resulting spatio-temporal dynamics corresponds to a large shock "eating" a small shock.

*This auxiliary Section can be dropped at the first reading. Material from this Section will not contribute midterm and finals.

Part III

Optimization

Chapter 6

Calculus of Variations

The main theme of this chapter is the relation of equations to minimal principles. Over-simplifying a bit: to minimize a function $S(q)$ is to solve $S'(q) = 0$. For a quadratic, $S(q) = \frac{1}{2}q^T Kq - q^T g$, where K is positive definite, one indeed has the minimum of $S(q)$ achieved at q_*, which solves $S'(q_*) = Kq_* - g = 0$.

In the example above, q is an n-(finite) dimensional vector, $q \in \mathbb{R}^n$. Consider extending the finite dimensional optimization to an infinite dimensional, continuous, problem where $q(x)$ is a function, say, $q(x) : \mathbb{R} \to \mathbb{R}$, and $S\{q(x)\}$ is a functional, typically an integral with the integrand dependent on $q(x)$ and its derivative, $q'(x)$, for example

$$S\{q(x)\} = \int dx \left(\frac{c}{2}(q'(x))^2 - g(x)q(x) \right).$$

The derivative of the functional over $q(x)$ is called the variational derivative, and by analogy with the finite dimensional example above, one finds that the Euler-Lagrange (EL) equation,

$$\frac{\delta S\{q\}}{\delta q(x)} = 0,$$

solves the problem of minimizing the functional. The goal of this section is to understand the variational derivative and other related concepts in theory and on examples.

6.1 Examples

To have a better understanding of the calculus of variations we start by describing four examples.

6.1.1 Fastest Path

Consider a robot navigating in the (x, y)-plane, and define the function $y = q(x)$ so it describes the path of the robot. For small δx, the arclength of the the robot's path from

(x, y) to $(x + \delta x, y + \delta y)$ can be approximated by $\sqrt{(\delta x)^2 + (\delta y)^2}$ by the Pythagorean theorem, which simplifies to $\sqrt{1 + (q'(x))^2}dx$ for infinitesimal δx . If the plane constitutes a rugged terrain, then the robot's speed may differ at each point in the plane, so we define the scalar-valued positive function $\mu^{-1}(x, y)$ to describe the speed of the robot at each point in the plane. The time it takes for the robot to move from $(x, q(x))$ to $(x + dx, q(x + dx))$ along the path $q(x)$ is

$$L(x, q(x), q'(x)) := \mu(x, q(x))\sqrt{1 + (q'(x))^2}dx.$$

The total time taken to travel along the path which starts at an initial point $(x_i, y_i) := (0, 0)$ and ends at a terminal point $(x_t, y_t) := (a, b)$, where $a > 0$ is

$$S\{q(x)\} = \int_0^a dx L(x, q(x), q'(x)).$$

Subject to a few modest conditions on $\mu(x, y)$, the calculus of variations provides a way to find the optimal path through the domain, that is, the path which minimizes the functional $S\{q(x)\}$, subject to $q(0) = 0$ and $q(a) = b$.

6.1.2 Minimal Surface

Consider making a three-dimensional bubble by dipping a wire loop into soapy water, and then asking whether there is an optimal bubble shape for a given loop. Physics suggests that the bubble will form in whatever shape minimizes the surface area of the soap film.

We formalize this setting as follows. The surface of a bubble is described by the continuously differentiable function, $q(x) : x = (x_1, x_2) \in \mathcal{D} \to q(x) \in \mathbb{R}$, where $\mathcal{D} \subset \mathbb{R}^2$ is bounded. We also assume that at the boundary of $\mathcal{D}$, denoted $\partial\mathcal{D}$ and representing a closed line in the (x_1, x_2) plane, $q(x)$ is fixed/known, i.e., $q(\partial\mathcal{D}) = g(\partial\mathcal{D})$, where $g(\partial\mathcal{D})$ describes the coordinate of the wire loop along the third dimension. Then the optimal bubble results from minimizing the functional

$$S\{q(x)\} = \int_{\mathcal{D}} dx \sqrt{1 + |\nabla_x q(x)|^2} \tag{6.1}$$

over $q(x)$, subject to $q(\partial\mathcal{D}) = g(\partial\mathcal{D})$.

Example 6.1.1. Show that Eq. (6.1) represents a general formula for the surface area of the graph of a continuously differentiable function, $q(x)$, where $x = (x_1, x_2) \in \mathcal{D} \subset \mathbb{R}^2$ and $\mathcal{D}$ is a region in the (x_1, x_2) plane with the smooth boundary.

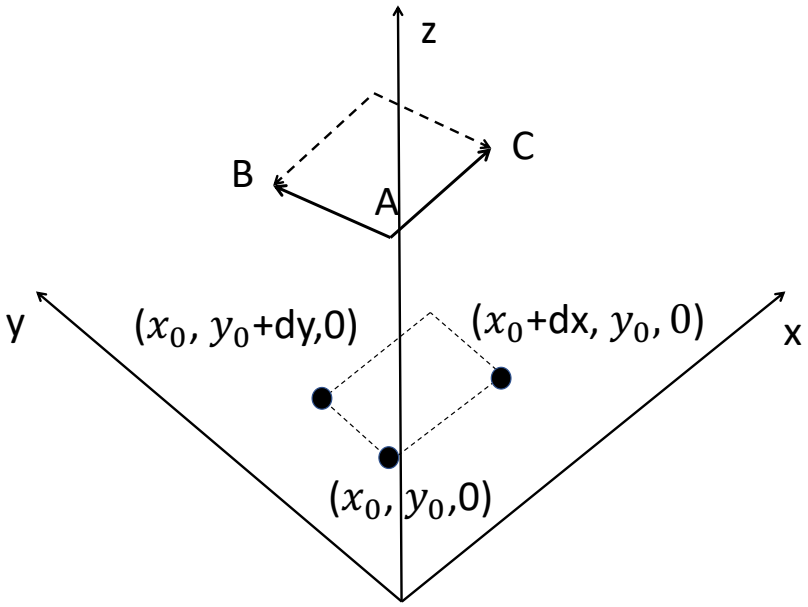


Figure 6.1: Illustration for construction of an infinitesimal surface element in the Example (6.1.1).

Solution. It is sufficient to derive the differential version of Eq. (6.1), i.e. to show that area of a surface element, represented by $q(x)$ around a point, $x^{(0)} = (x_1^{(0)}, x_2^{(0)}) \in \mathcal{D}$, is described by

$$\sqrt{1 + |\nabla_x q(x)|^2} dx = \sqrt{1 + (\partial_{x_1} q(x_1, x_2))^2 + (\partial_{x_2} q(x_1, x_2))^2} dx_1 dx_2,$$

where $\sqrt{1 + |\nabla_x q(x)|^2}$ is evaluated at $x = x^{(0)}$. In this (infinitesimal) case, we can represent the surface of the infinitesimal element by the plane

$$x_3 - x_3^{(0)} = a(x_1 - x_1^{(0)}) + b(x_2 - x_2^{(0)}),$$

in $\mathbb{R}^3$, where $x_3^{(0)} = q(x^{(0)})$, $a = \partial_{x_1} q|_{x=x^{(0)}}$ and $b = \partial_{x_2} q|_{x=x^{(0)}}$. Specifically, we can describe the plane in terms of the following three points in the three dimensional space (see Fig. (6.1) for the illustration)

$$\begin{aligned} \mathrm{A} &= \left(x_1^{(0)}, x_2^{(0)}, x_3^{(0)} \right), \\ \mathrm{B} &= \left(x_1^{(0)} + dx_1, x_2^{(0)}, x_3^{(0)} + a dx_1 \right), \\ \mathrm{C} &= \left(x_1^{(0)}, x_2^{(0)} + dx_2, x_3^{(0)} + b dx_2 \right). \end{aligned}$$

Then area of the surface of the infinitesimal element becomes the absolute value of the cross (vector) product of the following two infinitesimal three dimensional vectors:

$$|u \times v| = |(\mathrm{B} - \mathrm{A}) \times (\mathrm{C} - \mathrm{A})| = |(-a dx_1 dx_2, -b dx_1 dx_2, dx_1 dx_2)| = dx_1 dx_2 \sqrt{1 + a^2 + b^2},$$

where we have used standard vector calculus rules for the cross/vector product in three dimensions, $(y \times z)_i = \sum_{j,k=1,2,3} \varepsilon_{ijk} y_j z_k$, with ε_{ijk} being Levi-Civita (absolute anti-symmetric) tensor in three dimensions. □

6.1.3 Image Restoration

A gray-scale image is described by the function, $q(x) : [0,1]^2 \to [0,1]$, mapping a location, x within the square box, $[0,1]^2 \in \subset R^2$, into a real number between 0 (white), and 1 (black). However, the true image is often corrupted by a noise, and we only observe this noisy image. The task of image restoration is to restore the true image from the noisy observation.

Total Variation (TV) restoration [3] is a method built on the conjecture that the true image is reconstructed from the noisy image, $f(x)$, by minimization of the following functional

$$S\{q(x)\} = \int\limits_{U=[0,1]^2} dx \left((q(x) - f(x))^2 + \lambda |\nabla_x q(x)| \right), \tag{6.2}$$

subject to the Neumann boundary condition, $\boldsymbol{n} \cdot \nabla_x q(x) = 0$ for all $x \in \delta U$, where $\boldsymbol{n}$ is the (unit) vector normal to δU, which is the boundary of the domain U).

6.1.4 Classical Mechanics

Classical mechanics is described in terms of the function, $q(t) : \mathbb{R} \to \mathbb{R}^d$, mapping a time, $t \in \mathbb{R}$, into a d-dimensional real-valued spatial coordinate, $q \in \mathbb{R}^d$. The evolution of the coordinate in time is described in Hamiltonian mechanics by the minimal action, also called Hamiltonian, principle: trajectory, that is understood as describing the evolution of the coordinate in time, is governed by the minimum of the action,

$$S\{q\} := \int_{t_1}^{t_2} dt L\left(t, q(t), \dot{q}(t)\right), \tag{6.3}$$

where $L\left(t, q(t), \dot{q}(t)\right)$ is the system Lagrangian, and $\dot{q}(t) = dq(t)/dt$ is the momentum, under the condition that the values of the coordinate at the initial and final moment of time are fixed, $q(t_1) = q_1$, $q(t_2) = q_2$. An exemplary Hamiltonian dynamics is that of a (unit mass) particle in a potential, $V(q)$, then

$$L\left(t, q(t), \dot{q}(t)\right) = \frac{\dot{q}^2}{2} - V(q). \tag{6.4}$$

6.2 Euler-Lagrange Equations

All the examples can be stated as the minimization of the functional

$$S\{q(x)\} = \int_{\mathcal{D}\subseteq\mathbb{R}^n} dx L\left(x, q(x), \nabla_x q(x)\right),$$

over functions, $q(x)$, with the fixed value at the boundary, $x \in \partial\mathcal{D}: \; q(x) = g(x)$, where $\mathcal{D}$ is bounded with the known value at all points of the boundary, and the Lagrangian L is a given function

$$L : \mathcal{D} \subseteq \mathbb{R}^n \times \mathbb{R}^d \times \mathbb{R}^{d\times n} \to \mathbb{R},$$

of the three variables. It will also be convenient in deriving further relations to consider the three variables in the argument of L and then denoting the respective derivatives, L_x, L_q, and $L_{\nabla q}$. (Note that the variables are: $x \in \mathcal{D} \subseteq \mathbb{R}^n$, $q \in \mathbb{R}^d$, and $\nabla_x q \in \mathbb{R}^{d\times n}$, We will assume in the following that both L and g are smooth.

Theorem 6.2.1 (Necessary condition for optimality)**.** Suppose that $q(x)$ is the minimizer of S, that is

$$S\{\tilde{q}(x)\} \geq S\{q(x)\} \;\&\; \forall x \in \partial\mathcal{D}\; \tilde{q}(x) = q(x) \quad \left(\forall x \in \mathcal{D},\; \forall \tilde{q}(x) \in C^2(\bar{\mathcal{D}} = \mathcal{D} \cup \partial\mathcal{D})\right),$$

then L satisfies the so-called Euler-Lagrange (EL) equations

$$\nabla_x \left(L_{\nabla q}\left(x, q(x), \nabla_x q(x)\right)\right) - L_q\left(x, q(x), \nabla_x q(x)\right) = 0 \quad (\forall x \in \mathcal{D}). \tag{6.5}$$

Sketch of the proof: Consider the perturbation $q(x) \to q(x) + s\delta(x) = \tilde{q}(x)$, where $s \in \mathbb{R}$ and $\delta(x)$ sufficiently smooth and such that is does not change the boundary condition, i.e. $\delta(x) = 0$ $(\forall x \in \partial\mathcal{D})$. Then according to the assumption

$$S\{q(x)\} \leq S\{\tilde{q}(x)\} = S\{q(x) + s\delta(x)\} \quad (\forall x \in \mathcal{D},\; \forall s \in \mathbb{R}).$$

This means that

$$\frac{d}{ds} S\{q(x) + s\delta(x)\}\Big|_{s=0} = 0.$$

Notice that

$$S\{q(x) + s\delta(x)\} = \int\limits_{\mathcal{D}} dx L\left(x, q(x) + s\delta(x), \nabla_x q(x) + s\nabla_x \delta(x)\right)$$

Then, exchanging the orders of differentiation and integration, applying the differentiation (chain) rules to the Lagrangian, and evaluating one of the resulting integrals by parts and

removing the boundary term (because $\delta(x) = 0$ on $\partial\mathcal{D}$), one derives

$$\begin{aligned}
\frac{d}{ds}S\{q(x)+s\delta(x)\}\Big|_{s=0} &= \int_{\mathcal{D}} dx\ \frac{d}{ds}L\left(x, q(x)+s\delta(x), \nabla_x q(x)+s\nabla_x\delta(x)\right)\Big|_{s=0} \qquad (6.6)\\
&= \int_{\mathcal{D}} dx\left(L_q\left(x, q(x), \nabla_x q(x)\right)\cdot\delta(x) + L_p\left(x, q(x), \nabla_x q(x)\right)\cdot\nabla_x\delta(x)\right)\\
&= \int_{\mathcal{D}} dx L_q\left(x, q(x), \nabla_x q(x)\right)\cdot\delta(x) + \int_{\mathcal{D}} dx L_p\left(x, q(x), \nabla_x q(x)\right)\cdot\nabla_x\delta(x)\\
&= \int_{\mathcal{D}} dx\left(L_q\left(x, q(x), \nabla_x q(x)\right) - \nabla_x\cdot L_{\nabla q}\left(x, q(x), \nabla_x q(x)\right)\right)\cdot\delta(x).
\end{aligned}$$

Since the resulting integral should be equal to zero for any $\delta(x)$ one arrives at the desired statement.

□

Remark. Solutions to the EL Eqs. (6.5), $q(x)$, are stationary curves of the functional, $S\{q(x)\}$, and could be minimizers, maximizers or saddle-points. It's for this reason that theorem (6.2.1) provides a necessary, but not sufficient, condition for minimizing $S\{q(x)\}$.

Example 6.2.2. Find the Euler-Lagrange equations (conditions) for

(a) $\int dx\left((q'(x))^2 + \exp(q(x))\right)$,

(b) $\int dx\, q(x)q'(x)$

(c) $\int dx\, x^2(q'(x))^2$

where $q : \mathbb{R} \to \mathbb{R}$.

Solution. In one dimension the Euler-Lagrange equation simplifies to

$$\frac{d}{dx}\left(\frac{\partial L}{\partial q'}\right) - \frac{\partial L}{\partial q} = 0.$$

Since we are not asked to solve the equations this is only a matter of constructing the respective equation for each case.

(a) We identify $L(x, q, q') = (q')^2 + \exp(q)$ and derive

$$\begin{aligned}
\frac{d}{dx}\left(\frac{\partial L}{\partial q'}\right) &= \frac{d}{dx}\left(2q'\right) = 2q'',\\
\frac{\partial L}{\partial q} &= \exp(q).
\end{aligned}$$

Then the Euler-Lagrange equation is

$$2q'' - \exp(q) = 0.$$

(b) We identify $L(x, q, q') = qq'$ and derive

$$\frac{d}{dx}\left(\frac{\partial L}{\partial q'}\right) = \frac{d}{dx}(q) = q',$$
$$\frac{\partial L}{\partial q} = q'.$$

Then the Euler-Lagrange equation is

$$q' - q' = 0.$$

This implies that any function q satisfies the Euler-Lagrange equation, which in turn implies that the functional does not have any minima or maxima.

(c) We identify $L(x, q, q') = x^2(q')^2$ and derive

$$\frac{d}{dx}\left(\frac{\partial L}{\partial q'}\right) = \frac{d}{dx}\left(2x^2q'\right) = 2x^2q'' + 4xq',$$
$$\frac{\partial L}{\partial q} = 0.$$

Then the Euler-Lagrange equation is

$$x^2q'' + 2xq' = 0.$$

Example 6.2.3. Consider the shortest path version of the fastest path problem set in Section 6.1.1, that is the case of $g(x, y) = 1$:

$$\min_{\{q(x)|x\in[0,a]\}} \int_0^a dx\sqrt{1 + (q'(x))^2}dx\Bigg|_{q(0)=0,\ q(a)=b}.$$

Find the Euler-Lagrange (EL) condition on $q(x)$.

Solution. The Euler-Lagrange condition on $q(x)$ becomes

$$\begin{aligned}
0 &= \nabla_x\left(L_{\nabla q}\left(x, q(x), \nabla_x q(x)\right)\right) - L_q\left(x, q(x), \nabla_x q(x)\right) \\
&= \frac{d}{dx}\frac{q'(x)}{\sqrt{1 + (q'(x))^2}} - 0 \\
&\to \frac{q'(x)}{\sqrt{1 + (q'(x))^2}} = \text{constant} \\
&\to q'(x) = \text{constant} \\
&\to q(x) = \frac{b}{a}x,
\end{aligned}$$

where at the last step we accounted for the boundary condition. The shortest (optimal) path connects initial and final points by a straight line.

Exercise 6.1. (a) Write the Euler-Lagrange equation for the general case of the fastest path problem formulated in Section 6.1.1. (b) Find an example of, $\mu(x,y)$, resulting in the quadratic optimal path, i.e. $q(x) = \frac{b}{a^2}x^2$. (Your solution for $\mu(x,y)$ should be independent of a and b.)

Example 6.2.4. Let us derive the Euler-Lagrange condition for the Minimal Surface problem introduced in Section 6.1.2:

$$\min_{\{q(x)\}} \int_{\mathcal{D}} dx \sqrt{1+|\nabla_x q(x)|^2}\Big|_{q(\partial\mathcal{D})=g(\partial\mathcal{D})}.$$

Solution. In this case Eq. (6.5) becomes

$$\begin{aligned} 0 &= \nabla_x \left(L_{\nabla q}\left(x,q(x),\nabla_x q(x)\right)\right) - L_q\left(x,q(x),\nabla_x q(x)\right) \\ &= \nabla_x \cdot \left(\frac{\nabla_x q(x)}{\sqrt{1+|\nabla_x q(x)|^2}}\right) \\ &\to -\nabla_x q(x)\cdot\nabla_x^2 q \nabla_x q + (1+|\nabla_x q(x)|^2)\nabla_x^2 q = 0. \end{aligned} \tag{6.7}$$

6.3 Phase-Space Intuition and Relation to Optimization (finite dimensional, not functional)

Consider the special case of the fastest path problem of Section 6.1.1, which is still more general than the shortest path problem discussed in the Example 6.2.3, where $\mu(x)$ depends only on x. In this case the action is

$$S\{q(x)\} = \int_0^a dx\mu(x)\sqrt{1+(q'(x))^2} = \int_0^a ds\mu(x),$$

where ds is the element of arc-length of the curve $u(x)$:

$$ds = \sqrt{1+(q'(x))^2}dx = \sqrt{dx^2+dq^2}.$$

The Lagrangian and its partial derivatives are, $L(x;q(x);q'(x)) = \mu(x)\sqrt{1+(q'(x))^2}$, $L_q = 0$, $L_{q'} = \mu(x)q'/\sqrt{1+(q')^2}$. Then the Euler-Lagrange equation becomes

$$\frac{d}{dx}\left(\frac{\mu(x)q'(x)}{\sqrt{1+(q'(x))^2}}\right) = 0,$$

which results in

$$\frac{\mu(x)q'(x)}{\sqrt{1+(q'(x))^2}} = \mu(x)\sin(\theta) = \text{constant}, \tag{6.8}$$

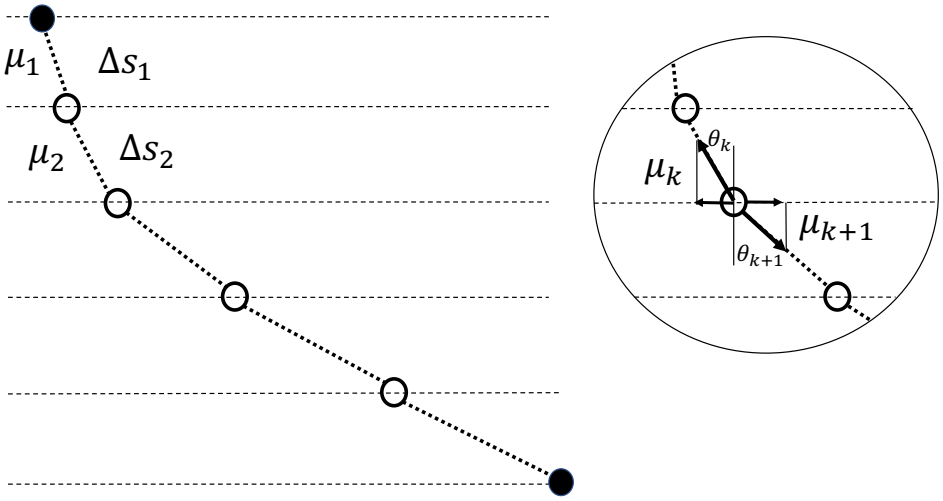


Figure 6.2: Variational Calculus via Discretization and Optimization.

where θ is the angle in the (q, x) space between the tangent to $q(x)$ and the x-axis.

It is instructive to derive Eq. (6.8) bypassing the variational calculus, taking instead perspective of standard optimization, that is optimizing over a finite number of continuous variables. To make this link we need, first, to **discretize** the action, $S\{q(x)\}$:

$$\begin{aligned} S\{q(x)\} &\approx S_k(\cdots, q_k, \cdots) = \sum_k \mu_k \Delta s_k = \sum_k \mu_k \sqrt{1 + \left(\frac{q(x_k) - q(x_{k-1})}{\Delta}\right)^2}\,\Delta \\ &= \sum_k \mu_k \sqrt{1 + \left(\frac{q_k - q_{k-1}}{\Delta}\right)^2}\,\Delta \end{aligned}$$

where Δ is the size of a step in x. i.e. $\Delta = x_{k+1} - x_k$, $\forall k$, and Δs_k is the length of the k-th segment of the discretized curve, illustrated in Fig. (6.2). Then, second, we look for **extrema** of S_k over q_k, i.e. require that $\forall\ k:\ \partial_{q_k} S_k = 0$. The result is the discretized version of the Euler-Lagrange Eqs. (6.8):

$$\begin{aligned} \forall k: \qquad & \frac{\mu_{k+1}(q_{k+1} - q_k)}{\sqrt{1 + \left(\frac{q_{k+1} - q_k}{\Delta}\right)^2}} = \frac{\mu_k (q_k - q_{k-1})}{\sqrt{1 + \left(\frac{q_k - q_{k-1}}{\Delta}\right)^2}} \\ \to \quad & \mu_{k+1} \sin\theta_{k+1} = \mu_k \sin\theta_k. \end{aligned}$$

6.4 Towards Numerical Solutions of the Euler-Lagrange Equations *

Here * we discuss the image restoration problem set up in Section 6.1.3. We will derive the Euler-Lagrange equations and observe that the resulting equations are difficult to solve. We will then use this case to illustrate the theoretical part (philosophy) of solving the Euler-Lagrange equations numerically. Following [4], we will use the example to discuss gradient descent in this Section and then also primal-dual method below in Section 6.7.

6.4.1 Smoothing Lagrangian

The TV functional (6.2) is not differentiable at $\nabla_x q(x) = 0$, which creates difficulty for variations. One way to bypass the problem is to smooth the Lagrangian, considering

$$S_\varepsilon\{q\} = \int\limits_{[0,1]^2} dx \left(\frac{(q(x) - f(x))^2}{2} + \lambda \sqrt{\varepsilon^2 + (\nabla_x q(x))^2} \right), \tag{6.9}$$

where ε is small and positive. The Euler-Lagrange equations for the smoothed action (6.9) are

$$\forall x \in [0,1]^2 : \quad q - \lambda \nabla_x \cdot \frac{\nabla_x q}{\sqrt{\varepsilon^2 + (\nabla_x q(x))^2}} = f, \tag{6.10}$$

with the homogeneous Neumann boundary conditions, $\forall x \in \partial[0,1]^2 : \quad \partial q(x)/\partial n = 0$, where n denotes normal to the boundary of the $[0,1]^2$ domain. Finding analytical solutions to Eq. (6.10) for an arbitrary f is not possible. We will discuss ways to solve Eq. (6.10) numerically in the following.

6.4.2 Gradient Descent and Acceleration

We will start this part with a disclaimer. The discussion below of the numerical procedure for solving Eq. (6.10) is not fully comprehensive. We add it here for completeness, delegating details to Math 589, and also aiming to emphasize connections between numerical PDE analysis and optimization algorithms.

A standard numerical scheme for solving Eq. (6.10) originating from optimization of the action is gradient descent. It is useful to think about the gradient descent algorithm by introducing an extra "computational time" dimension, which will be discrete in implementation but can also be thought of (for the purpose of analysis and gaining intuition) as

*This auxiliary Section can be dropped at the first reading. Material from this Section will not contribute midterm and finals.

continuous. Consider the following equation

$$\forall x \in [0,1]^2,\ t>0: \quad \partial_t \upsilon + \upsilon - \lambda \nabla_x \cdot \frac{\nabla_x \upsilon}{\sqrt{\varepsilon^2 + (\nabla_x \upsilon(x))^2}} = f, \tag{6.11}$$

for, $\upsilon(t;x)$, representing estimation at the computational time t for $q(x)$ solving Eq. (6.10), with the initial conditions, $\forall x: \quad \upsilon(0;x) = f(x)$, and the boundary conditions, $\forall x \in \partial[0,1]^2: \quad \partial\upsilon(x)/\partial n = 0$. Eq. (6.11) is a nonlinear heat equation. Close to the equilibrium the equation can be linearized. Discretizing the linear diffusion equation on the spatio-temporal grid with spacing, Δt, and, Δx, and looking for the dynamic (time-derivative) term balancing the diffusion term (containing second order spatial-derivative) one arrives at the following rough empirical estimation

$$\Delta t \sim \frac{\varepsilon(\Delta x)^2}{\lambda}.$$

The estimation suggests that the temporal step needs to be really small (square of the spatial step) to guarantee that the numerical scheme is proper (not stiff). The condition becomes even more demanding with decrease of the regularization parameter, ε.

One way to improve the gradient scheme (to make it less stiff) is to replace the diffusion Eq. (6.11) by the (damped) wave equation

$$\forall x \in [0,1]^2,\ t>0: \quad \partial_t^2 \upsilon + a\partial_t \upsilon + \upsilon - \lambda \nabla_x \cdot \frac{\nabla_x \upsilon}{\sqrt{\varepsilon^2 + (\nabla_x \upsilon(x))^2}} = f, \tag{6.12}$$

where a is the damping coefficient. Acting by analogy with the diffusive case, let us make an empirical estimate for the balanced choice of the spatial discertization step, Δx, temporal discretization step, Δt, and of the damping coefficient. Linearising the nonlinear wave Eq. (6.12) and then requiring that the ∂_t^2 (temporal oscillation) term, the $a\partial_t$ (damping) term and the $(\lambda/\varepsilon)\nabla_x^2$ (diffusion) term are balanced one arrives at the following estimate

$$(\Delta t)^2 \sim \frac{\Delta t}{a} \sim \frac{\varepsilon(\Delta x)^2}{\lambda},$$

which results in a much less demanding linear scaling, $\Delta t \sim \Delta x$.

This transition from the overdamped relaxation to balancing damping with oscillations corresponds to the Polyak's heavy-ball method [5] and Nesterov's accelerated gradient descent method [6], which are now used extensively (often with addition of a stochastic component) in training of the modern Neural Networks. Both methods will be discussed later in the course, and even more in the companion Math 575 course. Notice also that an additional material on modern, continuous-time interpretation of the acceleration method and other related algorithms can be found in [7, 8]. See also Sections 2.3 and 3.6 of [4]

We will come back to the image-restoration problem one more time in Section 6.7.2 where we discuss an alternative, primal-dual algorithm.

6.5 Dependence of the action on the end-points

Consider $x \in \mathbb{R}$ and let the points $A_0 := (x_0, q_0)$ and $A_1 := (x_1, q_1)$ be given, and let $\mathcal{F}$ be the family of continuously differentiable functions on $[x_0, x_1]$ that satisfy $q(x_0) = q_0$ and $q(x_1) = q_1$. For a given Lagrangian $L(x, q, q')$, let $S : \mathcal{F} \to \mathbb{R}$ be the functional

$$S\{q(x)\} = \int_{x_0}^{x_1} L(x, q(x), q'(x))\, dx. \tag{6.13}$$

In section 6.2, we showed that if a function $q(x)$ satisfies the EL equations, $\frac{d}{dx}L_{q'} - L_q = 0$, then $q(x)$ is a stationary curve of S and therefore a candidate for a minimizer of S. (See also discussion of the EL Theorem 6.2.1 and the following remark.)

Notation. In the following, we abuse notation a little and use the symbol S to denote both the action-functional

$$S\{q(x)\} = \int_{x_0}^{x_1} L(x, q(x), \dot{q}(x))\, dx$$

and also the action-function

$$S(A_0, A_1) = S(x_0, q_0, x_1, q_1) = \min_{q(x)\in\mathcal{F}} \int_{x_0}^{x_1} L(x, q(x), \dot{q}(x))\, dx.$$

Therefore, $S(A_0, A_1)$ corresponds to $S\{q(x)\}$ evaluated at solutions $q(x)$ of the associated EL equations and should thus be understood as a function of the end-points, A_0 and A_1, of $q(x)$.

The following statement gives a very intuitive, geometrical interpretation for the derivatives of the action over the end-point parameters

Theorem 6.5.1 (End-point derivatives of the action)**.**

$$(a) \qquad \partial_{x_1} S(A_0; A_1) = (L - q'L_{q'})\big|_{x=x_1} \qquad \partial_{x_0} S(A_0; A_1) = -(L - q'L_{q'})\big|_{x=x_0}, \tag{6.14}$$

$$(b) \qquad \partial_{q_1} S(A_0; A_1) = L_{q'}\big|_{x=x_1} \qquad \partial_{q_0} S(A_0; A_1) = -L_{q'}\big|_{x=x_0}. \tag{6.15}$$

Proof. Here (and as custom in this course) we will only sketch the proof. Let us focus, without loss of generality, on the part of the theorem concerning derivatives with respect to x_1 and q_1, i.e. the final end-point of the critical path.

Let us first keep the final independent variable fixed at x_1 but move the final position by dq, as shown in Fig. 6.5 Left. The trajectory $q(x)$ will vary by $\delta q(x)$, where $\delta q(x_0) = 0$ and $\delta q(x_1) = dq$.

At each time x, we can estimate the value of $L(x, q + \delta q, q' + \delta q')$ from the first-order Taylor expansion

$$L(x, q + \delta q, q' + \delta q') \approx L(x, q, q') + L_q(x, q, q') \cdot \delta q + L_{q'}(x, q, q') \cdot \delta q'. \tag{6.16}$$

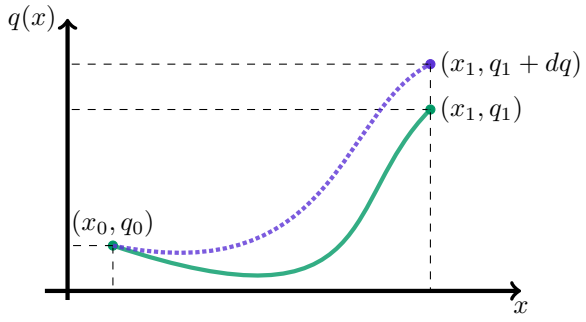


Figure 6.3: Critical curves from (x_0, q_0) to (x_1, q_1) (green) and to $(x_1, q_1 + dq)$ (purple).

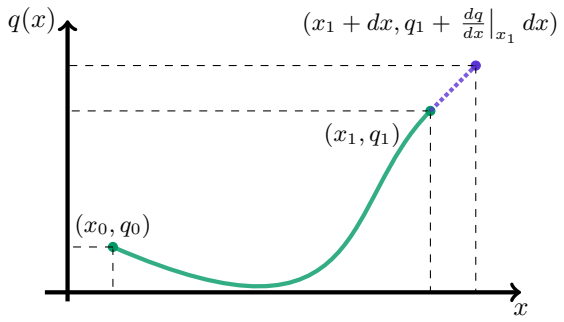


Figure 6.4: Critical curves from (x_0, q_0) to (x_1, q_1) (green) and to $(x_1 + dx, q_1 + dq)$ (purple) where $dq = \frac{dq}{dx}\Big|_{x_1} dx$.

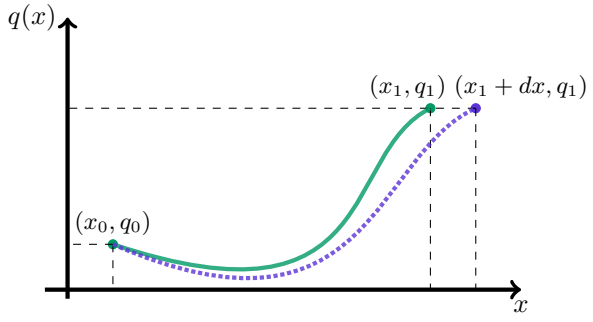


Figure 6.5: Critical curves from (x_0, q_0) to (x_1, q_1) (green) and to $(x_1 + dx, q_1)$ (purple).

Therefore, the variation of the Lagrangian is given by

$$dL = L_q\, d\delta q + L_{q'}\, d\delta q', \tag{6.17}$$

and the variation of the action is

$$dS = \int_{x_0}^{x_1} dx \left(L_{q'} \delta q' + L_q \delta q \right). \tag{6.18}$$

The relation $\delta q' = d\delta q/dx$, together with the Euler-Lagrange Eqs. (6.21), allows us to rewrite Eq. (6.18) as

$$dS = \int_{x_0}^{x_1} dx \left(L_{q'} \frac{d}{dx} \delta q + \delta q \frac{d}{dx} L_{q'} \right) = \int_{x_0}^{x_1} d\left(L_{q'} \delta q \right) = \left(L_{q'} \delta q \right) \Big|_{x_0}^{x_1} = L_{q'} \big|_{x_1} dq. \tag{6.19}$$

Therefore, as we kept the final independent variable fixed, $dS = \partial_{q_1} S dq$, and one arrives at the desired statement

$$\frac{\partial S}{\partial q_1} = L_{q'} \big|_{x_1}. \tag{6.20}$$

Now consider variation of the action extended from $A_1 = (x_1, q_1)$ to $(x_1 + dx, q_1 + q'(x_1)dx)$:

$$dS = Ldx = \frac{\partial S}{\partial x_1} dx + \frac{\partial S}{\partial q_1} q'(x_1) dx = \frac{\partial S}{\partial x_1} dx + L_{q'} \big|_{x_1} q'(x_1) dx = \left(\frac{\partial S}{\partial x_1} + q' L_{q'} \right)_{x_1} dx,$$

where we utilize Eq. (6.20). Finally, we derive

$$\frac{\partial S}{\partial x_1} = \left(L - q' L_{q'} \right) \big|_{x_1}.$$

□

Example 6.5.2. Find the minimzers of the functional

$$S\left\{q(x)\right\} = \int_0^1 (q'^2 + xq) dx$$

for the case where (a) $q(0) = 1$ and $q(1) = 0$, and where (b) $q(0) = 1$ and $q(1)$ is free.

Solution. The stationary curve (by definition) satisfies the EL equations:

$$\frac{d}{dx} L_{q'} - L_q = 0 \;\Rightarrow\; \frac{d}{dx} 2q' - x = 0 \;\Rightarrow\; q'' - x = 0,$$

thus resulting in $q(x) = \frac{1}{6}x^3 + c_1 x + c_2$. For part (a), the values of c_1 and c_2 are determined from the requirements that $q(0) = 1$ and $q(1) = 0$ giving $c_1 = \frac{7}{6}$ and $c_2 = 1$. For part (b), given that the value $q(1)$ is free, we must find the optimal value of $q(1)$ by solving $\partial_{q_1} S = 0$:

$$0 = \frac{\partial S}{\partial q_1} = L_{q'} \big|_{x=1} = (2q') \big|_{x=1}.$$

Therefore, in this case the optimal value of $q(1)$ occurs when $q'(1) = 0$. In this case, the corresponding values of c_1 and c_2 are $c_1 = -\frac{1}{2}$ and $c_2 = 1$.

Exercise 6.2. Find all the critical function(s), $q(x)$, of the functional

$$S\left\{q(x)\right\} = \int_0^a \left(q^2 + 2qq' + (q')^2 + 1\right) dx$$

where $q(0) = 0$ and at some positive but not fixed, a (i.e., such, a, that is by itself subject to optimization), $q(a) = 1$.

6.6 Variational Principle of Classical Mechanics

In this Section we apply the principle of minimal action (also called variational principle, or Hamiltonian principle), to the case of the classical mechanics, already highlighted in Section 6.1.4. (See also [9], which we follow in this Section.)

6.6.1 Noether's Theorem & time-invariance of space-time derivatives of action

In the case of the classical mechanics, introduced in Section 6.1.4, the Euler-Lagrange Eqs. (6.5) are

$$\frac{d}{dt}L_{\dot{q}} - L_q = 0, \tag{6.21}$$

where $L(t, q(t), \dot{q}(t)) : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Let us consider the case when the Lagrangian does not depend explicitly on time. (It may still depend on time implicitly via $q(t)$ and $\dot{q}(t)$, i.e. $L(q(t), \dot{q}(t))$.) In this case, and quite remarkably, the Euler-Lagrange equation can be rewritten as a conservation law. Indeed,

$$\frac{d}{dt}\left(\dot{q}\cdot L_{\dot{q}} - L\right) = \ddot{q}\cdot L_{\dot{q}} + \dot{q}\cdot\frac{d}{dt}L_{\dot{q}} - L_q\cdot\dot{q} - L_{\dot{q}}\cdot\ddot{q} = \dot{q}\cdot\left(\frac{d}{dt}L_{\dot{q}} - L_q\right) = 0,$$

where the last equality is due to Eq. (6.21).

We have just introduced the Hamiltonian, $H = \dot{q}\cdot L_{\dot{q}} - L$, representing energy stored within the mechanical system instantaneously, and proved that if the Lagrangian (and thus Hamiltonian) does not have explicit dependence on time, the Hamiltonian (and energy) is conserved. This is a particular case of the Noether's theorem.

Notice, that symmetry under a parametrically continuous change, such as one just explored (consisting in invariance of the Lagrangian under the time shift), is generally a stronger property than a conservation law.

To state theorems expressing the invariance(s) we need the following definition.

Definition 6.6.1 (Invariance of the Lagrangian)**.** Consider a family of transformations of $\mathbb{R}^d$, $h_s(q) : \mathbb{R}^d \to \mathbb{R}^d$, where $s \in \mathbb{R}$ and $h_s(q)$ is continuous in both, q, and (parameter), s, and $h_0(q) = q$. We say that a Lagrangian, $L(q(t), \dot{q}(t)) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, is invariant under the action of the family of transformations of $\mathbb{R}^d$, $h_s(q) : \mathbb{R}^n \to \mathbb{R}^d$, if $L(q, \dot{q})$ does not change when $q(t)$ is replaced by $h_s(q(t))$, i.e. if for any function $q(t)$ we have

$$L(h_s(q(t)), \frac{d}{dt} h_s(q(t))) = L(q(t), \frac{d}{dt} q(t)).$$

Common examples of $h_s(q(t))$ in the classical mechanics include

- translational invariance, $h_s(q(t)) = q(t) + se$, where e is the unit vector in $\mathbb{R}^n$ and s is the distance of the transformation;
- rotational invariance, $h_s(q(t)) = R_e(s)q(t)$, around the line through the origin defined by the unit vector e;
- combination of translational invariance and rotational invariance (cork-screw motion): $h_s(q(t)) = aes + R_e(s)q(t)$, where a is a constant.

Theorem 6.6.2 (Noether's theorem (1915))**.** If the Lagrangian L is invariant under the action of a one-parameter family of transformations, $h_s(q(t))$, then the quantity,

$$I(q(t), \dot{q}(t)) := L_{\dot{q}} \cdot \frac{d}{ds} \left(h_s(q(t)) \right)_{s=0}, \tag{6.22}$$

is constant along any solution of the Euler-Lagrange Eq. (6.21). Such a constant quantity is called an integral of motion.

Proof. Following discussion of Section 6.5, we consider the action-function, $S(t_0, q_0, t_1, q_1)$, i.e. a minimum value of the action-functional, analyzed as a function of the end points, (t_0, q_0) to (t_1, q_1). Theorem 6.5.1, applied to the case of the classical mechanics, states that

$$\begin{aligned} &(a) \qquad \partial_{t_1} S(A_0; A_1) = (L - \dot{q} L_{\dot{q}})_{t=t_1} \qquad \partial_{t_0} S(A_0; A_1) = -(L - \dot{q} L_{\dot{q}})_{t=t_0}, \\ &(b) \qquad \partial_{q_1} S(A_0; A_1) = L_{\dot{q}}|_{t=t_1} \qquad \partial_{q_0} S(A_0; A_1) = -L_{\dot{q}}|_{t=t_0}. \end{aligned}$$

The assumptions of Noether's theorem require that the Lagrangian is invariant under the transformation $q(t) \to h_s(q(t)$, which gives

$$S(t_0, h_s(q_0); t_1, h_s(q_1)) = S(t_0, q_0; t_1, q_1), \quad \forall s. \tag{6.23}$$

Differentiating both sides of Eq. 6.23 with respect to s, applying Theorem 6.5.1, and evaluating the result at $s = 0$, leads us to

$$\begin{aligned} 0 &= \partial_{q_0} S \cdot \frac{d}{ds} h_s(q(t_0)) \Big|_{s=0} + \partial_{q_1} S \cdot \frac{d}{ds} h_s(q(t_1)) \Big|_{s=0} \\ &= -L_{\dot{q}}(q(t_0), \dot{q}(t_0)) \cdot \frac{d}{ds} h_s(q(t_0)) \big|_{s=0} + L_{\dot{q}}(q(t_1), \dot{q}(t_1)) \cdot \frac{d}{ds} h_s(q(t_1)) \big|_{s=0} \end{aligned}$$

Since t_1 can be chosen arbitrarily, it proves that Eq. (6.22) is constant along the solution of the Euler-Lagrange Eq. (6.21). □

Exercise 6.3. For $q(t) \in \mathbb{R}^3$, where $t \in [t_0, t_1]$, and each of the following families of transformations, find the explicit form of the conserved quantity given by the Noether's theorem (assuming that the respective invariance of the Lagrangian holds)

(a) space translation in the direction, e: $h_s(q(t)) = q(t) + se$.

(b) rotation through angle s around the vector, $e \in \mathbb{R}^3$: $h_s(q(t)) = R_e(s)q(t)$.

(c) helical symmetry, $h_s(q(t)) = aes + R_e(s)q(t)$, where a is a constant.

6.6.2 Hamiltonian and Hamilton Equations: the case of Classical Mechanics

Let us utilize the specific structure of the classical mechanics Lagrangian which is split, according to Eq. (6.4), into a difference of the kinetic energy, $\dot{q}^2/2$, and the potential energy, $V(q)$. Making the obvious observation, that the minimum of the functional

$$\int dt\, \tfrac{1}{2}\left(\dot{q} - p\right)^2,$$

over $\{p(t)\}$ is achieved at $\forall t:\ \dot{q} = p$, and then stating the kinetic term of the classical mechanics action, that is the first term in Eq. (6.4), in terms of an auxiliary optimization

$$\int dt \frac{\dot{q}^2}{2} = \max_{\{p(t)\}} \int dt \left(p\dot{q} - \frac{p^2}{2}\right), \tag{6.24}$$

and substituting the result in Eqs. (6.3,6.4), one arrives at the following, alternative, variational formulation of the classical mechanics

$$\min_{\{q(t)\}} \max_{\{p(t)\}} \int dt \left(p\dot{q} - H(q;p)\right) \tag{6.25}$$

$$H(q;p) := \frac{p^2}{2} + V(q), \tag{6.26}$$

where p and H are defined as the momentum and Hamiltonian of the system. Turning the second (Hamiltonian) principle of the classical mechanics into the equations (which, like EL equations, are only sufficient conditions of optimality) one arrives at the so-called Hamiltonian equations

$$\dot{q} = \frac{\partial H(q;p)}{\partial p}, \quad \dot{p} = -\frac{\partial H(q;p)}{\partial q}. \tag{6.27}$$

Example 6.6.3. (a) [Conservation of Energy] Show that in the case of the time independent Hamiltonian (i.e. in the case of $H(q;p)$ considered so far), H, is also the energy which is conserved along the solution of the Hamiltonian equations (6.27).
(b) [Conservation of Momentum] Show that if the Lagrangian does not depend explicitly on one of the coordinates, say $q^{(1)}$ where $q = (q^{(1)}, \cdots)$, then the corresponding momentum, $\partial L/\partial \dot{q}^{(1)}$, is constant along the physical trajectory, given by the solutions of either EL or Hamiltonian equations.

Solution. **(a)** Full time derivative of the Hamiltonian is

$$\frac{dH}{dt} = \frac{\partial H}{\partial q}\dot{q} + \frac{\partial H}{\partial p}\dot{p}.$$

The Hamiltonian equations are

$$\dot{q} = \frac{\partial H}{\partial p}, \quad \dot{p} = -\frac{\partial H}{\partial q}.$$

Combining the two expressions

$$\frac{dH}{dt} = \frac{\partial H}{\partial q}\dot{q} + \frac{\partial H}{\partial p}\dot{p} = -\dot{p}\dot{q} + \dot{q}\dot{p} = 0,$$

we arrive at the desired statement.
(b) Suppose that L does not depend on $q^{(1)}$, then $\partial L/\partial q^{(1)} = 0$. Then, the EL equations

$$0 = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^{(1)}}\right) - \frac{\partial L}{\partial q^{(1)}} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^{(1)}}\right),$$

imply that $\partial L/\partial \dot{q}^{(1)}$ is constant in time.

Notice, that the Hamiltonian system of Eqs. (6.27) becomes even more elegant in the vector form

$$\dot{z} = -J\nabla_z H(z) = -\nabla_z J H(z), \quad z := \begin{pmatrix} q \\ p \end{pmatrix}, \quad J := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{6.28}$$

where the 2×2 matrix represents two-dimensional rotation (clock-wise in the (q,p)-space).

6.6.3 Hamilton-Jacobi equation

Let us work a bit more with the critical/optimal trajectory/path, $\{q(t); t \in [t_0 = 0, t_1]\}$, solving the Euler-Lagrange Eqs. (6.21), choosing the initial time to be equal to 0, fixing the initial position, $q(0) = q_0$, then analyzing dependence of the action on the final time, t_1, and position, q_1. That is we continue the thread of the Theorem 6.5.1 and consider the action-function as a function of $A_1 = (t_1, q_1)$, i.e., of the final position of the critical path.

Indeed, let us re-derive in a bit different, but equivalent, form the main results of the Theorem 6.5.1. Assuming that the action-function is a sufficiently smooth function of the arguments, t_1 and q_1, one would like to introduce (and interpret) derivatives of action over t_1 and q_1, and then check if the derivatives are related to each other. Consider, first, derivative of the action-function over t_1:

$$\mathcal{S}_{t_1} := \partial_{t_1}\mathcal{S}(t_1;q_1) = \partial_t \int\limits_0^{t_1} dtL\left(q(t),\dot{q}(t)\right) = L(t_1,q_1) + \int\limits_0^{t_1} dt\left(L_q\partial_{t_1}q(t) - L_{\dot{q}}\partial_{t_1}\dot{q}(t)\right)$$

$$= L(t_1,q_1) + \int\limits_0^{t_1} dt\partial_{t_1}q(t)\left(L_q + \frac{d}{dt}L_{\dot{q}}\right) - L_{\dot{q}}\partial_t q(t)|_0^{t_1} = \left(L - L_{\dot{q}}\dot{q}\right)_{t=t_1}, \tag{6.29}$$

where we have made an integration by parts and used that, $\partial_t q(t)|_{t=0} = 0$, $\partial_{t_1}q(t)|_{t=t_1} = \dot{q}(t_1)$, utilized the Euler-Lagrange equations Eq. (6.5), $\forall t\in[0,t_1]:\ L_q - \frac{d}{dt}L_{\dot{q}} = 0$.

Next, let us evaluate the derivative of the action-function over the coordinate at the final position, q_1:

$$\mathcal{S}_{q_1} := \partial_q\mathcal{S}(t_1;q_1) = \partial_{q_1}\int\limits_0^{t_1} dtL\left(q(t),\dot{q}(t)\right) = \int\limits_0^{t_1} dt\left(L_q\partial_{q_1}q(t) + L_{\dot{q}}\partial_{q_1}\dot{q}(t)\right)$$

$$= \int\limits_0^{t_1} dt\partial_{q_1}q(t)\left(L_q - \frac{d}{dt}L_{\dot{q}}\right) + L_{\dot{q}}\partial_{q_1}q(t)|_0^{t_1} = L_{\dot{q}}|_{t=t_1}. \tag{6.30}$$

In the case of the classical mechanics, when the Lagrangian is factorized into a difference of the kinetic energy and the potential energy terms, the object on right hand sides of Eq. (6.29) turns into the minus Hamiltonian, defined above in Eq. (6.26), and the right hand side of Eq. (6.30) becomes the momentum, then $p = \dot{q}$. In the case of a generic (not factorizable) Lagrangian, one can use the right hand side of and Eq. (6.29) and Eq. (6.30) as the definitions of the minus Hamiltonian of the system and of the system momentum, respectively,

$$\forall t:\ p := L_{\dot{q}},\quad H(t;q;p) := L_{\dot{q}}\dot{q} - L, \tag{6.31}$$

where the Hamiltonian is considered as a function of the current time, t, coordinate, $q(t)$, and momentum, $p(t)$.

Combining Eqs. (6.29,6.30,6.31), that is (a) and (b) of the Theorem 6.5.1 and the definitions of the momentum and the Hamiltonian, one arrives at the Hamilton-Jacobi (HJ) equation

$$\mathcal{S}_{t_1} + H(t_1;q_1;\partial_{q_1}S) = 0, \tag{6.32}$$

which provides a nonlinear first order PDE representation of the classical mechanics.

It is important to stress that if one knows the initial (at $t = 0$) values of the action-function, S, and of its derivative, $\partial_q S$, and also of the explicit expression of the Hamiltonian in terms of the time, coordinate and momentum at all moments of time, Eq. (6.32) combined with the initial conditions represents a Cauchy initial value problem, therefore resulting in solving of the HJ equation unambiguously. This is a rather remarkable and strong statement with many important consequences and generalizations. The statement is remarkable because because one gets the unique solution of the optimization problem in spite of the fact that solution of the EL equation is not necessarily unique (remember it is a sufficient but not necessary condition for the minimum action, i.e. there may be multiple solutions of the EL equations). Consequences of the HJ equations will be seen later when we will discuss its generalization to the case of theoptimal control, called the Bellman-Hamilton-Jacobi (BHJ) equation. HJ equation, discussed here, and BHJ discussed in Section are linked ultimately to the concept of the Dynamic Programming (DP), also discussed later in the course.

Let us re-emphasize, that the schematic derivation of the HJ-equation (just provided) has revealed the meaning of the action derivative over (final) time and over the (final) coordinate. We have learned that, $\partial_{t_1} S$, is nothing but minus Hamiltonian, while $\partial_{q_1} S$, is simply momenta, $p_1 = p(t_1)$ (also equal to velocity as in these notes we follow the convention of unit mass).

Let us provide an alternative (and as simple) derivation of the HJ-equation, based primarily on the differentials. Given transformation from representation of the action as a functional, of $\{q(t); t \in [0, t_1]\}$, to its representation as a function of t_1 and q_1, $\mathcal{S}\{q(t)\} \to \mathcal{S}(t_1; q_1)$, one rewrites Eqs. (6.3,6.4)

$$\mathcal{S} = \int p\, dq - \int H\, dt,$$

which then implies the following differential form

$$d\mathcal{S} = \frac{\partial \mathcal{S}}{\partial t_1} dt_1 + \frac{\partial \mathcal{S}}{\partial q_1} dq_1,$$

so that

$$\partial_{t_1} \mathcal{S} = -H(t_1; q_1; p_1), \quad \partial_{q_1} \mathcal{S} = p_1,$$

resulting (in combination) in the HJ Eq. (6.32).

So far it was important to differentiate the current moment of time, $t \in [0, t_1]$ and the final moment of time, t_1. However, and once the HJ equations are derived we may return, simplifying notations (and when it does not lead to a confusion), to using t interchangeably for both.

Example 6.6.4. Find and solve the HJ equation for a free particle.

In this case

$$H = \frac{p^2}{2}.$$

Therefore, the HJ equation becomes

$$\frac{(\partial_q \mathcal{S})^2}{2} = -\partial_t \mathcal{S}.$$

Look for solution of the HJ equation in the form $\mathcal{S} = f(q) - Et$. One derives $f(q) = \sqrt{2E}q - c$, and therefore the general solution of the HJ equation becomes

$$S(t;q) = \sqrt{2E}q - Et - c.$$

Exercise 6.4. Find and solve the HJ equation for a two dimensional oscillator (unit mass and unit elasticity) in spherical coordinates, i.e. for the Hamiltonian system with the action functional

$$\mathcal{S}\{r(t), \varphi(t)\} = \int dt \left(\frac{1}{2} \left(\dot{r}^2 + r^2 \dot{\varphi}^2 \right) - \frac{1}{2} r^2 \right).$$

We conclude this very brief discussion of the classical/Hamiltonian mechanics by mentioning that in addition to its relevance to the concepts of Optimal Control and Dynamic Programming (to be discussed in Section 7), the HJ-equations are also most useful in establishing (and using in practical setting) the transformation from the pair of the coordinate-momentum variables (q, p) to the so-called canonical variables for which paths of motion reduce to single points, i.e. variables for which the (re-defined) Hamiltonian is simply zero.

6.7 Legendre-Fenchel Transform *

This Section * is devoted to the Legendre-Fenchel (LF) transform, which was in fact used in its relatively simple but functional (infinite dimensional) form in Eq. (6.24). Given LF importance in variational calculus (already mentioned) and finite dimensional optimization (yet to be discussed), we have decided to allocate a special section for this important transformation and its consequences. We will also mention in the end of this Section two applications of the LF transform: (a) to solving the image restoration problem by a primal-dual algorithm, and (b) to estimating integrals with the Laplace method.

Definition 6.7.1 (Legendre-Fenchel (LF) transform)**.** Legendre-Fenchel transform of a function, $f : \mathbb{R}^n \to \mathbb{R}$, is

$$f^*(k) := \sup_{x \in \mathbb{R}^n} \left(x \cdot k - f(x) \right). \tag{6.33}$$

*This auxiliary Section can be dropped at the first reading. Material from this Section will not contribute midterm and finals.

Often LF transform also refers to as "dual" transform. Then $f^*(k)$ is dual to $f(x)$.

Example 6.7.2. Find the LF transform of the quadratic function, $f(x) = x \cdot A \cdot x/2 - b \cdot x$, where A is symmetric positive definite matrix, $A \succ 0$.
Solution: The following sequence of transformations show that the LF transform of the positively define quadratic function is another positively defined quadratic function

$$\begin{aligned} &\sup_x \quad \left(x \cdot k - \frac{1}{2} x \cdot A \cdot x + b \cdot x \right) \\ &= \quad \sup_x \left(-\frac{1}{2}(x - (k+b) \cdot A^{-1}) \cdot A \cdot (x - A^{-1}(k+b)) + \frac{1}{2}(b+k) \cdot A^{-1} \cdot (b+k) \right) \\ &= \quad \frac{1}{2}(b+k) \cdot A^{-1} \cdot (b+k), \end{aligned} \tag{6.34}$$

where the maximum is achieved at $x_* = A^{-1}(k+b)$.

Definition 6.7.3 (Convex function over $\mathbb{R}^n$)**.** A function, $u : \mathbb{R}^n \to \mathbb{R}$ is convex if

$$\forall x, y \in \mathbb{R}^n,\ \lambda \in (0,1): \quad u(\lambda x + (1-\lambda)y) \le \lambda u(x) + (1-\lambda)u(y). \tag{6.35}$$

The combination of these two notions (the Legendre-Fenchel transform and the convexity) results in the following bold statements (which we only state here, delegating proofs to Math 527).

Theorem 6.7.4 (Convexity and Involution of Legendre-Fenchel)**.** The Legendre-Fenchel transform of a convex function is convex, and it is also an involution, i.e. $(f^*)^* = f$.

6.7.1 Geometric Interpretation: Supporting Lines, Duality and Convexity

Once the formal definitions and statements are made, let us consider the one dimensional case, $n = 1$, to develop intuition about the LF and convexity. In one dimension, the LF transform has a very clear geometrical interpretation (see e.g. [10]) stated in terms of the supporting lines.

Definition 6.7.5 (Supporting Lines)**.** $f : \mathbb{R} \to \mathbb{R}$ has a supporting line at $x \in \mathbb{R}$ if

$$\forall x' \in \mathbb{R}: \quad f(x') \ge f(x) + \alpha(x' - x).$$

If the inequality is strict at all $x' \neq x$, the line is called strictly supporting.

Notice that as defined above supporting lines are defined locally, i.e. not globally for all $x \in \mathbb{R}$, but locally for a particular/fixed, x.

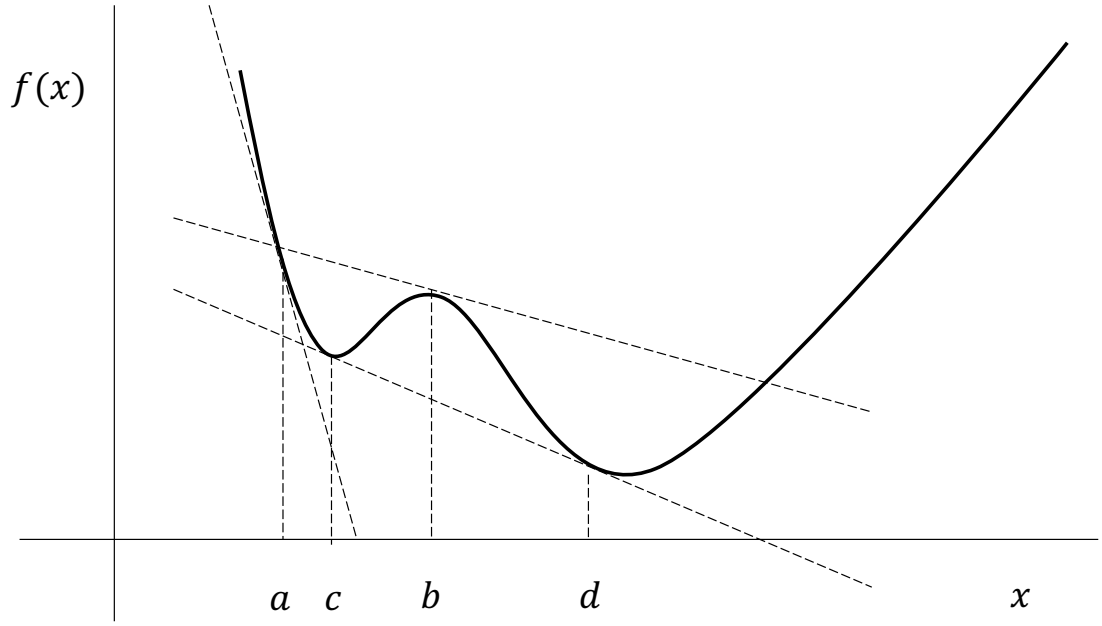


Figure 6.6: Geometric interpretation of supporting lines.

Example 6.7.6. Find $f^*(k)$ and the supporting line(s) for $f(x) = ax + b$.
Solution: Notice that we cannot draw any straight line which do not cross $f(x)$ unless they have the same slope. Therefore, $f(x)$ is the supporting line for itself. We also observe that the LF transform of the straight line is finite only at a single point $k = a$, corresponding to the slope of the line, i.e.

$$f^*(k) = \begin{cases} -b, & k = a \\ \infty, & \text{otherwise.} \end{cases}$$

Example 6.7.7. Consider the quadratic, $f(x) = ax^2/2 - bx$. Find $f^*(k)$, supporting line(s) for $f(x)$, and supporting line(s) for $f^*(k)$.
Solution: The solution, given by one dimensional version of Eq. (6.34), is $f^*(k) = (b + k)^2/(2a)$, where the maximum (in the LF transform) is achieved at $x_* = (b + k)/a$. We observe that $f^*(k)$ is well defined (finite) for all $k \in \mathbb{R}$. Denote by $f_x(y)$ the supporting line of, $f(x)$, at x. In this case of a nice (smooth and convex) $f(x)$, one derives, $f_x(x') = f(x) + f'(x)(x' - x) = ax^2/2 - bx + (ax - b)(x' - x)$, representing the Taylor series expansion of, $f(x)$, around, $x = y$, truncated at the first (linear) term. Similarly, $f^*_k(k') = f^*(k) + (f^*)'(k)(k' - k) = (b + k)^2/(2a) + (b + k)(k' - k)/a$.

What we see in this example generalizes into the following statements (given without proof):

Proposition 6.7.8. Assume that $f(x)$ admits a supporting line at x and $f'(x)$ exists at x, then the slope of the supporting line at x should be $f'(x)$, i.e. for a differentiable function the supporting line is always a tangient line.

Theorem 6.7.9. If $f(x)$ admits a supporting line at x with slope k, then $f^*(k)$ admits supporting line at k with the slope x.

Example 6.7.10. Draw supporting lines for the example of a smooth non-convex function shown in Fig. (6.6).
Solution: Sketching supporting lines for this smooth, non-convex and bounded from below example of a function with two local minima we arrive at the following observations:

- The point a admits a supporting line. The supporting line touches f at point a and the touching line is beneath the graph of $f(x)$, hence the term supporting is justified.
- The supporting line at a is strictly supporting because it touches the graph of f only at $x = a$.
- The point b does not admit a supporting line, because any line passing through $(b, f(b))$ crosses the line $f(x)$ at some other point.
- The point c admits a supporting line which is supporting, but not strictly supporting, as it touches $f(x)$ at another point, d. In this case c and d share the same supporting line.

The supporting line analysis yields a number of other useful statements listed below (without proof and only with limited discussion):

Theorem 6.7.11. $f^*(k)$ is always convex in k.

Corollary 6.7.12. $f^{**}(x)$ is always convex in x.

The last statement tells us, in particular, that f^{**} is not always convolutive, because f^{**} is always convex even for non-convex f, when $f \neq f^{**}$. This observation generalizes to

Theorem 6.7.13. $f^{**}(x) = f(x)$ iff $f(x)$ admits a supporting line at x.

The following two statements are immediate corollaries of the theorem.

Corollary 6.7.14. $f^{**} = f$ if f is convex.

Corollary 6.7.15. If $f^*(k)$ is differentiable for all k then $f^{**}(x) = f(x)$.

The following two statements are particularly useful for visualization of $f^{**}(x)$

Corollary 6.7.16. A convex function can always be written as a LF transform of another function.

Theorem 6.7.17. $f^{**}(x)$ is the largest convex function satisfying $f^{**}(x) \leq f(x)$.

Because of the last statement we call $f^{**}(x)$ the convex envelope of $f(x)$.

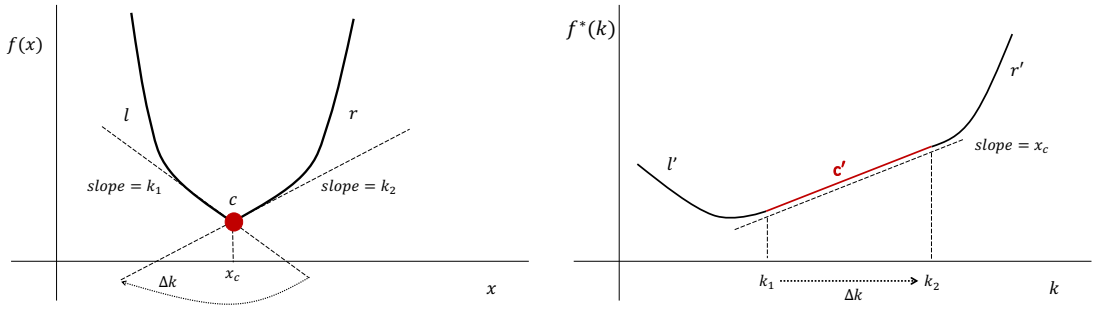


Figure 6.7: Function having a singularity cusp (left) and its LF transform (right).

Below we continue to illustrate the notion of supporting lines, as well as convexity and duality, on illustrative examples.

Example 6.7.18. Consider function containing a non-differentiable point (cusp), as shown in Fig. (6.7a). Utilizing the notion of supporting lines, draw and explain $f^*(k)$. Is $f^{**}(x) = f(x)$?

Solution: When a function has a non-differentiable point it is natural to split the analysis in two, discussing the differentiable and non-differentiable parts separately.

- (Differentiable part of $f(x)$:) Each point $(x, f(x))$ on the differentiable part of the function curve (parts a and b in Fig. (6.7a) admits a strict supporting line with slope $f'(x) = k$. These points maps under the LF transformation into $(k, f^*(k))$ points admitting supporting lines of slopes $(f^*)'(k) = x$, shown as l' and r' branches in Fig. (6.7b). Overall left (l) and right (r) branches in Fig. (6.7a) transform into left (l') and right (l') branches in Fig. (6.7b)).

- (The cusp of $f(x)$ at $x = x_c$:) The nondifferentiable point x_c admits not one but infinetily many supporting lines with slopes in the range $[k_1, k_2]$. This means that $f^*(k)$ with $k \in [k_1, k_2]$ must admit a supporting line with the constant slope x_c, shown as branch (c') in Fig. (6.7b), i.e. (c') branch is linear (affine).

The example is convex, therefore according to Corollary 6.7.14, $f^{**}(x) = f(x)$.

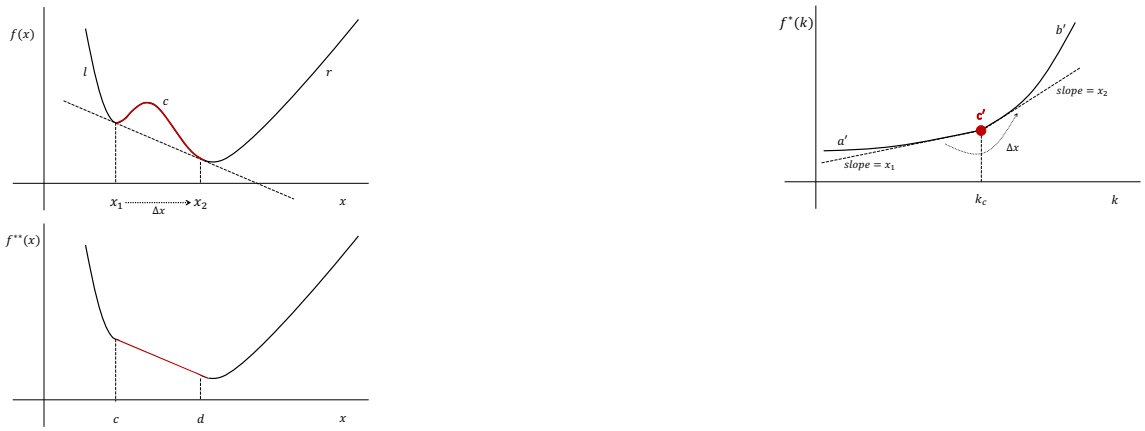


Figure 6.8: (a) An exemplary nonconvex function, $f(x)$; (b) its LT transform, $f^*(k)$; (c) its double LT transform $f^{**}(x)$.

Example 6.7.19. Show schematically $f^*(k)$ and $f^{**}(x)$ for $f(x)$ shown in Fig. (6.6).
Solution: We split curve of the function into three branches (l-left), (c-center) and (r-right), and then built LF and double-LT transform separately for each of the branches, as before relying in this construct of building supporting lines. The result in shown in Fig. (6.8) and the details are as follows.

- Branch (l) and branch (r) are strictly convex thus admitting strict supporting lines. LT transforms of the two branches are smooth. Double LF transform returns exactly the same function we have started from.

- Branch (c) is not convex and as a result none of the points within this branch, extending from x_1 to x_2, admits supporting lines. This means that the points of the branch are not represented in $f^*(k)$. We see it in Fig. (6.8b) as a collapse of the branch under the LF transform to a point. Supporting line with slope k_c connects end-points of the branch. The supporting line is not strict and it translates in $f^*(k)$ into a single $(k_c, f^*(k_c))$ point. This point of $f^*(k)$ is not differentiable. Notice that $f^*(k)$ is convex, as well as, $f^{**}(x)$. LF transformation extends $(k_c, f^*(k_c))$ into a straight line with slope k_c (shown red in Fig. (6.8c). This straight line may be thought as a convex extrapolation, envelope, of $f(x)$ in its non-convex branch.

Example 6.7.20. (a) Find the supporting lines and build the LF transform of

$$f(x) = \begin{cases} p_1 x + b_1, & x \le x_* \\ p_2 x + b_2, & x \ge x_* \end{cases}$$

where $x_* = (b_2 - b_1)/(p_1 - p_2)$, and $b_2 > b_1,\ p_2 > p_1$; and find the respective $f^{**}(x)$

(b) Suggest an example of a convex function defined on a bounded domain with diverging (infinite) slopes at the boundary. Show schematically $f^*(k)$ and $f^{**}(x)$ for the function.

6.7.2 Example of Dual Optimization in Variational Calculus

Now we are ready to return back to the image restoration problem set up in Section 6.1.3. Our task becomes to by-pass ε-smoothing discussed in Section 6.4.2 by using LF transform. This neat theoretical trick will then in developing computationally advantageous primal-dual algorithm. We will use Theorem 6.7.4 to accomplish this, transformation-to-dual, goal.

In fact, let us consider a more general set up than one discussed in Section 6.1.3. Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and consider

$$\min_{\{q(x)\}} \int_U dx \left(g(x, q(x)) + f(\nabla_x q(x))\right) \Big|_{n^T \cdot \nabla_x q = 0,\ \forall x \in \partial U}, \tag{6.36}$$

where $q : U \to \mathbb{R}$ and as before n is the normal vector to ∂U. Let us now restate the formulation in terms of the Legendre-Fenchel transform of f, thus utilizing Theorem 6.7.4:

$$\min_{\{q(x)\}} \max_{\{p(x)\}} \int_U dx \left(g(x, q(x)) + p(x) \cdot \nabla_x q(x) - f^*(p(x))\right) \Big|_{n^T \cdot \nabla_x q = n^T \cdot p = 0,\ \forall x \in \partial U}, \tag{6.37}$$

where $p : U \to \mathbb{R}^n$ and we also require that p (which is a vector) does not have a projection to the boundary, i.e. it is elongated with $\nabla_x q$. $q(x)$ is called the primal variable and $p(x)$ is called dual variable. We "dualize" only the second term in the integrand on the right hand side of Eq. (6.36) which is non-smooth, leaving the first (smooth) term unchanged. The optimization problem (6.37) is also called saddle-point formulation, due to its min-max structure.

Given the boundary conditions, we can apply integration by parts to the term in the middle in Eq. (6.37) then arriving at

$$\min_{\{q(x)\}} \max_{\{p(x)\}} \int_U dx \left(g(x, q(x)) - q(x) \nabla \cdot p(x) - f^*(p(x))\right) \Big|_{n^T \cdot p = 0,\ \forall x \in \partial U}. \tag{6.38}$$

We can attempt to solve Eq. (6.37) or Eq. (6.38) by the primal-dual method which consists in alternating minimization and maximization steps in either of the two optimizations. Implementations may be, for example, via alternating gradient descent (for minimization) and gradient ascent (for maximization).

However in the original problem we are trying to solve – the image restoration problem defined in Section 6.1.3 – we can carry over the primal-dual min-max formulation further

by exploring the structure of the argument (effective action), evaluating minimization over $\{q(x)\}$ explicitly and thus arriving at the dual formulation. This is our plan for the remainder of the section.

The case of the Total Variation image restoration corresponds to setting

$$g(x,q) = \frac{(q-f(x))^2}{2\lambda}, \quad f(w = \nabla_x q(x)) = |w|,$$

in Eq. (6.36) thus arriving at the following optimization

$$\min_q \int\limits_U dx \left(\frac{(q-f)^2}{2\lambda} + |\nabla_x q| \right) \Bigg|_{n^T \cdot \nabla_x q = 0, \forall x \in \partial U}. \tag{6.39}$$

Notice that $f(w) = |w|$ is convex and thus, according to the high-dimensional generalization of what we have learned about LF transform, $f^{**}(w) = f(w)$. The LF dual of $f(w)$ can be easily computed

$$f^*(p) = \sup_{w \in \mathbb{R}^n} (p \cdot w - |w|) = \begin{cases} 0, & |w| \leq 1 \\ \infty, & |w| > 1. \end{cases} \tag{6.40}$$

And then convexity of $f(w) = |w|$ allows us, according to Theorem 6.7.4, to "invert" Eq. (6.40)

$$f(w) = |w| = \sup_p \left(p \cdot w - \begin{cases} 0, & |p| \leq 1 \\ \infty, & |p| > 1. \end{cases} \right) = \max_{|p| \leq 1} p \cdot w. \tag{6.41}$$

Then min-max Eq. (6.38) becomes

$$\min_q \max_{|p| \leq 1} \int\limits_U dx \left(\frac{(q-f)^2}{2\lambda} - q \nabla_x \cdot p) \right) \Bigg|_{n^T \cdot p = 0, \ \forall x \in \partial U}. \tag{6.42}$$

Remarkably we can swap min and max in Eq. (6.42). This is guaranteed by the strong convexity theorem (see Appendix)

$$\max_{|p| \leq 1} \min_q \int\limits_U dx \left(\frac{(q-f)^2}{2\lambda} - q \nabla_x \cdot p \right) \Bigg|_{x \in \partial U: \ n^T \cdot p = 0}. \tag{6.43}$$

This trick is very useful because the "ultra-local" optimization over q can be done explicitly. One finds that the minimum of the quadratic over q function in the integrand of the objective in Eq. (6.43) is achieved at

$$q = f + \lambda \nabla_x \cdot p, \tag{6.44}$$

and then substituting the optimal value back in the objective we arrive at

$$\max_{|p|\leq 1} \int_U dx \left(f\nabla_x \cdot p - \frac{\lambda}{2}(\nabla_x \cdot p)^2 \right) \Bigg|_{n^T \cdot p = 0,\ \forall x \in \partial U:} . \tag{6.45}$$

which is thus the optimization dual to the primal optimization (6.39). If we are to ignore the constraint in Eq. (6.45), the objective is minimal at $\nabla \cdot p = f/\lambda$. To handle the constraint [11] has suggested to use the so-called projected gradient ascent algorithm

$$\forall x: \quad p^{k+1}(x) = \frac{p^k + \tau \nabla_x \cdot (\nabla_x \cdot p^k - f/\lambda)}{1 + \tau |\nabla_x \cdot p^k - f/\lambda|}, \tag{6.46}$$

initiated with p^0 satisfying the constraint, $|p^0| < 1$, iterating in time with step $\tau > 0$ and taking appropriate spatial discretization of the $\nabla_x \cdot$ operation on a grid with spacing Δx. Introduction of the denominator in the ratio on the right hand side of Eq. (6.46) guarantees that the condition is enforced in iterations, $|p^k| < 1$. When the iterations converge and the optimal p is found, the optimal pattern, u is reconstructed from Eq. (6.44).

6.7.3 More on Geometric Interpretation of the LF transform

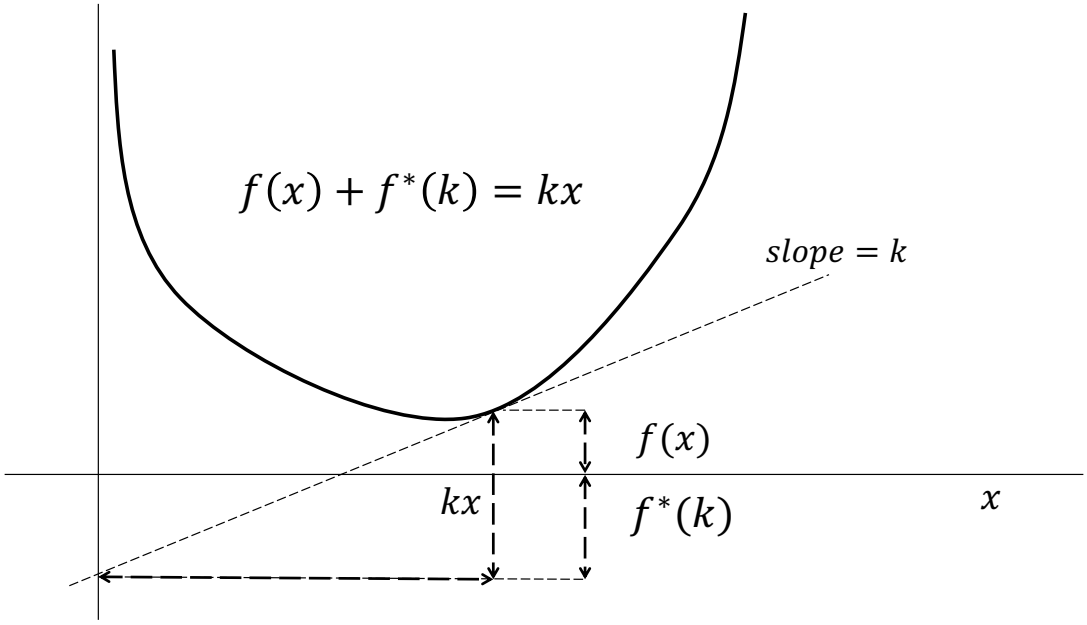


Figure 6.9: Graphic representation of the LF transform.

Here we inject some additional geometric meaning in the LF transform following [12]. We continue to draw our intuition/inspiration from a one dimensional example.

First, notice that if the function is $f : \mathbb{R} \to \mathbb{R}$ is strictly convex than $f'(x)$ is increasing, monotonically and strictly, with x. This means, in particular, that the relation between the original variable, x, and the respective optimal dual variable, k, is one-to-one, therefore providing additional explanation for the self-inverse feature of the LT transform in the case of convexity (strict convexity, to be precise, but we know that is also holds in the convex case).

Second, consider relation, illustrated in Fig. (6.9), between the original function, $f(x)$, at x allowing strict supporting line and the respective LT transform, $f^*(k)$, evaluated at $k = f'(x)$, i.e. $f^*(f'(x))$:

$$\forall x : \quad kx = f(x) + f^*(k), \qquad \text{where } k = f'(x). \tag{6.47}$$

As seen clearly in the figure the LF relation explains $f^*(k)$ as $f(x)$ extended by kx (where the latter term is associated with the supporting line). Notice remarkable symmetry of Eq. (6.47) under $x \leftrightarrow k$ and $f \leftrightarrow f^*$ transformation, also assuming that the variables, x and k, are not independent - one of the two is to be selected as tracking the change while the other (conjugated) variable will depend on the first one, according to $k = f'(x)$ or $x = (f^*)'(k)$

6.7.4 Hamiltonian-to-Lagrangian Duality in Classical Mechanics

LF transform is also the key to understanding relation between Hamiltonian and Lagrangian in classical mechanics. Let us illustrate it on a "no q"-example, i.e. on the case when the Hamiltonian, generally dependent on t, q and p depends only on p. Specifically consider example of a free relativistic particle, where $H(p) = \sqrt{p^2 + m^2}$, m is the particle mass and the speed of light is set to unity, $c = 1$. In this case, $\dot{q} = \partial_p H = dH/dp = p/\sqrt{p^2 + m^2}$, according the Hamilton equation, and the Lagrangian, which generally depends on $\dot{q}$ and q but now only depends on q, is, $L(\dot{q}) = p\dot{q} - H(p)$. This relation, rewritten in the symmetric form,

$$p\dot{q} = L(\dot{q}) + H(p),$$

should be compared with the LF relation Eq. (6.47). We observe that p and $\cdot q$, like x and k, are conjugated variables while L should be viewed as the LF transform of the Hamiltonian, $L = H^*$, or vice versa, $H = L^*$.

See [12] for further discussion of other examples of LF transform in physics, for example in statistical thermodynamics (where inverse temperature and energy are conjugated variables, while free energy is the LF dual of the entropy, and vice versa).

6.7.5 LF Transformation and Laplace Method

Consider the integral

$$F(k,n) = \int_{\mathbb{R}} dx \exp\left(n\left(kx - f(x)\right)\right).$$

When $n \to \infty$ the Laplace methods of approximating the integral (discussed in Math 583a in the fall) consists in

$$\log F(k,n) = n \sup_{x\in\mathbb{R}} \left(kx - f(x)\right) + o(n).$$

6.8 Second Variation *

Finding extrema of a function involves more than finding its critical points *. A critical point may be a minimum, a maximum or a saddle-point. To determine the critical point type one needs to compute the Hessian matrix of the function. Similar consideration applies to functionals when we want to characterize solutions of the Euler-Lagrange equations.

We naturally start the discussion of the second variation from the finite dimensional case. Let $f : U \subset \mathbb{R}^n \to \mathbb{R}$ be a $\mathbb{C}^2$ function (with existing first and second derivatives). The Hessian matrix of f at x is a symmetric bi-linear form (on the tangent vector space $\mathbb{R}^n_x$ to $\mathbb{R}^n$ at x) defined by

$$\forall \epsilon, \eta \in \mathbb{R}^n_x : \quad \mathrm{Hess}_x(\epsilon,\eta) = \left.\frac{\partial^2 f(x + s\epsilon + w\eta)}{\partial s \partial w}\right|_{s=w=0}. \tag{6.48}$$

If the Hessian is positive-definite, i.e. if the respective matrix of second-derivatives has only positive eigenvalues, then the critical point is the minimum.

Let us generalize the notion of the Hessian to the action, $\mathcal{S} = \int dt L(q,\dot{q})$ and the Lagrangian, $L(q,\dot{q})$, where $q(t) : \mathbb{R} \to \mathbb{R}^n$ is a $\mathbb{C}^2$ function. Direct generalization of Eq. (6.48)

*This auxiliary Section can be dropped at the first reading. Material from this Section will not contribute midterm and finals.

becomes

$$\begin{aligned}
\mathrm{Hess}_x\{\epsilon(t),\eta(t)\} &= \left.\frac{\partial^2\mathcal{S}\{u(t)+s\epsilon(t)+w\eta(t)\}}{\partial s\partial w}\right|_{s=w=0} \\
&= \left(\frac{\partial}{\partial w}\left(\frac{\partial\mathcal{S}\{q(t)+s\epsilon(t)+w\eta(t)\}}{\partial s}\right)_{s=0}\right)_{w=0} \\
&= \int dt\sum_{i=1}^{n}\left(\frac{\partial L(q+s\epsilon;\dot{q}+s\dot{\epsilon})}{\partial q^i}-\frac{d}{dt}\frac{\partial L(q+s\epsilon;\dot{q}+s\dot{\epsilon})}{\partial\dot{q}^i}\right)\eta^i\Bigg|_{s=0} \\
&= \int dt\sum_{i,j=1}^{n}\left(\frac{\partial^2 L}{\partial q^j\partial q^i}\epsilon^j+\frac{\partial^2 L}{\partial\dot{q}^j\partial q^i}\dot{\epsilon}^j-\frac{d}{dt}\left(\frac{\partial^2 L}{\partial q^j\partial\dot{q}^i}\epsilon^j+\frac{\partial^2 L}{\partial\dot{q}^j\partial\dot{q}^i}\dot{\epsilon}^j\right)\right)\eta^i \\
&:= \int dt\sum_{i,j=1}^{n}J_{ij}\epsilon^j\eta^i,
\end{aligned} \tag{6.49}$$

where J_{ij} is the matrix of differential operators called the Jacobi operator. To determine if the bilinear form is positive definite is usually hard, but in some simple cases the question can be resolved.

Consider, $q:\mathbb{R}\to\mathbb{R}$, $q\in\mathbb{C}^2$, and quadratic action,

$$\mathcal{S}\{q(t)\}=\int_0^T dt\left(\dot{q}^2-q^2\right). \tag{6.50}$$

with zero boundary conditions, $q(0)=q(T)=0$. To get some intuition about how the landscape of action (6.50) looks like, let us consider a subclass of functions, for example oscillatory functions consisting of only one harmonic,

$$\bar{q}(t)=a\sin\left(n\frac{\pi t}{T}\right), \tag{6.51}$$

where $a\in\mathbb{R}$ (any real) and $n\in\mathbb{Z}\setminus 0$ (any nonzero integer). Substituting Eq. (6.51) into Eq. (6.50) one derives,

$$\begin{aligned}
\mathcal{S}\{\bar{q}(t)\} &= \frac{n^2\pi^2a^2}{T^2}\left(\int_0^T dt\cos^2\left(\frac{n\pi t}{T}\right)\right) \\
&\quad -a^2\left(\int_0^T dt\sin^2\left(\frac{n\pi t}{T}\right)\right)=\frac{Ta^2}{2}\left(\frac{n^2\pi^2}{T^2}-1\right).
\end{aligned}$$

One observes that at $T<\pi$, the action, $\mathcal{S}$, considered on this special class of functions, is positive. However, when some of these probe functions will result in a negative action when $T>\pi$. This means that at $T>\pi$, the functional quadratic form, correspondent to the action (6.50), is certainly not positive definite.

One thus came out of this "probe function" exercise with the following question: can it be that the functional quadratic form, correspondent to the action (6.50), is not positive definite? The analysis so far (restricted to the class of single harmonic test functions) is not conclusive. Quite remarkably one can prove that the action (6.50) is always positive (over the class of zero boundary condition, twice differentiable functions), and thus the respective quadratic form is always positive definite, if $T < \pi$.

Example 6.8.1. Prove that the action $\mathcal{S}\{q(t)\}$ given by Eq. (6.50) is positive at, $T < \pi$, for any twice differentiable function, $q \in \mathbb{C}^2$ with zero boundary conditions, $q(0) = q(T) = 0$. (Hint: Represent the function as Fourier Series and show that the action is a sum of squares.)

Solution. Consider the general Fourier series expansion of $q(t)$, that is

$$q(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n \cos\left(\frac{2\pi n t}{T}\right) + b_n \sin\left(\frac{2\pi n t}{T}\right) \right].$$

Calculating $\int_0^T (\dot{q}^2 - q^2) dt$ and noting the orthogonality of the terms in the expansion, one arrives at

$$\int_0^T (\dot{q}^2 - q^2) dt = -T\frac{a_0^2}{4} + T\sum_{n=1}^{\infty} \frac{a_n^2 + b_n^2}{2} \left(\frac{4\pi^2 n^2}{T^2} - 1\right).$$

If we consider the 'worst-case' scenario of $T = \pi$ we then have to show that

$$\sum_{n=1}^{\infty} \frac{a_n^2 + b_n^2}{2} \left(4n^2 - 1\right) \geq \frac{a_0^2}{4},$$

to ensure positivity of the action. Note that from the boundary conditions we have that, $a_0/2 + \sum_n a_n = 0$. Without loss of generality let us scale $q(t)$ such that $a_0 = -2$. Then it will suffice for us to show that

$$\sum_{n=1}^{\infty} \frac{a_n^2}{2} \left(4n^2 - 1\right) \geq 1.$$

We can do this by constructing the dual problem and demonstrating that the minimal value of the left-hand-side is 1 when varying the a_n. Specifically, our problem is

$$\min_{a=(a_1,\ldots,a_k)} \sum_{n=1}^{k} \frac{a_n^2}{2} \left(4n^2 - 1\right),$$
$$\text{s.t.} \sum_{n=1}^{k} a_n = 1,$$

where we first consider the partial sum case and then take the $k \to \infty$ limit. The Lagrangian is

$$\mathcal{L}(a, \mu) = \sum_{n=1}^{k} \frac{a_n^2}{2} \left(4n^2 - 1\right) - \mu \left(\sum_{n=1}^{k} a_n - 1 \right).$$

We compute $\nabla_a \mathcal{L} = 0$ to get

$$a_n(4n^2 - 1) - \mu = 0, \qquad \forall n = 1, \ldots, k.$$

This implies $a_n = \mu/(4n^2 - 1)$. If the equality constraint is enforced we derive

$$\sum_{n=1}^{k} \frac{\mu}{4n^2 - 1} = 1.$$

One can compactly represent the partial sums as

$$\sum_{n=1}^{k} \frac{\mu}{4n^2 - 1} = \frac{\mu k}{2k + 1}$$

which yields $\mu = (2k + 1)/k$. Substituting this back into the original objective function we arrive at

$$\sum_{n=1}^{k} \frac{a_n^2}{2} \left(4n^2 - 1\right) = \frac{2k + 1}{2k} \geq 1.$$

Therefore in the $k \to \infty$ limit we are still larger than 1, thus proving positivity of the action. □

6.9 Methods of Lagrange Multipliers

So far we have only discussed unconstrained variational formulations. This Section is devoted to generalizations where variational problems with constraints are formulated and resolved.

6.9.1 Functional Constraint(s)

Consider the shortest path problem discussed in Example 6.2.3, however constrained by the area, A as follows

$$\min_{\{q(x)|x\in[0,a]\}} \int_0^a dx\sqrt{1 + (q'(x))^2}dx \Bigg|_{q(0)=0,\ q(a)=b,\ \int_0^a q(x)dx=A}.$$

The area constraint can be built in the optimization by adding,

$$\lambda \left(\int_0^a dxq(x)dx - A \right),$$

to the optimization objective, where λ is the Lagrangian multiplier. The Euler-Lagrange equations for this "extended" action are

$$\begin{aligned} 0 &= \nabla_x \left(L_{\nabla q} \left(x, q(x), \nabla_x q(x) \right) \right) - L_q \left(x, q(x), \nabla_x q(x) \right) - \lambda \\ &= \frac{d}{dx} \frac{q'(x)}{\sqrt{1 + (q'(x))^2}} - 0 - \lambda \\ &\rightarrow \frac{q'(x)}{\sqrt{1 + (q'(x))^2}} = \text{constant} + \lambda x \end{aligned}$$

Example 6.9.1. The principle of maximum entropy, also called principle of the maximum likelihood (distribution), selects the probability distribution that maximizes the entropy, $\mathcal{S} = -\int_D dx P(x) \log P(x)$, under normalization condition, $\int_D dx P(x) = 1$.

- (a) Consider $D \in \mathbb{R}^n$. Find optimal $P(x)$.
- (b) Consider $D = [a, b] \subset \mathbb{R}$. Find optimal $P(x)$, assuming that the mean of x is known, $\mathbb{E}_{\{P(x)\}}(x) := \int_D dx x P(x) = \mu$.

Solution:
(a) The effective action is,

$$\tilde{\mathcal{S}} = \mathcal{S} + \lambda \left(1 - \int_D dx P(x) \right),$$

where λ is the (constant, i.e. not dependent on x, Lagrangian multiplier. Variation of $\tilde{\mathcal{S}}$ over $P(x)$ results in the following EL equation

$$\frac{\delta \tilde{S}}{\delta P(x)} = 0 : \quad -\log(P(x)) - 1 - \lambda = 0.$$

Accounting for the normalization condition one finds that the optimum is achieved at the equ-distribution:

$$P(x) = \frac{1}{\|D\|},$$

where $\|D\|$ is the size of D.
(b) The effective action is,

$$\tilde{\mathcal{S}} = \mathcal{S} + \lambda \left(1 - \int_D dx P(x) \right) + \lambda_1 \left(\mu - \int_D dx x P(x) \right),$$

where λ and λ_1 are two (constant) Lagrangian multipliers. Variation of $\tilde{\mathcal{S}}$ over $P(x)$ results in the following EL equation

$$\frac{\delta \tilde{S}}{\delta P(x)} = 0 : \quad -\log(P(x)) - 1 - \lambda - \lambda_1 x = 0 \ \rightarrow \ P(x) = e^{-1-\lambda} \exp(-\lambda_1 x).$$

λ and λ_1 are constants which can be expressed via a, b and μ resolving the normalization constraint and the constraint on the mean,

$$e^{-1-\lambda}\left(-\frac{\exp(-\lambda_1 x)}{\lambda_1}\right)\Bigg|_a^b = 1, \quad e^{-1-\lambda}\left(-\frac{x\exp(-\lambda_1 x)}{\lambda_1} - \frac{\exp(-\lambda_1 x)}{\lambda_1^2}\right)\Bigg|_a^b = \mu.$$

Exercise 6.5. Consider the setting of Example 6.9.1b with $a = -\infty$, $b = \infty$. Assuming that mean and variance of the probability distribution are known, i.e. $\mathbb{E}_{\{P(x)\}}(x) = \mu$ and $\mathbb{E}_{\{P(x)\}}(x^2) = \sigma^2 + \mu^2$, find $P(x)$ which maximizes the entropy.

6.9.2 Function Constraints

The method of Lagrange multipliers in the calculus of variations extends to other types of constrained optimizations, where the condition is not a functional as in the cases discussed so far but a function. Consider, for example, our standard one-dimensional example of the action functional,

$$\mathcal{S}\{q(t)\} = \int dt L(t; q(t); \dot{q}(t)), \tag{6.52}$$

over $q : \mathbb{R} \to \mathbb{R}$, however constrained by the functional,

$$\forall t : \quad G(t; q(t); \dot{q}(t)) = 0. \tag{6.53}$$

Let us also assume that $L(t; q; \dot{q})$ and $G(t; q; \dot{q})$ are sufficiently smooth functions of their last argument, $\dot{q}$. The idea then becomes to introduce the following "modified" action

$$\tilde{\mathcal{S}}\{q(t), \lambda(t)\} = \int dt \left(L(t; q(t); \dot{q}(t)) - \lambda(t) G(t; q(t); \dot{q}(t))\right), \tag{6.54}$$

which is now a functional of both $q(t)$ and $\lambda(t)$, and extremize it over both $q(t)$ and $\lambda(t)$. One can show that solutions of the EL equations, derived as variations of the action (6.54) over both $q(t)$ and $\lambda(t)$, will give a sufficient condition for the minimum of Eq. (6.52) constrained by Eq. (6.53).

Let us illustrate this scheme and derive the Euler-Lagrange equation for a Lagrangian $L(q; \dot{q}; \ddot{q})$ which depends on the second derivative of a $\mathbb{C}^3$ function, $q : \mathbb{R} \to \mathbb{R}$ and does not depend on t explicitly. In full analogy with Eq. (6.54) the modified action in this case becomes

$$\tilde{\mathcal{S}}\{q(t), \lambda(t)\} = \int dt \left(L(q(t); \dot{q}; \dot{v}) - \lambda(t)\left(v(t) - \dot{q}(t)\right)\right). \tag{6.55}$$

Notice Then the modified Euler-Lagrange equations are

$$\frac{\partial L}{\partial q} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}} + \lambda\right), \quad -\lambda = \frac{d}{dt}\frac{\partial L}{\partial \dot{v}}, \quad v = \dot{q}. \tag{6.56}$$

Eliminating λ and v one arrives at the desired modified EL equations stated solely in terms of derivatives of the Lagrangian over $q(t)$ and its derivatives:

$$\frac{\partial L}{\partial q} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}} + \frac{d^2}{dt^2}\frac{\partial L}{\partial \ddot{q}} = 0. \tag{6.57}$$

Exercise 6.6. Find extrema of $\mathcal{S}\{q(t)\} = \int_0^1 dt \|\dot{q}(t)\|$ for $q : [0,1] \to \mathbb{R}^3$ subject to the norm constraint, $\forall t : \quad \|q(t)\|^2 = 1$, and with generic boundary conditions, i.e. $q(0) = q_0$ and $q(1) = q_1$, where both boundary points satisfy the norm constraint.

We will see more of the calculus of variations with (function) constraints later in the optimal control section of the course.

Chapter 7

Optimal Control and Dynamic Programming

Optimal control problem shall be considered as a special case of a general variational calculus problem, where the (vector) fields evolve in time, i.e. reside in one dimensional real space equipped with a direction, and constrained by a system of ODEs, possibly with algrebraic constraints added too. We will learn how to analyze the problems by the methods of the variational calculus from Section 6, using optimization approaches, e.g. convex analysis and duality, described in Appendix A.1, and also adding to arsenal of tools a new one called "Dynamic Programming" (DP) in Section 7.4.

Let us start with an illustrative (as sufficiently simple) optimal control problem.

Example 7.0.1. Consider trajectory of a particle in one dimension: $\{q(\tau) : \ [0,t] \to \mathbb{R}\}$ which is subject to control $\{u(\tau) : \ [0,t] \to \mathbb{R}\}$. Specifically, the control means that at any moment in time, the velocity of the particle can be set to any value less than or equal to one. That is $\dot{q}(t) = u(t)$ where $u(t) \leq 1$. Solve the following constrained problem of the variational calculus type:

$$\min_{\{u(\tau),q(\tau)\}} \int_0^t d\tau (q(\tau))^2 \Bigg|_{\tau\in(0,t]:\ \dot{q}(\tau)=u(\tau),\ u(\tau)\leq 1} \tag{7.1}$$

where $t > 0$ and the initial position, $q(0) = q_0$, is known (fixed).

Solution. If $q_0 > 0$, one can guess the optimal solution right away: jump to $q = 0$ immediately (at $\tau = 0^+$) and then stay zero. To justify the solution, one first drops all the constraints in Eq. (7.1), observe that the minimal solution of the unconstrained problem is, $\tau \in (0,t] : \ q(\tau) = u(\tau) = 0$, and then verify that constraints dropped are satisfied. (Notice

that the resulting discontinuity of the optimal $q(\tau)$ at $\tau = 0$ is not a problem, as it was not required in the problem formulation.)

The analysis in the case of $q_0 \leq 0$ is more elaborate. Let us exclude the control variable, turning the pair of constraints in Eq. (7.1) into one, $\forall \tau : \ \dot{q} \leq 1$. Our next step will be to use the Duality Theory and KKT conditions, discussed in Math 584 (see also Appendix A) in the context of finite dimensional optimization. We can therefore view what follows as a practical lesson (just a use case, no proofs) on how to extend this important approach from the case of finite dimensional optimization to the variational calculus.

We introduce the Lagrangian function,

$$L(q(\tau), \mu(\tau)) = q^2 + \mu(\dot{q} - 1),$$

and then write the following four KKT conditions:

1. KKT-1: Primal Feasiblity: $\dot{q}(\tau) \leq 1$ for $\tau \in (0, t]$.
2. KKT-2: Dual Feasibility: $\mu(t) \geq 0$ for $\tau \in (0, t]$.
3. KKT-3: Stationary point in primal variables - which is simply the Euler-Lagrange condition of the variational calculus: $2q = \dot{\mu}$ for $\tau \in (0, t]$.
4. KKT-4: Complementary Slackness: $\mu(t)(\dot{q}(t) - 1) = 0$ for $\tau \in (0, t]$.

We find that,

$$q(\tau) = \tau + q_0, \quad \mu(\tau) = \tau^2 + 2q_0\tau + c, \tag{7.2}$$

where c is a constant, satisfy both the KKT conditions and the initial condition, $q(0) = q_0$. Can we have another solution different from Eqs. (7.2) but satisfying the KKT conditions? How about a discontinuous control? Consider the following probe functions, bringing q to zero first with the maximal allowed control, and then switching off the control:

$$q(\tau) = \begin{cases} q_0 + \tau, & 0 < \tau \leq -q_0 \\ 0, & -q_0 < \tau \leq t \end{cases}, \quad \mu(\tau) = \begin{cases} \tau^2 + 2q_0\tau + q_0^2, & 0 < \tau \leq -q_0 \\ 0, & -q_0 < \tau \leq t \end{cases}. \tag{7.3}$$

We observe that, indeed, in the regime where the probe function is well defined, i.e. $0 < -q_0 < t$, Eqs. (7.3) solves the KKT conditions (7.3), therefore providing an alternative to the solution (7.2). Comparing objectives in Eq. (7.1) for the two alternatives one finds that at, $0 < -q_0 < t$, the solution (7.3) is optimal while the solution (7.2) is optimal if $t < -q_0$. □

Exercise 7.1. Solve Example 7.0.1 with the condition $u \leq 1$ replaced by $|u| \leq 1$.

7.1 Linear Quadratic Control via Calculus of Variations *

Our next extra-curricular topic * is Linear Quadratic (LQ) control. Consider d-dimensional real vector representing evolution of the system state in time, $\{q(\tau) \in \mathbb{R}^d | \tau \in [0,t]\}$, governed by the following system of linear ODEs

$$\forall \tau \in (0,t]: \quad \dot{q}(\tau) = Aq(\tau) + Bu(\tau), \quad q(0) = q_0, \tag{7.4}$$

where A and B are constant (time independent) square, nonsingular (invertible) and possibly asymmetric, thus $A \neq A^T$ and $B \neq B^T$, real matrices, $A, B \in \mathbb{R}^d \times \mathbb{R}^d$, and $\{u(\tau) \in \mathbb{R}^d | \tau \in [0,t]\}$ is a time-dependent control vector of the same dimensionality as q. Introduce a combined action, often called *cost-to-go*:

$$\mathcal{S}\{q(\tau), u(\tau)\} := \mathcal{S}_{eff}\{u(\tau)\} + \mathcal{S}_{des}\{q(\tau)\} + \mathcal{S}_{fin}(q_t), \tag{7.5}$$

$$\mathcal{S}_{eff}\{u(\tau)\} := \frac{1}{2}\int_0^t d\tau u^T(\tau) R u(\tau), \tag{7.6}$$

$$\mathcal{S}_{des}\{q(\tau)\} := \frac{1}{2}\int_0^t d\tau q^T(\tau) Q q(\tau), \tag{7.7}$$

$$\mathcal{S}_{fin}(q(t)) := \frac{1}{2} q^T(t) Q_{fin} q(t), \tag{7.8}$$

where $\mathcal{S}_{eff}$, dependent only on $\{u(\tau)\}$, represents required *efforts* of control; $\mathcal{S}_{des}$, dependent only on $\{q(\tau)\}$, expresses the cost of maintaining *desired* state of the system $\{q(t)\}$ proper; and $\mathcal{S}_{fin}$, dependent only on $q(t)$, expresses the cost of achieving the *final* state, $q(t)$. We assume that R, Q and Q_{fin} are symmetric real positive definite matrices. We aim to optimize the *cost-to-go* over $\{q(\tau)\}$ and $\{u(\tau)\}$ constrained by the governing ODEs and respective initial condition in Eqs. (7.4).

As custom in the variational calculus with function constraints, let us extend the action (7.5) with a Lagrangian multiplier function associated with the ODE constraints (7.4) and then formulate necessary conditions for the optimality stated as an unconstrained variation of the following effective action

$$\mathcal{S}\{q, u, \lambda\} := \mathcal{S}\{q, u\} + \int_0^t d\tau \lambda^T(\tau) \left(-\dot{q} + Aq + Bu\right), \tag{7.9}$$

where $\{\lambda(\tau)\}$ is the time-dependent vector of the Lagrangian multipliers, also called the adjoint vector. Euler-Lagrange (EL) equations and the primal feasibility equations following

*This auxiliary Section can be dropped at the first reading. Material from this Section will not contribute midterm and finals.

from variations of the effective action (7.9) over q, u and λ are

$$\text{Euler-Lagrange}: \quad \tfrac{\delta\mathcal{S}\{q,u,\lambda\}}{\delta q} = 0: \quad \forall\tau\in(0,t]: \; Qq + \dot\lambda + A^T\lambda = 0, \tag{7.10}$$

$$\tfrac{\delta\mathcal{S}\{q,u,\lambda\}}{\delta u} = 0: \quad \forall\tau\in[0,t]: \; Ru + B^T\lambda = 0, \tag{7.11}$$

$$\text{primal feasibility}: \quad \tfrac{\delta\mathcal{S}\{q,u,\lambda\}}{\delta \lambda} = 0: \quad \text{Eqs. (7.4)}. \tag{7.12}$$

The equations should also be complemented with the boundary condition,

$$\text{boundary condition at } \tau = t, \quad \frac{\partial\mathcal{S}\{q,u,\lambda\}}{\partial q(t)} = 0 : \lambda(t) = Q_{fin}q(t), \tag{7.13}$$

derived by variations of the effective action over q at the final point, $q(t)$. The simplest way to derive the boundary condition Eq. (7.13) is through discretization: turning temporal integrals into discrete sums, specifically

$$\int_0^t d\tau\lambda^T(\tau)\dot q(\tau) \to \lambda^T(\Delta)(q(\Delta)-q(0)+\cdots+\lambda^T(t)(q(t-\Delta))-q(t)), \tag{7.14}$$

where Δ is the discretization step, and then looking for a stationary point over $q(t)$.

Notice that alternatively the boundary conditions can be derived from the "main theorem of classical mechanics – Theorem 6.5.1 – proving that the gradient of S with respect to the terminal point (t_1, q_1) is $\nabla S = \langle\, L - \dot q L_{\dot q}\,,\, L_{\dot q}\,\rangle|_{(t_1,q_1)}$. (See Theorem 6.5.1). This result suggests that given that $q(\tau)$ is optimal, the action cannot be improved by variations in the terminal value, $q(t)$ and therefore the boundary condition at $\tau = t$ is:

$$\text{boundary condition at } \tau = t, \quad 0 = \frac{\partial\mathcal{S}\{q,u,\lambda\}}{\partial q(t)} = \lambda(t) + L_{\dot q}|_{\tau=t} = \lambda(t) - Q_{fin}q(t), \tag{7.15}$$

Observe that Eqs. (7.11) are algebraic, thus allowing to express the control vector, u, via the adjoint vector, λ

$$u = -R^{-1}B^T\lambda. \tag{7.16}$$

Substituting it into Eqs. (7.10,7.12) one arrives at the following joint system of the original and adjoint equations

$$\begin{pmatrix} \dot q \\ \dot\lambda \end{pmatrix} = \begin{pmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{pmatrix}\begin{pmatrix} q \\ \lambda \end{pmatrix}, \quad \begin{pmatrix} q(0) \\ \lambda(t) \end{pmatrix} = \begin{pmatrix} q_0 \\ Q_{fin}q(t) \end{pmatrix}. \tag{7.17}$$

The system of ODEs (7.17) is a two-point Boundary Value Problem (BVP) because it has two boundary conditions at the opposite ends of the time interval. In general, two-point BVPs are solved by the shooting method, which requires multiple iterations forward and backward in time (hoping for convergence). However for the LQ Control problems,

the system of equations is linear, and we can solve it in one shot – with only one forward iteration and one backward iteration. Indeed, integrating the linear ODEs (7.17) one derives

$$\begin{pmatrix} q(\tau) \\ \lambda(\tau) \end{pmatrix} = W(\tau) \begin{pmatrix} q(0) \\ \lambda(0) \end{pmatrix}, \tag{7.18}$$

$$W(\tau) = \begin{pmatrix} W^{1,1}(\tau) & W^{1,2}(\tau) \\ W^{2,1}(\tau) & W^{2,2}(\tau) \end{pmatrix} := \exp\left(\tau \begin{pmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{pmatrix} \right), \tag{7.19}$$

which allows to express $\lambda(0)$ via $q(0) = q_0$

$$\lambda(0) = Mq_0, \quad M := -\left(W^{2,2}(t) + Q_{fin}W^{1,2}(t)\right)^{-1}\left(W^{2,1}(t) + Q_{fin}W^{1,1}(t)\right). \tag{7.20}$$

Substituting Eqs. (7.18,7.20) into Eq. (7.16) one arrives at the following expression for the optimal control via q_0

$$u(\tau) = -R^{-1}B^T\left(W^{2,1}(\tau) + W^{2,2}(\tau)M\right)q_0. \tag{7.21}$$

A control of this type, dependent on the initial state, is called *open loop* control. That is, the control policy $u(\tau)$ doesn't explicitly depend on the current state $q(\tau)$, and instead only on the initial state q_0. While in ideal conditions this does not pose an issue, under uncertainty it is often better for the control policy to depend on the current state of the system $q(\tau)$. Such a control scheme is called *feedback loop* control, which may also be called the *closed loop* control. The feedback loop version of Eq. (7.21) requires us to express $u(\tau)$ in terms of $q(\tau)$. To do this we seek a map from $q(\tau)$ to $\lambda(\tau)$, which for LQ control is a matrix $P(\tau)$ such that $P(\tau)q(\tau) = \lambda(\tau)$. This allows us to write

$$u(\tau) = -R^{-1}B^T\lambda(\tau) = -R^{-1}B^TP(\tau)q(\tau).$$

We derive this matrix by expressing $\lambda(\tau)$ and $q(\tau)$ via q_0 according to Eq. (7.18,7.20) and then substituting the result in Eq. (7.16):

$$\begin{aligned} \forall\tau \in (0,t]: \quad u(\tau) &= -R^{-1}B^T\lambda(\tau) = -R^{-1}B^T\left(W^{2,1}(\tau) + W^{2,2}(\tau)M\right)q_0 \\ &= -R^{-1}B^T\left(W^{2,1}(\tau) + W^{2,2}(\tau)M\right)\left(W^{1,1}(\tau) + W^{1,2}(\tau)M\right)^{-1}q(\tau) \\ &= -R^{-1}B^TP(\tau)q(\tau), \end{aligned} \tag{7.22}$$

$$P(\tau) := \left(W^{2,1}(\tau) + W^{2,2}(\tau)M\right)\left(W^{1,1}(\tau) + W^{1,2}(\tau)M\right)^{-1}. \tag{7.23}$$

At any moment of time τ, i.e. as we go along, the feedback loop control responds to the current measurement of the system state, $q(\tau)$, at the same time, τ.

Notice that in the deterministic case without uncertainty/perturbation (and this is what we have considered so far) the open loop and the feedback loop are equivalent. However,

the two control schemes/policies give very different results in the presence of uncertainty/perturbation (consistently with what was already mentioned above). We will investigate this phenomenon and have a more extended comparison of the two controls in the probability/statistics/data science section of the course

Example 7.1.1. Show, utilizing derivations and discussions above, that the matrix, $P(\tau)$, defined in Eq. (7.23), satisfies the so-called Riccati equations:

$$\dot{P} + A^T P + PA + Q = PBR^{-1}B^T P, \tag{7.24}$$

supplemented with the terminal/final ($\tau = t$) condition, $P(t) = Q_{fin}$.

Solution. To solve this problem we will not actually be using the explicit expression for P given in (7.23). Instead we will use the important relation $Pq = \lambda$. First, we note that

$$\dot{q} = Aq - BR^{-1}B^T\lambda, \quad \text{and} \quad \dot{\lambda} = -Qq - A^T\lambda.$$

With these three relations we may write

$$\begin{aligned}
&\dot{P}q + P\dot{q} = \dot{\lambda},\\
\implies \quad &\dot{P}q + P(Aq - BR^{-1}B^T\lambda) = -Qq - A^T\lambda,\\
\implies \quad &\dot{P}q + P(Aq - BR^{-1}B^T Pq) = -Qq - A^T Pq,\\
\implies \quad &\dot{P}q + A^T Pq + PAq + Qq = PBR^{-1}B^T Pq,\\
\implies \quad &(\dot{P} + A^T P + PA + Q)q = PBR^{-1}B^T Pq.
\end{aligned}$$

Since this holds for arbitrary q we must have that

$$\dot{P} + A^T P + PA + Q = PBR^{-1}B^T P.$$

Example 7.1.2. Consider an unstable one dimensional process

$$\tau \in [0, \infty[: \quad \dot{q}(\tau) = Aq(\tau) + u(\tau),$$

where $u \in \mathbb{R}$ and A is a positive constant, $A > 0$. Design an LQ controller $u(\tau) = -Pq(\tau)/R$ that minimizes the action

$$\mathcal{S}\{q(\tau), u(\tau)\} = \int_0^\infty d\tau \left(q^2 + Ru^2\right),$$

where P is a constant (need to find) and R is a positive known constant. Discuss/explain what happens with P when $R \to 0$ or $R \to \infty$.

Solution. We note that since P is a constant, $\dot{P} = 0$. Moreover, since we are in one dimension the Riccati Eq. (7.24) can be simplified to

$$2AP + Q = \frac{B^2}{R}P^2.$$

Since $B = Q = 1$ the quadratic form for P becomes

$$P = RA \pm \sqrt{R^2A^2 + R},$$

that is it shows two branches. We must consider both of these branches in the context of the optimization problem. Substituting u in the ODE, we arrive at

$$q(\tau) = \exp\left((A - P/R)\tau\right) = \exp\left(\left(\mp\sqrt{A^2 + 1/R}\right)\tau\right).$$

Observe that the cost is infinite if we do not take $P = RA + \sqrt{R^2A^2 + R}$. If we consider $R \to 0$ it results in $P \to \sqrt{R}$. However, if we consider $R \to \infty$ the result is $P \to 2RA$.

We can even write down the action explicitly:

$$\begin{aligned}
\mathcal{S} &= \int_0^\infty (q^2 + Ru^2)d\tau \\
&= \int_0^\infty (1 + 2A^2 + 1/R - 2\sqrt{A^2 + 1/R}) \exp\left(-2\tau\sqrt{A^2 + 1/R}\right) d\tau \\
&= \frac{1 + 2A^2 + 1/R - 2\sqrt{A^2 + 1/R}}{2\sqrt{A^2 + 1/R}}.
\end{aligned}$$

7.2 From Variational Calculus to Bellman-Hamilton-Jacobi Equation

Next we consider optimal control problem which is more general, in terms of governing equations and optimization objective, than what was considered so far. We study controlled dynamical system, which is nonlinear in our primal variable, $\{q(\tau) : [0, t] \to \mathbb{R}^d\}$, but still linear in the control variable, $\{u(\tau) : [0, t] \to \mathbb{R}^d\}$

$$\forall \tau \in [0, t] : \quad \dot{q}(\tau) = f(q(\tau)) + u(\tau). \tag{7.25}$$

As above, we will formulate a control problem as an optimization. We aim to minimize the objective

$$\int_0^t d\tau \left(\frac{1}{2}u^T(\tau)u(\tau) + V(q(\tau))\right), \tag{7.26}$$

over $\{u(\tau)\}$ which satisfies the ODE (7.25). Here in Eq. (7.26) we shortcut notations and use $(u(\tau))^2$ for $u^T(\tau)u(\tau)$. Notice that the cost-to-go objective (7.26) is a sum of two terms: (a) the cost of control, which is assumed quadratic in the control efforts, and (b) the bounded from below "potential", which defines preferences or penalties imposed on where the particle may or may not go. The potential may be soft or hard. An exemplary soft potential is the quadratic potential

$$V(q) = \frac{1}{2} q^T \Lambda q = \frac{1}{2} \sum_{i=1}^{d} q_i \Lambda_{ij} q_j, \tag{7.27}$$

where Λ is a positive semi-definite matrix. This potential encourages $q(\tau)$ to stay close to the origin, $q = 0$, penalizing (but softly) for deviation from the origin. An exemplary hard constraint may be

$$V(q) = \begin{cases} 0, & |q| < a \\ \infty, & |q| \geq a \end{cases}, \tag{7.28}$$

completely prohibiting $q(\tau)$ to leave the ball of size a around the origin. Summarizing, we discuss the optimal control problem:

$$\min_{\{u(\tau),q(\tau)\}} \int_0^t d\tau \left(\frac{u^T(\tau)u(\tau)}{2} + V(q(\tau)) \right) \Bigg| \begin{array}{l} \forall \tau \in [0,t]: \quad \dot{q}(\tau) = f(q(\tau)) + u(\tau) \\ q(0) = q_0, \quad q(t) = q_t \end{array} \tag{7.29}$$

where initial and final states of the system are assumed fixed.

In the following we restate Eq. (7.29) as an unconstrained variational calculus problem. (Notice, that we do not count the boundary conditions as constraints.) We will assume that all the functions involved in the formulation (7.29) are sufficiently smooth and derive respective Euler-Lagrange (EL) equations, Hamiltonian equations and Hamilton-Jacobi (HJ) equations.

To implement the plan, let us, first of all, exclude $\{u(\tau)\}$ from Eq. (7.29). The resulting "q-only" formulation becomes

$$\min_{\{q(\tau)\}} \int_0^t d\tau \left(\frac{\left(\dot{q}(\tau) - f(q(\tau))\right)^T \left(\dot{q}(\tau) - f(q(\tau))\right)}{2} + V(q(\tau)) \right) \Bigg|_{q(0)=q_0, \quad q(t)=q_t}. \tag{7.30}$$

Following Lagrangian and Hamiltonian approaches, described in details in the variational calculus portion of the course, see Section 6, one identifies action, Lagrangian, momentum

and Hamiltonian for the functional optimization (7.30) as follows

$$\begin{aligned} S\{q(\tau),\dot{q}(\tau)\} &= \int_0^t d\tau \frac{(\dot{q}-f(q))^T(\dot{q}-f(q))}{2}+V(q), && (7.31)\\ L &= \frac{(\dot{q}-f(q))^T(\dot{q}-f(q))}{2}+V(q), && (7.32)\\ p &\equiv \frac{\partial L}{\partial \dot{q}^T}=\dot{q}-f(q), && (7.33)\\ H &\equiv \dot{q}^T\frac{\partial L}{\partial \dot{q}^T}-L=\frac{\dot{q}^T\dot{q}}{2}-\frac{(f(q))^Tf(q)}{2}-V(q)\\ &= \frac{p^Tp}{2}+p^Tf(q)-V(q). && (7.34) \end{aligned}$$

Then the Euler-Lagrange equations are

$$\begin{aligned} \forall i=1,\cdots,d: \quad & \frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i}=\frac{\partial L}{\partial q_i} && (7.35)\\ & \frac{d}{dt}\left(\dot{q}_i-f_i(q)\right)=-\sum_{j=1}^{d}\left(\dot{q}-f(q)\right)_j\partial_{q_i}f_j(q)+\partial_{q_i}V(q), \end{aligned}$$

where we stated the vector equation by components for clarity. The Hamilton equations are

$$\begin{aligned} \forall i=1,\cdots,d:\dot{q}_i &= \frac{\partial H}{\partial p_i}=p_i+f_i(q), && (7.36)\\ \dot{p}_i &= -\frac{\partial H}{\partial q_i}=-p_i\nabla_{q_i}f(q)+\nabla_{q_i}V(q). && (7.37) \end{aligned}$$

Considering the action, $\mathcal{S}$, as a function (not a functional!) of the final time, t, and of the final position, q_t, and recalling that,

$$\frac{\partial S}{\partial t}=-H|_{\tau=t}\,,\quad \frac{\partial S}{\partial q_t}=\left.\frac{\partial L}{\partial \dot{q}}\right|_{\tau=t}=p|_{\tau=t}\,,$$

one arrives at the Hamilton-Jacobi (HJ) equations

$$\frac{\partial S}{\partial t}=-H|_{\tau=t}=-H\left(q_t,\frac{\partial S}{\partial q_t}\right)=-\frac{1}{2}\left(\frac{\partial S}{\partial q_t}\right)^T\left(\frac{\partial S}{\partial q_t}\right)-\left(\frac{\partial S}{\partial q_t}\right)^T f(q_t)+V(q_t). \quad (7.38)$$

We will see later on that it may be useful to consider the HJ equation backwards in time. In this case we consider the action, $\mathcal{S}=\int_\tau^t d\tau' L$, as the function of τ and $q(\tau)=q$. This results in the following (backwards in time) modification of Eq. (7.38)

$$-\frac{\partial S}{\partial \tau}=-\frac{1}{2}\left(\frac{\partial S}{\partial q}\right)^T\left(\frac{\partial S}{\partial q}\right)+\left(\frac{\partial S}{\partial q}\right)^T f(q)+V(q), \quad (7.39)$$

where we use the relations, $\partial_\tau \mathcal{S} = H|_\tau$ and $\partial_q \mathcal{S} = -\partial_{\dot q} L|_\tau$. (Check Theorem 6.5.1 to recall how differentiation of the action with respect to time and coordinates at the beginning and at the end of a path are related to each other.)

Notice, that the HJ equation, in the control formulation, is called Bellman or Bellman-Hamilton-Jacobi (BHJ) equation to commemorate contribution of Bellman to the field, who has formulated the problem and resolved it deriving the BHJ equation.

In Section 7.4 we derive the BHJ equation in a more general setting.

Exercise 7.2. Consider one-dimensional optimal control problem where we know where we want to arrive and we must learn how to steer there from any location, formally:

$$\{u^*(\tau), q^*(\tau)\} = \underset{\{u(\tau),q(\tau)\in\mathbb{R}\}}{\arg\min} \int_0^t d\tau \frac{(u(\tau))^2 + \beta^2 (q(\tau))^2}{2} \Bigg| \begin{array}{l} \forall \tau \in [0,t]: \quad \dot q(\tau) = -\alpha q(\tau) + u(\tau) \\ q(0) = q_0, \quad q(t) = 0 \end{array} . \tag{7.40}$$

Use a substitution to eliminate the control variable, and derive the Euler-Lagrange equations, the Hamiltonian equations, and the Hamilton-Jacobi equations. Find the optimal trajectory $q^*(\tau)$ and verify that it is consistent with the three equations. Reconstruct the optimal controller $u^*(\tau)$ and express it via (a) $q^*(\tau)$ (closed loop control); (b) q_0 (open loop control). [Hint: In your solution, you many find it convenient to define $\gamma^2 = \beta^2 + \alpha^2$.]

7.3 Pontryagin Minimal Principle

Let us now consider the following even more general optimal control problem formulated for a dynamical system in a state, $\{q(\tau) \in \mathbb{R}^d | \forall \tau \in [0,t]\}$:

$$\{u^*(\tau), q^*(\tau)\} = \tag{7.41}$$
$$\underset{\{u(\tau),q(\tau)\}}{\arg\min} \left(\phi(q(t)) + \int_0^t d\tau L\left(\tau, q(\tau), u(\tau)\right) \right) \Bigg| \begin{array}{l} \dot q(\tau) = f(\tau, q(\tau), u(\tau)), \ \forall \tau \in (0,t]; \\ u(\tau) \in U \subseteq \mathbb{R}^d, \ \forall \tau \in [0,t]; \\ q(0) = q_0 \end{array} ,$$

where the control $\{u(\tau) \in U \in \mathbb{R}^d | \tau \in [0,1]\}$ is restricted to the domain U of the d-dimensional space at all the times considered.

The analog of the standard variational calculus approach, consisting in the necessary Euler-Lagrange (EL) conditions over $\{u(\tau)\}$ and $\{q(\tau)\}$, is called Pontryagin Minimal Principle (PMP), commemorating the contribution of Lev Pontryagin to the subject [13] (see

also [14] for extended discussion of the PMP bibliography, circa 1963). We present it here without much elaboration, as it follows the same variational logic repeated by now many times in this Section. Begin by introducing the effective action,

$$\tilde{\mathcal{S}} := \phi(q(t)) + \int_0^t d\tau L\left(\tau, q(\tau), u(\tau)\right) + \int_0^t d\tau \lambda(\tau)\big(f(\tau, q(\tau), u(\tau)) - \dot{q}(\tau)\big), \tag{7.42}$$

where $\{\lambda(\tau)\}$ is a Lagrangian multiplier that is a function of τ. Then, by optimizing over $\{u\}$ and $\{q\}$, we arrive at the following set of variational equations for the candidate solution of Eq. (7.41):

$$\forall \tau \in [0,t]: \quad \min_{\{u\}} \tilde{S}: \tag{7.43}$$

$$u^*(\tau) = \arg\min_{\tilde{u}} \left(L\left(\tau, q^*(\tau), \tilde{u}(\tau)\right) + \lambda^*(\tau) f(\tau, q^*(\tau), \tilde{u}(\tau))\right),$$

$$\left.\frac{\delta \tilde{\mathcal{S}}}{\delta q(\tau)}\right|_{q(\tau)=q^*(\tau)} = 0: \tag{7.44}$$

$$\dot{\lambda}^*(\tau) = -\frac{\partial}{\partial q^*}\left(L\left(\tau, q^*(\tau), u^*(\tau)\right) + \lambda^*(\tau) f(\tau, q^*(\tau), u^*(\tau))\right),$$

$$\tau = t: \qquad \left.\frac{\partial \tilde{\mathcal{S}}}{\partial q(t)}\right|_{q(\tau)=q^*(\tau)} = 0: \tag{7.45}$$

$$\lambda^*(t) = \partial\phi(q^*(t))/\partial q^*(t).$$

Notice that Eq. (7.46) is the result of variation of $\tilde{S}$ over $q(t)$, providing the boundary conditions at $\tau = t$ by relating $q(t)$ and $\lambda^*(t)$. (Derivation of Eq. (7.46) is equivalent to the derivation of the respective boundary condition (7.15) at $\tau = t$ in the case of the LQ control discussed in Section 7.1.) Combination of Eqs. (7.43,7.44,7.46) with the (primal) dynamic equations and the initial condition on $q(0)$ (from the first line in the conditions of Eq. (7.41) completes description of the PMP approach. This PMP system of equations, stated as a Boundary Value (BV) problem, with two boundary conditions on the opposite ends of the temporal interval, is too difficult to allow an analytic solution in the general case. The system of equations is normally solved numerically by the shooting method. Solution of the PMP system of equations is not guaranteed to be unique.

Exercise 7.3. Consider a rocket, modeled as a particle of constant (unit) mass moving in zero gravity (empty) two dimensional space. Assume that the thrust/force acting on the rocket, $f(\tau)$ is a known (prescribed) function of time (dependent on, presumably pre-calculated, rate of the fuel burn), and that the direction of the thrust can be controlled. Then equations of motion for the controlled rocket are

$$\forall \tau \in (0,t]: \quad \ddot{q}_1 = f(\tau)\cos u(\tau), \quad \ddot{q}_2 = f(\tau)\sin u(\tau).$$

(a) Assume that $\forall \tau \in [0,t],\ f(\tau) > 0$. Show that $\min_{\{u(\tau),q_1(\tau),q_2(\tau)\}} \phi(q(t))$, where $\phi(q)$ is an arbitrary function, always result in the optimal control stated in the following, so-called bi-linear tangent, form:

$$\tan\left(u^*(\tau)\right) = \frac{a+b\tau}{c+d\tau}.$$

(b) Assume that the rocket starts at rest at the origin and that we want to drive it to a given height, $q_2(t) = q_*$, at the final moment of time t, such that the final velocity in the horizontal direction, $\dot{q}_1(t)$, is maximized, while $\dot{q}_2(t) = 0$:

$$\max_{\{q_1(\tau),q_2(\tau),u(\tau)\}} \dot{q}_1(t)\Big|_{q_1(0)=q_2(0)=0,\ q_2(t)=q_*,\ \dot{q}_2(t)=0}.$$

Show that the optimal control is reduced to a linear tangent law,

$$\tan\left(u^*(\tau)\right) = a + b\tau.$$

7.4 Dynamic Programming in Optimal Control and Beyond

7.4.1 Discrete Time Optimal Control

Discretizing Eq. (7.41) in time one arrives at

$$\min_{u_{0:n-1},q_{1:n}} \left(\phi(q_n) + \Delta \sum_{k=0}^{n-1} L(\tau_k, q_k, u_k)\right)\Bigg|_{k=0,\cdots,n-1:\ q_{k+1}=q_k+\Delta f(\tau_k,q_k,u_k)}, \tag{7.46}$$

where $k = 1,\cdots,n:\ \tau_k := kt/n,\ q_k := q(\tau_k),\ u_{k-1} := u(\tau_k),\ \Delta := t/n$, and q_0 is assumed fixed.

The main idea of Dynamic Programming (DP) consists in performing the optimization in Eq. (7.46), not over all the variables at once, but sequentially, one after another, that is in a greedy fashion. Specifically, let us first optimize over q_n and u_{n-1}. In fact, optimization over q_n consists simply in the substitution of q_n by $q_{n-1} + \Delta f(\tau_{n-1}, q_{n-1}, u_{n-1})$, according to the condition in Eq. (7.46) evaluated at $k = n-1$. One derives

$$S(n, q_n) := \phi(q_n), \tag{7.47}$$

$$u^*_{n-1} := \arg\min_{u_{n-1}\in U} S\left(n, q_{n-1} + \Delta f\left(\tau_{n-1}, q_{n-1}, u_{n-1}\right)\right) + \Delta L\left(\tau_{n-1}, q_{n-1}, u_{n-1}\right), \tag{7.48}$$

$$S(n-1, q_{n-1}) := S\left(n, q_{n-1} + \Delta f\left(\tau_{n-1}, q_{n-1}, u^*_{n-1}\right)\right) + \Delta L\left(\tau_{n-1}, q_{n-1}, u^*_{n-1}\right), \tag{7.49}$$

where making optimization over u_{n-1} we took advantage of the locality in the causal structure of the objective in Eq. (7.46), therefore taking into account only terms in the objective

dependent on u_{n-1}. Repeating the same scheme by first excluding, q_{n-1}, and second optimizing over u_{n-2}, and then repeating the two sub-steps (by induction) $n-1$ times (backwards in discreet time) we arrive at the following recurrent generalization of Eqs. (7.48,7.49), $k = n, \cdots, 1$:

$$u^*_{k-1} := \arg \min_{u_{k-1} \in U} \left(S\left(k, q_{k-1} + \Delta f\left(\tau_{k-1}, q_{k-1}, u_{k-1}\right)\right) + \Delta L\left(\tau_{k-1}, q_{k-1}, u_{k-1}\right)\right), \quad (7.50)$$

$$S(k-1, q_{k-1}) := S\left(k, q_{k-1} + \Delta f\left(\tau_{k-1}, q_{k-1}, u^*_{k-1}\right)\right) + \Delta L\left(\tau_{k-1}, q_{k-1}, u^*_{k-1}\right), \quad (7.51)$$

where Eq. (7.47) sets initial condition for the backward in (discrete) time iterations. It is now clear that $S(0, q_0)$ is exactly the solution of Eq. (7.46). $S(k, q_k)$, which is defined in Eq. (7.50), is called the *cost-to-go*, or the *value* function, evaluated at the (discrete) time τ_k. $L(\tau, q, u)$ and $f(\tau, q, u)$ are called (incremental) reward and (incremental) state correction. Eqs. (7.47,7.50,7.51) are summarized in Algorithm 1.

Algorithm 1 Dynamic Programming [Backward in time Value Iteration]

Input: $L(\tau, q, u),\ f(\tau, q, u), \quad \forall \tau, q, u.$

1: $\mathcal{S}(n, q) \leftarrow \phi(q)$
2: **for** $k = n, \cdots, 0$ **do**
3: $\quad u^*_k(q) \leftarrow \arg\min_u \left(\Delta L(\tau_k, q, u) + \mathcal{S}(\tau_k + 1, q_k + \Delta f(\tau_k, q_k, u))\right), \quad \forall q$
4: $\quad \mathcal{S}(k, q) \leftarrow \Delta L(\tau_k, q, u^*_k(q)) + \mathcal{S}(k+1, q_k + \Delta f(\tau_k, q, u^*_k(q))), \quad \forall q$
5: **end for**

Output: $u^*_k(q), \quad \forall q, k = n-1, \cdots, 0.$

The scheme just explained and the resulting DP Algorithm 1 were introduced in the famous paper of Richard Bellman from 1952 [15].

In accordance with the greedy nature of the DP algorithm construction—one step at a time, backward in time—is an example of what is called a greedy algorithm in Computer Science, that is an algorithm that attempts to find a globally optimal solution by making choices at each step that are only locally optimal. In general, greedy algorithms offer only a heuristic, i.e. an approximate (sub-optimal), solution. However, the remarkable feature of the optimal control problem, which we just sketched a proof of (through the sequence of transformations of Eqs. (7.47,7.50,7.51) resulted in the optimal solution of Eq. (7.46)) is that the greedy algorithm in this case is optimal/exact.

7.4.2 Continuous Time Optimal Control

Taking a continuous limit of Eqs. (7.47,7.50,7.51) one arrives at the Bellman (also called Bellman-Hamilton-Jacobi) equation (which is already familiar from the discussion of the Section 7.2, where it was derived in a special case)

$$-\partial_\tau \mathcal{S}(\tau, q) = \min_{u \in U} \Big(L(\tau, q, u) + \partial_q \mathcal{S}(\tau, q) \cdot f(\tau, q, u) \Big). \tag{7.52}$$

Then expression for the optimal control, that is continuous time version of the line 3 in the Algorithm 1, is

$$\forall \tau \in (0, t] : \; u^*(\tau, q) = \arg\min_{u \in U} \Big(L(\tau, q, u) + \partial_q \mathcal{S}(\tau, q) \cdot f(\tau, q, u) \Big). \tag{7.53}$$

Notice that the special case considered in Section 7.2, where

$$L(\tau, q, u) \to \frac{u^2}{2} + V(q), \quad f(\tau, q, u) \to f(q) + u,$$

and $U \to \mathbb{R}^d$, leads, after explicit evaluation of the resulting quadratic optimization, to Eq. (7.39).

Example 7.4.1 (Bang-Bang control of an oscillator)**.** Consider a particle of unit mass on the spring, subject to a bounded amplitude control:

$$\tau \in (0, t] : \ddot{x}(\tau) = -x(\tau) + u(\tau), \quad |u(\tau)| < 1, \tag{7.54}$$

where particle and control trajectories are $\{x(\tau) \in \mathbb{R} | \tau \in (0, t]\}$ and $\{u(\tau) \in \mathbb{R} | \tau \in (0, t]\}$. Given $x(0) = x_0$ and $\dot{x}(0) = 0$, i.e. particle is at rest initially, find the control path $\{u(\tau)\}$ such that particle position at the final moment, $x(t)$ is maximal. (t is assumed known too.) Describe optimal control and optimal solution for the case of $x(0) = 0$ and $t = 2\pi$.

Solution. First, we change from a single second order (in time) ODE to the two first order ODEs

$$\forall \tau \in (0, t] : \qquad q = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} := \begin{pmatrix} x \\ \dot{x} \end{pmatrix}, \quad \dot{q} = Aq + Bu, \tag{7.55}$$

$$A := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad B := \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{7.56}$$

We arrive at the optimal control problem (7.41) where, $\phi(q) = C^T q$, $C^T := (-1, 0)$, $L(t, q, u) = 0$, $f(t, q, u) = Aq + Bu$. Then, Eq. (7.52) becomes

$$\forall \tau \in (0, t] : \; -\partial_\tau \mathcal{S} = (\partial_q S)^T Aq - \Big| (\partial_q \mathcal{S})^T B \Big|. \tag{7.57}$$

The absolute value here comes from the fact that the optimal value for $u(\tau)$ is at one of it's extremes, either $+1$ or -1, depending on the sign of $(\partial_q \mathcal{S})^T B$. Let us look for solution by the (standard for HJ) method of variable separation, $\mathcal{S}(\tau, q) = (\psi(\tau))^T q + \alpha(\tau)$. Substituting the ansatz into Eq. (7.57) one derives

$$\forall \tau \in (0,t]: \quad \dot{\psi} = -A^T \psi, \quad \dot{\alpha} = |\psi^T B|. \tag{7.58}$$

These equations must be solved for all τ, with the terminal/final conditions: $\psi(t) = C$ and $\alpha(t) = 0$. Solving the first equation and then substituting the result in Eq. (7.53) one derives

$$\forall \tau \in (0,t]: \ \psi(\tau) = \begin{pmatrix} -\cos(\tau - t) \\ \sin(\tau - t) \end{pmatrix}, \quad u(\tau, q) = -\text{sign}(\psi_2(\tau)) = -\text{sign}\left(\sin(\tau - t)\right), \tag{7.59}$$

that is the optimal control depends only on τ (does not depend on q) and it is ± 1.

Consider for example $q_1(0) = x(0) = 0$ and $t = 2\pi$. In this case the optimal control is

$$u(\tau) = \begin{cases} -1, & 0 < \tau < \pi \\ 1, & \pi < \tau < 2\pi \end{cases}, \tag{7.60}$$

and the optimal trajectory is

$$q^T = (q_1, q_2) = \begin{cases} (\cos(\tau) - 1, -\sin(\tau)) & 0 < \tau < \pi \\ (3\cos(\tau) + 1, -3\sin(\tau)) & \pi < \tau < 2\pi \end{cases} \tag{7.61}$$

The solution consists in, first, pushing the mass down, and then up, in both cases to the extremes, i.e. to $u = -1$ and $u = 1$, respectively. This type of control is called bang-bang control, observed in the cases, like the one considered, without any (soft) cost associated with the control but only (hard) bounds.

Exercise 7.4. Consider a soft version of the problem discussed in Example 7.4.1:

$$\min_{\{u(\tau)\},\{q(\tau)\}} \left(C^T q(t) + \frac{1}{2} \int_0^t d\tau (u(\tau))^2 \right) \Bigg|_{\forall \tau \in (0,t]: \ \dot{q}(\tau) = Aq(\tau) + Bu(\tau)}, \tag{7.62}$$

where $(q(0))^T = (x_0, 0)$ and A, B and C are defined above (in the formulation and solution of the Example 7.4.1). Derive Bellman/BHJ equation, build a generic solution and illustrate it on the case of $t = 2\pi$ and $q_1(0) = x_0 = 0$. Compare your result with solution of the Example 7.4.1.

7.5 Dynamic Programming in Discrete Mathematics

Let us take a look at the Dynamic Programming (DP) from the prospective of discrete mathematics, usually associated with combinations of variables (thus combinatorics) and graphs (thus graph theory). In the following we start exploring this very rich and modern field of applied mathematics on examples.

7.5.1 LaTeX Engine

Consider a sequence of words of varying lengths, $w_1, \ldots, w_n$, and pose the question of choosing locations for breaking the sequence at $j_1, j_2, \cdots$ into multiple lines. Once the sequence is chosen, spaces between words are stretched, so that the left margin and the right margins are aligned. We are interested to place the line breaks in a way which would be most pleasing for the eye. We turn this informally stated goal into optimization requiring that word stretching in the result of the line breaking is minimal [b].

To formalize the notion of the minimal stretching consider a sequence of words labeled by index $i = 1, \cdots, n$. Each word is characterized by its length, $w_i > 0$. Assume that the cost of fitting all words in between i and j, where $j > i$, in a row is, $c(i, j)$. Then the total cost of placing n words in (presumably) nice looking text consisting of l rows is

$$c(1, j_1) + c(j_1 + 1, j_2) + \cdots + c(j_l + 1, n), \tag{7.63}$$

where $1 < j_1 < j_2 < \cdots < j_l < n$. We seek an optimal sequence that minimizes the total cost. To make the description of the problem complete, one needs to introduce a plausible way of "pricing" the line breaks. Let us define the total length of the line as a sum of all lengths (of words) in the sequence plus the number of words in the line minus one (corresponding to the number of spaces in the line before stretching). Then, one requires that the total length of the line (before stretching) to be less then the widest allowed line length L, and define the cost to be a monotonically increasing function of the stretching factor, for example

$$c(i,j) = \begin{cases} +\infty, & L < (j-i) + \sum_{k=i}^{j} w_k \\ \left(\dfrac{L - (j-i) - \sum_{k=i}^{j} w_k}{j-i} \right)^3, & \text{otherwise} \end{cases} \tag{7.64}$$

(The cubic dependence in Eq. (7.64) is an empirical way to introduce preference for smaller stretching factors. Notice also that Eq. (7.64) assumes that $j > i$, i.e. any line contains more than one word, and it does not take into account the last string in the paragraph.)

[b] The exemplary Dynamical Programming problem is borrowed from [16]. See Section 3.3.1.

At first glance the problem of finding the optimal sequence seems hard, that is exponential in the number of words. Indeed, formally one has to make a decision whether to place a break (or not) after reading each word in the sequence, thus facing the problem of choosing an optimal sequence from 2^{n-1} of possible options.

Is there a more efficient way of finding the optimal sequence? Apparently the answer to this question is the affirmative, and in fact, as we will see below that the solution is of the Dynamic Programming (DP) type. The key insight is the relation between the optimal solution of the full problem and an optimal solution of a sub-problem consisting of an early portion of the full paragraph. One discovers that the optimal solution of the sub-problem is a sub-set of the optimal solution of the full problem. This means, in particular, that we can proceed in a greedy manner, looking for an optimal solution sequentially - solving a sequence of sub-problems, where each consecutive problem extends the preceding one incrementally. (In general the greedy algorithms follow this basic structure: First, we view the solving of the problem as making a sequence of "steps" such that every time we make a "step" we end up with a smaller version of the same basic problem. Second, we follow an approach of always taking whichever "step" looks best at the moment, and we never back up and change the "step".)

Let $f(i)$ denote the minimum cost of formatting a sequence of words which starts from the word i and runs to the end of the paragraph. Then, the minimum cost of the entire paragraph is

$$f(1) = \min_j (c(1,j) + f(j+1)). \tag{7.65}$$

while a partial cost satisfies the following recursive relation

$$\forall i: \quad f(i) = \min_{j:i\le j} (c(i,j) + f(j+1)), \tag{7.66}$$

which we also supplement by the boundary condition, $f(n+1) = 0$, stating formally that no word is available for formatting when we reach the end of the paragraph. Eq. (7.66) is a full analog of the Bellman equation (7.51). Algorithm 2 is a recursive algorithm for $f(i)$ implementing Eq. (7.66).

Algorithm 2 answers the formatting question in a way smarter than naive check mentioned above. However, it is still not efficient, as it recomputes the same values of f many times, thus wasting efforts. For example, the algorithm calculates $f(4)$ whenever it calculates $f(1), f(2), f(3)$. To avoid this unnecessary step, one should save the values already calculated, by placing the result just computed into the memory. Then, by storing the results we win calling, computing and storing the functions $f(i)$ sequentially. Since we have n different values of i and the loop runs through $O(n)$ values of j, the total running time of the algorithm, relaying on the previous values stored, is $O(n^2)$.

Algorithm 2 Dynamic Programming for LaTeX Engine

Input: $c(i,j), \forall i,j = 1, \cdots, n$, e.g. according to Eq. (7.64). $f(n+1) = 0$.

```
for i = n, ··· , 1 do
    f(i) = +∞
    for j = i, ··· , n do
        f(i) ← min (f(i), c(i, j) + f(j + 1))
    end for
end for
```

Output: $f(i),\ \forall i = 1, \cdots, n$

7.5.2 Cheapest Path over Grid

Let us now discuss another problem. There is a number placed in each cell of a rectangular grid, $N \times M$. One starts from the left-up corner and aims to reach the right-down corner. At every step one can move down or right, then "paying a price" equal to the number written into the cell. What is the minimum amount needed to complete the task?

Solution: You can move to a particular cell (i,j) only from its left-most $(i-1,j)$ or upper-most $(i,j-1)$ neighbor. Let us solve the following sub-problem — find a minimal price $p[i,j]$ of moving to the (i,j) cell. The recursive formula (Bellman equation again) is:

$$p(i,j) = \min(p(i-1,j), p(i,j-1)) + a(i,j),$$

where $a(i,j)$ is a table of initial numbers. The final answer is an element $p(n,m)$. Note, that you can manually add the first column and row in the table $a(i,j)$, filled with numbers which are deliberately larger than the content of any cell (this helps as it allows to avoid dealing with the boundary conditions). See Algorithm 3.

Algorithm performance is illustrated in Fig. (7.1).

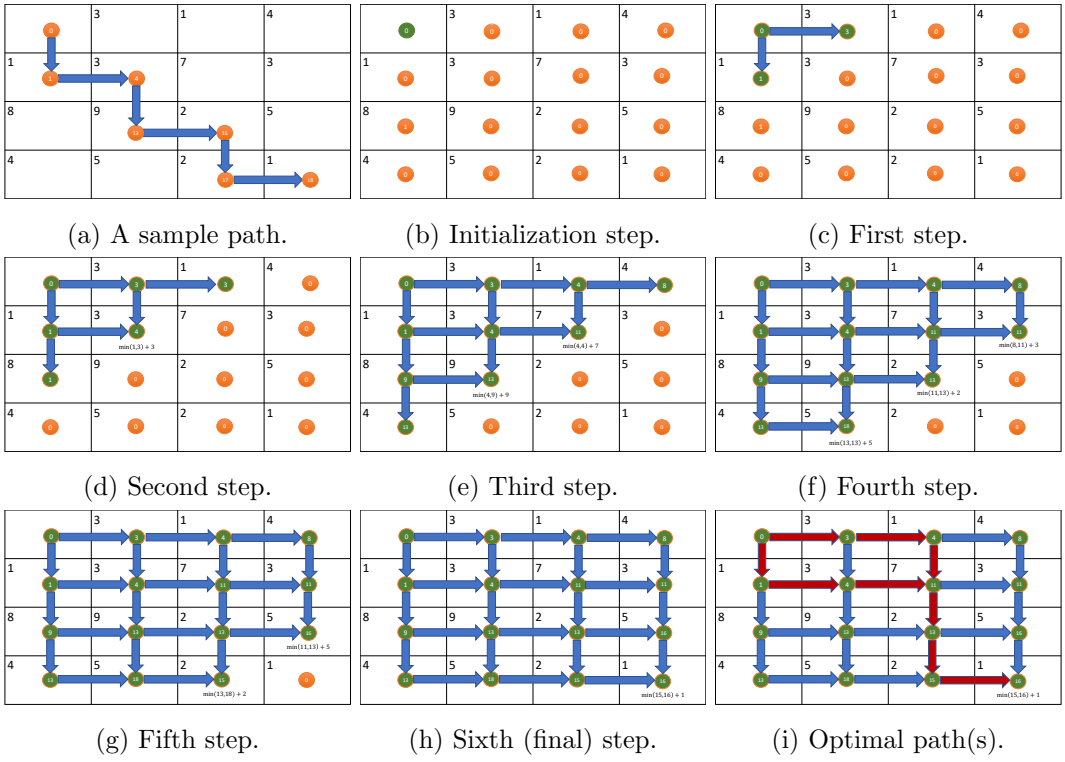
Exercise 7.5. Consider a directed acyclic graph with weighted edges (a directed acyclic graph is a directed graph with no cycles). One node is the start, and one node is the end. Construct an algorithm to compute the *maximum* cost path from the start node to the end node recursively. That is, beginning with the start node (say node 1), propagate by adjacency and keep an updated list of the max cost path from node 1 to node j. Your algorithm should *not* arbitrarily compute all possible paths. Provide pseudo-code for your algorithm and test it (by hand or with a computer) on the graph given in Fig. 7.2

Algorithm 3 Dynamic Programming for Minimum Cost Path over Grid

Input: Costs assigned: $a(i,j), \forall i = 1, \cdots, N;\ \forall j = 1, \cdots, M$. Boundary conditions fixed: $p(i,0) \leftarrow +\infty, \forall i = 1, \cdots, N$. $p(0,j) \leftarrow +\infty, \forall j = 1, \cdots, M$. Initialization: $p(1,1) \leftarrow 0$.

1: **for** $t = 2, \cdots, N+M$ **do**
2: **for** $i + j = t,\ i, j \geq 0$ **do**
3: $p(i,j) \leftarrow \min\left(p(i-1,j), p(i,j-1)\right) + a(i,j)$
4: **end for**
5: **end for**

Output: $p(i,j),\ \forall i = 1, \cdots, N;\ j = 1, \cdots, M$.



(a) A sample path. (b) Initialization step. (c) First step.

(d) Second step. (e) Third step. (f) Fourth step.

(g) Fifth step. (h) Sixth (final) step. (i) Optimal path(s).

Figure 7.1: Step-by-step illustration of the Cheapest-Path Algorithm 3 for an exemplary 4×4 grid. Number in the corner of each cell (except cell $(1,1)$ is respective a_{ij}. Values in the green circles are respective final, p_{ij}, corresponding to the cost of the optimal path from $(1,1)$ to (i,j).

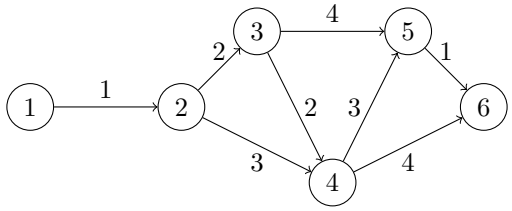


Figure 7.2: Example of a weighted directed acyclic graph.

7.5.3 DP for Graphical Model Optimization

The number of optimization problems which can be solved efficiently with DP is remarkably broad. In particular, it appears that the following combinatorial optimization problem, over binary n-dimensional variable, x:

$$E := \min_{x\in\{\pm 1\}^n} \sum_{i=1}^{n-1} E_i(x_i, x_{i+1}), \tag{7.67}$$

which requires optimization over 2^n possible states, can be solved efficiently by DP in efforts linear in n. Here, in Eq. (7.67) $E_i(x_i, x_{i+1})$ is an arbitrary, known, and possibly different for different i, real-valued function of its arguments, which are both binary. In the jargon of mathematical physics the problem just introduced is called "finding a ground state of the Ising model".

To explain the DP algorithm for this example it is convenient to represent the problem in terms of a linear graph (a chain) shown in Fig. (7.3). The components of x are associated with nodes and the "energy" of "pair-wise interactions" between neighboring components of x are associated with an edge, thus arriving at a linear graph (chain).

Let us illustrate the greedy, DP approach to solving optimization (7.67) on the example in Fig. (7.3). The greedy essence of the approach suggests that we should minimize over components sequentially, starting from one side of the chain and advancing to its opposite end. Therefore, minimizing over x_1 one derives

$$\begin{aligned} E &= \min_{x_2,\cdots,x_n}\left(\min_{x_1} E_1(x_1,x_2) + \sum_{i=2}^{n-1} E_i(x_i,x_{i+1})\right) \\ &= \min_{x_2,\cdots,x_n}\left(\tilde{E}_2(x_2,x_3) + \sum_{i=3}^{n-1} E_i(x_i,x_{i+1})\right), \end{aligned} \tag{7.68}$$

$$\tilde{E}_2(x_2,x_3) := E_2(x_2,x_3) + \min_{x_1} E_1(x_1,x_2), \tag{7.69}$$

where we took advantage of the objective factorization (into sum of terms each involving only a pair of neighboring components). Notice, that computing $\min_{x_1} E_1(x_1,x_2)$, we need

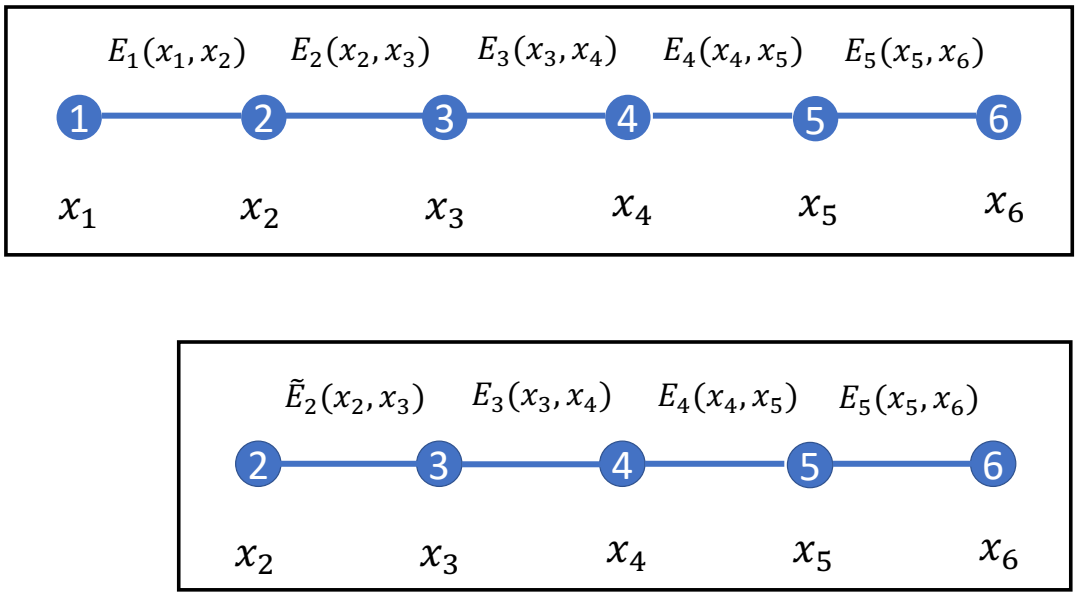


Figure 7.3: Top: Example of a linear Graphical Model (chain). Bottom: Modified GM (shorter chain) after one step of the DP algorithm.

to track the result for all possible (two) values of x_2. The end result of the (greedy) minimization (over x_1) we arrive at the problem with exactly the same structure we started with, i.e. a chain. However, the chain is shorter by one node (and edge). The only change in the new structure (when compared with the original structure) is "renormalization" of the pair-wise energy: $E_2(x_2,x_3) \to \tilde{E}_2(x_2,x_3)$. Graphical transformation associated with one greedy step is illustrated in Fig. (7.3). It shows transition from the original chain to the reduced (one node and one edge shorter) chain. Therefore, repeating the process sequentially (by induction) we will get the desired answer in exactly n steps. The DP algorithm is shown below, where we also generalize assuming that all components of x_i are drawn from and arbitrary (and not necessarily binary) set, Σ, often called "alphabet" in the Computer Science and Information Theory literature.

Consider generalization of the combinatorial optimization problem (7.69) to the case of a single-connected tree, $\mathcal{T} = (\mathcal{V},\mathcal{E})$, e.g. one shown in Fig. (7.4):

$$E := \min_{x\in\Sigma^{|\mathcal{V}|}} \sum_{\{i,j\}\in\mathcal{E}} E_{i,j}(x_i,x_j), \tag{7.70}$$

where $\mathcal{V}$ and $\mathcal{E}$ are the sets of nodes and edges of the tree respectively; $|\mathcal{V}|$ is the cardinality of the of nodes (number of nodes); and Σ is the set (alphabet) marking possible (allowed) values for any, $x_i,\ i\in\mathcal{V}$, component of x.

Algorithm 4 DP for Combinatorial Optimization over Chain

Input: Pair-wise energies, $E_i(x_i, x_{i+1}),\ \forall i = 1, \cdots, n-1$.

1: **for** $i = 1, \cdots, n-2$ **do**
2: **for** $x_{i+1}, x_{i+2} \in \Sigma$ **do**
3: $E_{i+1}(x_{i+1}, x_{i+2}) = E_{i+1}(x_{i+1}, x_{i+2}) + \min_{x_i} E_i(x_i, x_{i+1})$
4: **end for**
5: **end for**

Output: $E = \min_{x_{n-1}, x_n} E_{n-1}(x_{n-1}, x_n)$

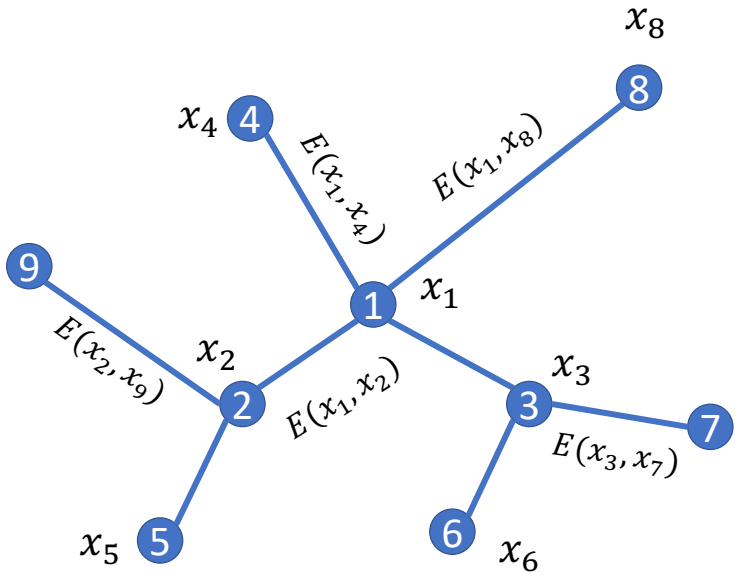


Figure 7.4: Example of a tree-like Graphical Model.

Exercise 7.6. Generalize Algorithm 4 to the case of the GM optimization problem (7.70) over a tree, that is compute E defined in Eq. (7.70). (Hint: one can start from any leaf node of the tree, and use induction as in any other DP scheme.)

Part IV

Mathematics of Uncertainty

Chapter 8

Basic Concepts from Statistics

8.1 Distributions and Random Variables

Consider a system that can exist in a number of different states. State spaces can be either continuous or discrete. An example of continuous state space is the angle between the two hands of a clock, measured clockwise from the hour hand, so $\Sigma = [0, 2\pi)$, and an example of a discrete state space is the number showing on the top of a die, so $\Sigma = \{1, 2, 3, 4, 5, 6\}$. If the state of the system is influenced by a source of randomness, then each state $x \in \Sigma$ is associated with a probability, $P(x)$, which describes the likelihood that state x is observed.

8.1.1 Discrete Random Variables

For discrete state spaces, P must satisfy

$$\forall x: \quad 0 \leq P(x) \leq 1 \tag{8.1}$$

$$\sum_{x \in \Sigma} P(x) = 1, \tag{8.2}$$

It is often useful to work with state spaces that are quantitative. For example, the set of possible outcomes of a single coin toss, $\{\text{Tail}, \text{Head}\}$, can be mapped to the quantitative state space, $\Sigma = \{0, 1\}$, by asking how many heads are observed after a single toss. For this example, the probability mass function associated with this binary discrete sample space is completely determined by a single parameter, call it β. So if $P(1) = \beta$, then $P(0) = 1 - \beta$ (See also Fig. 8.1.)

Terminology. In the example of the coin toss, we defined a random variable to be the number of heads after one coin toss. If we call this random variable X, then the notation $P(1) = \beta$ and $P(0) = 1 - \beta$ is really shorthand for $P(X = 1) = \beta$ and $P(X = 0) = 1 - \beta$.

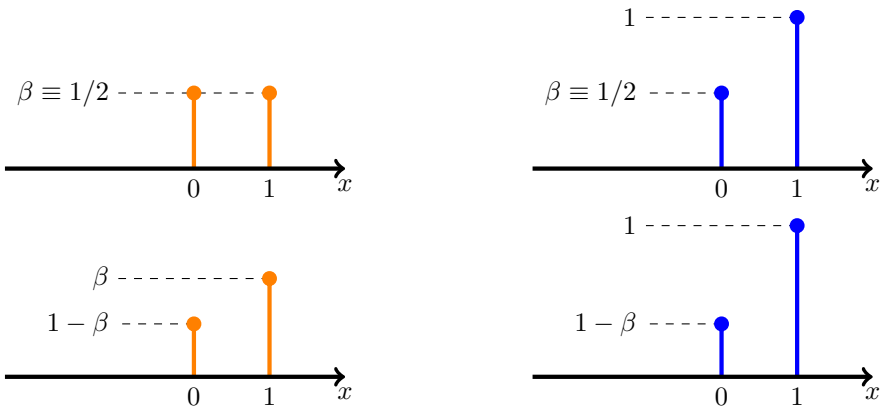


Figure 8.1: Probability mass function (left column) and cumulative distribution functions (right column) for a Bernoulli random variable with parameter $\beta \equiv 1/2$ (top) and for a Bernoulli random variable with parameter $\beta > 1/2$ (bottom).

Another common notation is $P_X(1) = \beta$ and $P_X(0) = 1-\beta$. All three notations mean "The probability that exactly one head is observed after a toss is β."

We could also write

$$P(X = k) = \begin{cases} 1-\beta, & \text{for } k = 0; \\ \beta, & \text{for } k = 1. \end{cases} \tag{8.3}$$

The probability distribution described by (8.3) is called *Bernoulli distribution* with the parameter β. A random variable X that follows the Bernoulli distribution is called a *Bernoulli random variable*, and we state it as, $X \sim \text{Bernoulli}(\beta)$.

Eq. (8.3) and Fig. 8.1 describe the Bernoulli distribution by its *Probability Mass Function* (PMF). A PMF is written in the form in the form $P(X = x) = \dots$ because it defines the probability that a random variable takes certain values. Distributions can be described by their *Cumulative Distribution Functions* (CDF). A CDF is written in the form $P(X \leq x) = \dots$ because it defines the probability that a random variable is less than or equal to a certain value. For example, the CDF of the Bernoulli distribution (8.3) is

$$P(X \leq k) = \begin{cases} 1-\beta, & \text{for } k = 0 \\ 1, & \text{for } k = 1 \end{cases} \tag{8.4}$$

See also Fig. 8.1 for illustration.

We can take our example further and ask what happens when we toss the coin n times. The set of possible outcomes of our experiment is the set of sequences of length n consisting

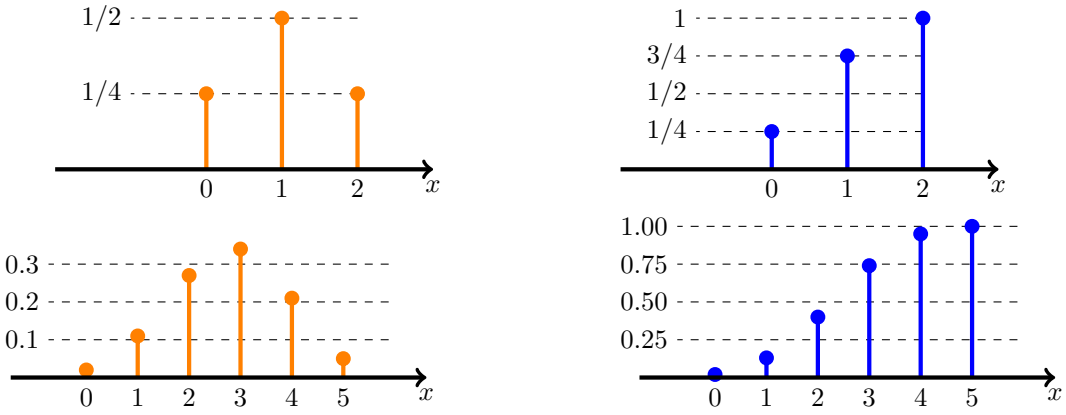


Figure 8.2: Probability Mass Functions (left column) and cumulative distribution functions (right column) for Binomial random variables with parameters $n = 2$ and $\beta = 1/2$ (top) and with parameters $n = 5$ and $\beta > 1/2$ (bottom).

of heads and tails, for example the sequence $(H, T, H, H, ..., T)$ is one possible outcome. If we define the random variable X_i to be the number of heads showing on the i^{th} toss, then the sequence $(X_i)_{i=1}^n$ is a (quantitative) sequence of ones and zeros that represents the outcome of n tosses.

For this example, we say that the random variables X_i are *independent* because the outcome of each coin toss does not depend on the previous tosses, and we say they are *identically distributed* because the underlying principles that determine the outcome of a toss do not change from toss to toss. Random variables that are both independent and identically distributed are given the shorthand *i.i.d.*

Let us define a new random variable, Y, to be the number of heads after n coin tosses, so $Y = X_1 + X_2 + \cdots + X_n$. In this situation, $\{X_i\}$ are *i.i.d.*, so the probability of tossing a sequence with exactly k heads is precisely the proportion of sequences that contain exactly k heads, which can be computed from the binomial formula giving

$$P(Y = k) = \binom{n}{k} \beta^k (1 - \beta)^{n-k}. \tag{8.5}$$

The probability distribution described by (8.5) is called a *binomial distribution* with parameters n and k (Figs. 8.2). A random variable Y that follows a Binomial distribution is called a *binomial random variable*, and we say that $Y \sim \mathrm{B}(n, \beta)$ or $Y \sim \mathrm{Binom}(n, \beta)$.

Let us now discuss an example of an unbounded discrete state space. Consider some event that occurs by chance, for example, observing a meteor in the night sky. Let K be the random variable that counts the number of such occurrences during a given period of time.

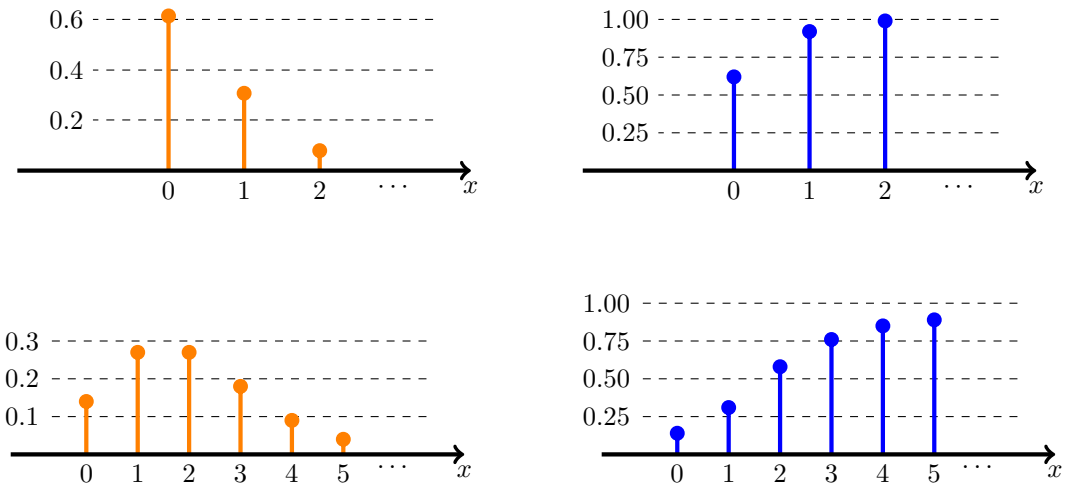


Figure 8.3: Probability mass functions (left column) and cumulative distribution functions (right column) for Poisson random variables with parameter $\lambda = 0.5$ (top) and with parameter $\lambda = 2$ (bottom).

Then $\Sigma = \{0, 1, 2, \dots\}$, (i.e, it is possible, that there are no occurrences, one occurrence, two occurrences, and so on). It can be shown that under certain conditions, K will have the probability distribution

$$P(K = k) = \frac{\lambda^k e^{-\lambda}}{k!}. \quad \text{for } k = 0, 1, 2, \dots \tag{8.6}$$

The probability distribution described by (8.6) is called a *Poisson distribution* with parameter λ (Fig. 8.3). A random variable K that follows a Poisson distribution is called a *Poisson random variable*, and we say that $K \sim \text{Pois}(\lambda)$. (Check that the probability defined in Eq. (8.6) satisfies Eq. (8.2).)

Other real-life examples of random processes associated with the Poisson distribution (called Poisson processes) are: the probability distribution of the number of phone calls received at a call center in an hour, the probability distribution of the number of customers arriving at the shop or bank, probability distribution of the number of typing errors per page page, and many more.

Example 8.1.1. Are the Bernoulli and Poisson distributions related? Can you "design" a Poisson process from Bernoulli process?

Solution. Consider repeating the Bernoulli process n times independently, thus drawing a sequence consisting of zeros and ones from a Binomial distribution. Check only for the ones and record the times (indexes) associated with their occurrences. Analyzing the probability

distribution of k arrivals in n steps in the limit $n \to \infty$ and assuming than $n\beta$ converge to a constant, i.e., $n\beta \to \lambda$ will recover a Poisson distribution. (The statement, also called in the literature the Poisson Limit Theorem, will be discussed in more detail in the following lectures.) □

8.1.2 Continuous Random Variables

The state space Σ can also be continuous. Random variables on continuous states spaces are associated with a probability density function that must satisfy

$$\forall x \in \Sigma : \quad p(x) \geq 0, \tag{8.7}$$

$$\int_{\Sigma} dx\, p(x) = 1, \tag{8.8}$$

It is customary to use lower case p for the Probability Density Function and upper case, P, to denote actual probabilities. The *Probability Density Function* (PDF) provides a means to compute probabilities that an outcome occurs in a given set or interval.

For example, for $\mathcal{A} \subset \Sigma$, then the probability of observing an outcome in the set $\mathcal{A}$ is given by

$$P(\mathcal{A}) = \int_{\mathcal{A}} p(x)\, dx.$$

Consider the probability that a real-valued X will take a value less than or equal to x:

$$P(X \leq x) = \int_{-\infty}^{x} p(x')dx'. \tag{8.9}$$

Eq. (8.9) extends to continuous space example the notion of CDF (we remind that the abbreviation stands for the already familiar from Section 8.1.1 *Cumulative Distribution Function*).

The setting can be extended from infinite to finite intervals. The uniform distribution on the interval $[a, b]$ is an example of a distribution on a bounded continuous state space:

$$\forall x \in [a, b] : \quad p(x) = \frac{1}{b-a}, \tag{8.10}$$

A random variable X with a probability distribution given by equation (8.10) can be described by the notation $X \sim \text{Unif}(a, b)$. Fig. 8.4 illustrates PDF and CDF of $\text{Unif}(a, b)$.

The Gaussian distribution is perhaps the most common (and also the most important) continuous distribution:

$$\forall x \in \mathbb{R} : \quad p(x|\sigma, \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{8.11}$$

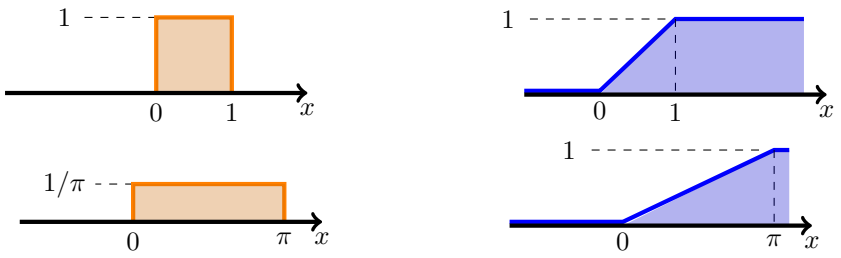


Figure 8.4: PDF (left column) and CDF (right column) for uniform random variables on $(0, 1)$ (top) and on $(0, \pi)$ (bottom).

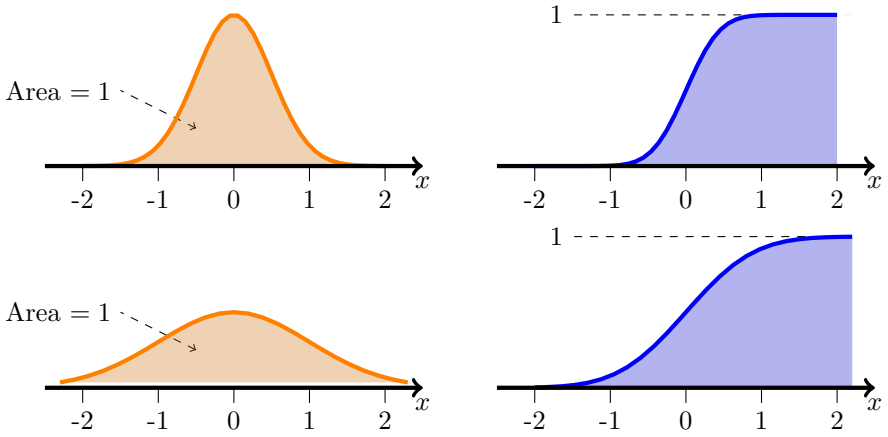


Figure 8.5: Probability density function (left column) and cumulative distribution function (right column) for normally distributed random variables $\mathcal{N}(0, 1)$ (top) and $\mathcal{N}(1, 0.5^2)$ (bottom).

$p_{\sigma,\mu}(x)$ another possible notation. The distribution is parameterized by the mean, μ, and by the variance σ^2. Standard notation for the Gaussian/normal distribution is $\mathcal{N}(\mu, \sigma^2)$ or $N(\mu, \sigma^2)$. Fig. 8.4 illustrates PDF and CDF of $\mathcal{N}(\mu, \sigma^2)$.

The probability distribution given by (8.11) is also called a *normal distribution*—where "normality" refers to the fact that the Gaussian distribution is a "normal/natural" outcome of summing up many random numbers, regardless of the distributions for the individual contributions. (We will discuss the law of large numbers and central limit theorem shortly.)

Let us make a brief remark about the notations. We will often write, $P(X = x)$, or a short-cut, $P(x)$ and sometimes you see in the literature, $P_X(x)$. By convention, upper case variables denote random variables. A random variable takes on values in some domain, and a particular observation of a random variable (that is, it has been sampled and observed

to have a particular value in the domain) is then a non-random value and is denoted by lower case e.g. x. $X \sim P(x)$ denotes the fact that the random variable X is drawn from the distribution, $P(x)$. Also, and when it is not confusing, we will streamline the notations (and thus abuse them a bit) and use the lower case variables across the board – for both a random variable and for a deterministic value the random variable takes.

8.1.3 Sampling. Histograms.

Random process generation. Random process is generated/sampled. Any computational package/software contains a random number generator (even a number of these). Designing a good random generation is important. In this course, however, we will mainly be using the random number generators (in fact pseudo-random generators) already created by others.

Histogram. To show distributions graphically, you may also "bin" it in the domain - thus generating the histogram, which is a convenient way of showing $p(\sigma)$ (see plots in the attached julia notebook with illustration breaking $[0, 1]$ interval in $N > 1$ bins).

8.2 Moments & Cumulants

8.2.1 Expectation & Variance

It is often useful to use as few numbers as possible to describe a probability distribution as meaningfully as possible.

The cumulants of a probability distribution are one of the most common set of descriptors of the distribution. The first two cumulants are well known: the mean measures the central tendency of a distribution and the variance measures the spread of a distribution about the mean. The third and fourth cumulant are less well known: the skewness measures asymmetry about the mean and the kurtosis measures whether extreme values are unusually rare or common (Fig. 8.6).

The cumulants of a distribution are often found by taking combinations of the distribution's moments, which can be computed directly by summation or integration of certain quantities (see below). Alternatively, the moments and the cumulants of a distribution can be derived from its moment generating function and its characteristic function (see below for details).

The moment generating function and the characteristic function are also frequently used in more theoretical analysis, which is beyond the scope of this course.

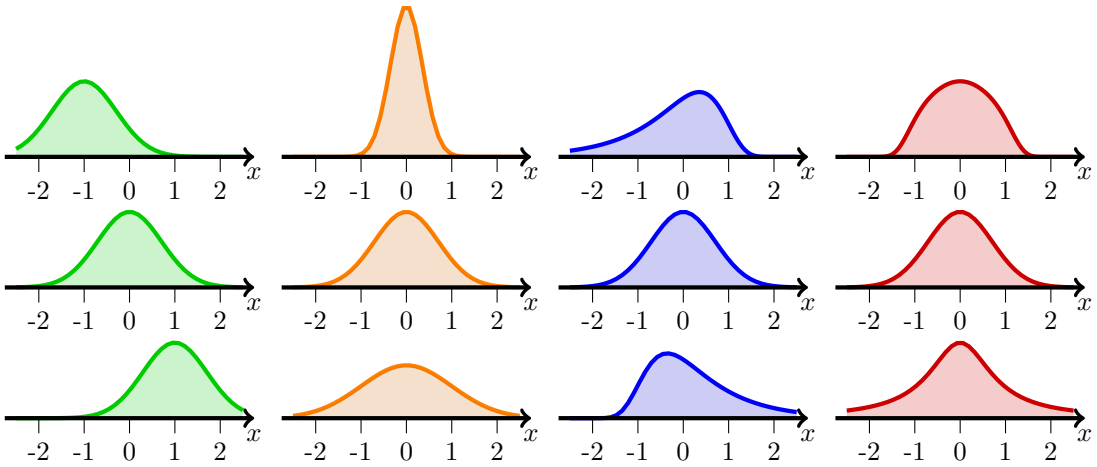


Figure 8.6: The first cumulant describes the central tendency (mean) of a probability distribution (green). The second cumulant describes the spread about the mean (orange). The third moment describes the asymmetry of the distribution (blue). The fourth moment describes whether extreme values are unusually rare or common (red).

For a random (discrete or continuous) variable X, the *expectation of* X is defined as

$$\mathbb{E}[X] := \sum_{x \in \Sigma} x\, P(x), \quad \text{(discrete)} \tag{8.12}$$

$$\mathbb{E}[X] := \int_{x \in \Sigma} dx x\, p(x), \quad \text{(continuous)}. \tag{8.13}$$

Notation. Common notation for the expectation of a random variable X with probability mass function P includes $\mathbb{E}[X]$, $\mathbb{E}_P[X]$, $\langle X \rangle$ and $\langle X \rangle_P$.

Example 8.2.1. Consider the example of tossing a pair of fair coins. The set of possible outcomes is $\{(T,T),(T,H),(H,T),(H,H)\}$, each outcome occurring with equal probability. Define the random variable X to be the number of heads that are observed, so $X \sim B(2,1/2)$ and

$$P(X = x) = \begin{cases} 1/4, & \text{for } x = 0 \\ 1/2, & \text{for } x = 1 \\ 1/4, & \text{for } x = 2 \end{cases}$$

The expected number of heads is

$$\mathbb{E}[X] = \sum_{x=\{0,1,2\}} x\, P(x) = 0 \cdot 1/4 + 1 \cdot 1/2 + 2 \cdot 1/4 = 1$$

The expectation can also be defined for functions of a random variable. Consider a function, $f(x)$, and its expectation over the probability, $P(x)$:

$$\mathbb{E}_P\left[f(x)\right] = \langle f(x)\rangle_P = \sum_{x\in\Sigma} f(x)P(x) \quad \text{(discrete)}$$
$$\mathbb{E}_p\left[f(x)\right] = \langle f(x)\rangle_p = \int_{x\in\Sigma} f(x)p(x)dx \quad \text{(continuous)}$$

Example 8.2.2. Consider a scenario where a gambler wins \$200 for tossing a pair of heads, but loses \$100 for any other outcome. If we define $f : \Sigma \to \mathbb{R}$ to be the earnings, then the expectation of f is calculated to be

$$\mathbb{E}[f] = -100 \cdot 3/4 + 200 \cdot 1/4 = -25$$

The variance of a random variable is defined as

$$\mathrm{Var}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \tag{8.14}$$

The *variance* of a random variable measures its expected spread about its mean. Note that the mean and the variance do not have the same units (the units of the variance are the square of the units of the mean), so it can be difficult to meaningfully interpret the variance. Consequently, it is common to consider the standard deviation of a random variable, σ, which is defined as the square root of the variance.

Example 8.2.3. Compute the variance and the standard deviation of X and f for the Example 8.2.1 and the Example 8.2.2.

Solution. $\mathrm{Var}[X] = (1-0)^2 \cdot 1/4 + (1-1)^2 \cdot 1/2 + (1-2)^2 \cdot 1/4 = 1/2$, and $\sigma = 1/4$.
$\mathrm{Var}[f(X)] = (-100+25)^2 \cdot 3/4 + (200+25)^2 \cdot 1/4 = 4275$, and $\sigma = 65.4$.

Example 8.2.4. The Cauchy distribution plays an important role in physics, since it describes the resonance behavior, e.g. shape of a spectral width of a laser. The probability density function of a Cauchy distribution with parameters $a \in \mathbb{R}$ and $\gamma > 0$ is given by

$$p(x|a,\gamma) = \frac{1}{\pi}\frac{\gamma}{(x-a)^2+\gamma^2}, \quad -\infty < x < +\infty. \tag{8.15}$$

Show that the probability distribution is properly normalized and find its mean. What can you say about its variance?

Solution. To verify that equation (8.15) is properly normalized, we must show that the probability density integrates to unity, which can be done with the trig-substitution, $\tan(\theta) = (x-a)/\gamma$.

To compute the mean of the Cauchy distribution, we evaluate

$$\text{mean} = \frac{\gamma}{\pi} \int_{-\infty}^{+\infty} \frac{x\,dx}{(x-a)^2+\gamma^2} = a.$$

which is calculated using using a principal value integral from residue calculus. To compute the second moment, we attempt to evaluate the integral

$$\text{variance} = \frac{\gamma}{\pi} \int_{-\infty}^{+\infty} \frac{(x-a)^2\,dx}{(x-a)^2+\gamma^2},$$

and find that it is unbounded. Since this integral is unbounded, we conclude that the variance of a Cauchy distribution does not exist (it it is infinite). □

8.2.2 Higher Moments

The concept of expectation and variance can be generalized to the *moments* of a distribution. For a discrete random variable with probability distribution $P(x)$ the *moments* of $P(x)$ are defined as follows

$$k = 0, \cdots, \quad \mu_k := \mathbb{E}_P\left[X^k\right] = \langle X^k \rangle_P = \sum_{x \in \Sigma} x^k P(x). \tag{8.16}$$

For a continuous random variable, X with probability density $p(x) = p_X(x)$, the moments of X are:

$$k = 0, \cdots, \quad \mu_k := \mathbb{E}_p\left[X^k\right] = \langle X^k \rangle_p = \int_{\Sigma} dx\, x^k p(x). \tag{8.17}$$

From the definitions in equations (8.16) and (8.14), it follows that the first moment of a random variable is equivalent to its mean, $\mathbb{E}[X] = \mu_1$, and the second moment is related to the variance according to $VarX = \mu_2 - \mu_1^2$.

Example 8.2.5. Give a closed-form expression for the moments of a Bernoulli distribution with parameter β. Use the first and second moment to find the mean and variance of the Bernoulli distribution.

$$p(x) = \beta\delta(1-x) + (1-\beta)\delta(x). \tag{8.18}$$

Solution.

$$k = 1, \cdots: \quad \mu_k = \langle X^k \rangle = \int_{-\infty}^{\infty} x^k\, p(x) dx = \beta. \tag{8.19}$$

The mean of a distribution is equal to it's first moment: $\mu = \mu_1$. In this case $\mu = \beta$. The variance of a distribution is equal to the combination of its first two moments given by $\sigma^2 = \mu_2 - \mu_1^2$. In this case the variance is $\sigma^2 = \beta - \beta^2 = \beta(1-\beta)$. □

Example 8.2.6. What is the mean number of the events in the Poisson process, Pois(λ). What is the variance of the Poisson distribution?

Solution. For this example, we will compute the mean and the variance of the Poisson distribution by first computing its first and second moments:

$$\mu_1 = \sum_{k=0}^{\infty} k\, P(k) = \sum_{k=0}^{\infty} \frac{k\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} = \lambda \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} = \lambda.$$

The second moment is

$$\mu_2 = \sum_{k=0}^{\infty} k^2\, P(k) = \sum_{k=0}^{\infty} \frac{k^2\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{k\lambda^k}{(k-1)!} e^{-\lambda} = \lambda \sum_{n=0}^{\infty} \frac{(n+1)\lambda^n}{n!} e^{-\lambda} = \lambda(\lambda+1),$$

Therefore, the mean (average) and the variance of the number of events is

$$\mu := \mu_1 = \lambda \quad \text{and} \quad \sigma^2 := \mu_2 - \mu_1^2 = \lambda$$

Note that the expectation and the variance of a Poisson distribution are both equal to the same value, λ. This is not generally true for other distributions.

8.2.3 Moment Generating Functions.

The *moment generating function* of a random variable is defined as

$$M_X(t) = \mathbb{E}\left[\exp(tx)\right] = \int_{-\infty}^{\infty} dx p(x) \exp(tx). \tag{8.20}$$

where $t \in \mathbb{R}$ and all integrals are assumed well defined. When $\exp(tx)$ is expressed by its Taylor series, we find that the moment generating function can be expressed as an infinite sum involving the moments, μ_k, of the random variable.

$$M_X(t) = \int_{-\infty}^{\infty} dx p(x) \exp(tx) = \int_{-\infty}^{\infty} dx p(x) \sum_{k=0}^{\infty} \frac{(tx)^k}{k!} = \sum_{k=0}^{\infty} \frac{\mu_k t^k}{k!}. \tag{8.21}$$

The name 'moment generating function' arises from the observation that differentiating $M_X(t)$ k times and evaluating the result at $t = 0$ recovers the k^{th} moment of X.

$$\left.\frac{d^k}{dt^k}\right|_{t=0} M_X(t) = \left.\frac{d^k}{dt^k}\right|_{t=0} \left(\sum_{k=0}^{\infty} \frac{\mu_k t^k}{k!} \right) = \mu_k \tag{8.22}$$

Example 8.2.7. Consider standard example of Boltzmann distribution from statistical mechanics, where the probability density, $p(x)$, of a random state (variable), X is

$$p(x) = \frac{1}{Z} e^{-\beta E(x)}, \quad Z(\beta) = \sum_x e^{-\beta E(x)}, \tag{8.23}$$

where $\beta = 1/T$ is the inverse temperature and $E(x)$ is a known function of x, called energy of the state x. The normalization factor Z is called the *partition function*. Suppose we know the partition function, $Z(\beta)$ as a function of the inverse temperature, β. Compute the expected mean value and the variance of the energy.

Solution. The mean (average) value of the energy is

$$\langle E(X)\rangle = \sum_x p(x)E(x) = \frac{1}{Z}\sum_x E(x)e^{-\beta E(x)} = -\frac{1}{Z}\frac{\partial Z}{\partial \beta} = -\frac{\partial \ln Z}{\partial \beta}. \tag{8.24}$$

The variance of the energy (energy fluctuations) is

$$\mathrm{Var}\left[E(X)\right] = \langle (E(X) - \langle E(X)\rangle)^2\rangle = \frac{\partial^2 \ln Z}{\partial \beta^2}, \tag{8.25}$$

Notice that up to sign inversion of the argument of the partition function is equivalent to the moment generating function (8.20), $Z(\beta) = M_{E(X)}(-\beta)$. □

8.2.4 Characteristic Functions

The *characteristic function* of a random variable is defined as the Fourier transform of its probability density:

$$G(t) := \mathbb{E}_p\left[\exp(itx)\right] = \int_{-\infty}^{+\infty} dx p(x)\exp(itx), \tag{8.26}$$

where $i^2 = -1$. The characteristic function exists for any real t and it obeys the following relations

$$G(0) = 1, \quad |G(t)| \leq 1. \tag{8.27}$$

The characteristic function contains information about all the moments μ_k. Moreover it allows the Taylor series representation in terms of the moments:

$$G(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!}\langle X^k\rangle, \tag{8.28}$$

and thus

$$\langle X^k\rangle = \frac{1}{i^k}\frac{\partial^k}{\partial t^k}G(t)\Big|_{t=0}. \tag{8.29}$$

This implies that derivatives of $G(t)$ at $t = 0$ exist up to the same m as the moments μ_k.

Example 8.2.8. Find the characteristic function of a Bernoulli distribution with parameter β.

Solution. Substituting Eq. (8.18) into the Eq. (8.26) one derives

$$G(t) = 1 - \beta + \beta e^{it}, \tag{8.30}$$

and thus

$$\mu_k = \frac{\partial^k}{i^k \partial t^k} \left(1 - \beta + \beta e^{it}\right) \Big|_{t=0} = \beta. \tag{8.31}$$

The result is naturally consistent with Eq. (8.19). □

Exercise 8.1. The exponential distribution has a probability density function given by

$$p(x) = \begin{cases} Ae^{-\lambda x}, & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{8.32}$$

where the parameter $\lambda > 0$. Calculate

(a) The normalization constant A of the distribution.

(b) The *mean* and the *variance* of the probability distribution.

(c) The characteristic function $G(t)$ of the exponential distribution.

(d) The k^{th} moment of the distribution (utilizing $G(t)$).

8.2.5 Cumulants

The *cumulants* κ_k of a random variable X are defined by the characteristic function as follows

$$\ln G(t) = \sum_{k=1}^{\infty} \frac{(it)^k}{k!} \kappa_k. \tag{8.33}$$

According to Eq. (8.27) the Taylor series in Eq. (8.33) start from unity. Utilizing Eqs. (8.28) and (8.33), one derives the following relations between the cumulants and the moments

$$\kappa_1 = \mu_1, \tag{8.34}$$

$$\kappa_2 = \mu_2 - \mu_1^2 = \sigma^2. \tag{8.35}$$

The procedure naturally extends to higher order moments and cumulants.

Notice that moments determine the cumulants in the sense that if the all the moments of any two probability distributions are identical then all the cumulants will be identical as well, and similarly the cumulants determine the moments. In some cases theoretical treatments of problems in terms of cumulants are simpler than those using moments.

Example 8.2.9. Find the characteristic function and the cumulants of the Poisson distribution (8.6).

Solution. The respective characteristic function is

$$G(t) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} e^{itk} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^{it})^k}{k!} = \exp\left[\lambda(e^{it} - 1)\right], \tag{8.36}$$

and then

$$\ln G(t) = \lambda(e^{it} - 1). \tag{8.37}$$

Next, using the definition (8.33), one finds that $\kappa_k = \lambda$, $k = 1, 2, \ldots$. □

Example 8.2.10. Birthday Problem Assume that a year has 366 days. What is the probability, p_m, that m people in a room all have different birthdays?

Solution. Let $(b_1, b_2, \ldots, b_m)$ be a list of people birthdays, $b_i \in \{1, 2, \ldots, 366\}$. There are 366^m different lists, and all are distributed identically (equiprobable). We should count the lists, which have $b_i \neq b_j$, $\forall i \neq j$. The amount of such lists is $\prod_{i=1}^{m}(366 - i + 1)$. Then, the final answer

$$p_m = \prod_{i=1}^{m} \left(1 - \frac{i-1}{366}\right). \tag{8.38}$$

The probability that at least 2 people in the room have the same birthday day is $1 - p_m$. Note that $1 - p_{23} > 0.5$ and $1 - p_{22} < 0.5$. □

Exercise. (not graded) Choose, at random, three points on the circle of unit radius. Interpret them as cuts that divide the circle into three arcs. Compute the expected length of the arc that contains the point $(1, 0)$.

8.3 Probabilistic Inequalities.

Here are some useful probabilistic inequalities.

- (Markov Inequality) For all non-negative random X

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \tag{8.39}$$

- (Chebyshev's inequality)

$$P(|X - \mu| \geq b) \leq \frac{\sigma^2}{b^2}, \tag{8.40}$$

  where μ and σ^2 are the mean and the variance of X.

- (Chernoff bound)

$$P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}}, \tag{8.41}$$

where $X \in \mathbb{R}$ and $t \geq 0$.

Example 8.3.1. Prove Markov, Chebyshev and Chernoff inequalities/bounds.

Solution. We prove the Markov inequality (8.39) in two steps. First, let us introduce the indicator function, $\mathbb{1}(y)$, which returns unity if $y \geq 0$, and is zero otherwise, and observe that the left hand side of Eq. (8.39) can be restated as $\mathbb{E}\left[\mathbb{1}(a - X)\right]$. Second, notice that $\forall X \geq 0,\ \mathbb{1}(X - a) \leq X/a$. Taking expectation of (averaging) the inequality we arrive at the desired result (8.39). □

Notice that $\mathbb{1}(X - a) = X/a$ only if $X = 0$ or $X = 1$. Therefore, the Markov inequality becomes equality iff, $P(X \in \{0, a\}) = 1$.

The Chebyshev inequality (8.40) and the Chernoff bound (8.41) are corollaries of the Markov inequality (8.39).

Indeed, to prove the Chebyshev inequality (8.40) we consider an unbounded random X, $X \in \mathbb{R}$, and apply the Markov inequality (8.39) to the following auxiliary random variable, $Y = (X - \mathbb{E}[X])^2 = (X - \mu)^2$, which is thus non-negative by construction. Then substituting a by b^2, and observing that, $\mathbb{E}[Y] = \sigma^2$, and $P(Y \geq b^2) = P(\sqrt{Y} \geq b)$, we arrive at the Chebyshev inequality (8.40). □

Similarly, to prove the Chernoff bound (8.41) we consider $Z = \exp(tX)$, where X is unbounded random variable, $X \in \mathbb{R}$, and $t \geq 0$. Z is positive by construction. Therefore, observing that, $P(X \geq a) = P(e^{tX} \geq e^{ta})$ and applying the Markov inequality (8.39) to Z, where a is substituted by e^{ta}, we arrive at the Chernoff bound (8.41). □

Notice that the Chernoff bound (8.41) can be viewed as the Markov inequality applied to the moment generating function.

We will get back to discussion of some other useful probabilistic inequalities in the lecture devoted to entropy and to how compare probabilities.

8.4 Random Variables: from one to many.

Transition from one to many random variables is natural. We will review notions which are instrumental for the transition in this Section. However, it is also useful to note that we have already touched on the issue of the multi-variate probability distributions in the preceding Section when we constructed more complex (but still single-variate) probability distributions from simpler ones. In particular, we have generated a sequence of *independent*

$P(H,H) = 1/4$ $P(H,T) = 1/4$ $P(T,H) = 1/4$ $P(T,T) = 1/4$

Figure 8.7

random variables $X_1, X_2, \cdots$ and then created a new variable by taking a function, e.g. a sum, of the original random variables. The assumed independence was useful for reaching the goal of designing a new random variable (e.g. for transitioning from Bernoulli random variable to Poisson random variable), however not all random variables are independent. In the coming Section we will learn how to describe dependencies or correlations, that is opposite of the independence, in high-dimensional statistics.

8.4.1 Multivariate Distributions. Marginalization. Conditional Probability.

Consider an n-component random vector, $\boldsymbol{X}$, and let $P(\boldsymbol{x})$ be the probability that the state $\boldsymbol{x}$ is observed, where $\sum_{\boldsymbol{x}} P(\boldsymbol{x}) = 1$. (Recall: $\boldsymbol{x}$ (lower-case) represents a particular realization of the random variable $\boldsymbol{X}$ (upper-case).) So if each component X_i takes values in $\Sigma = \{0, 1\}$, then $\boldsymbol{X}$ is the random vectors of length n with entries zero or one, and $\boldsymbol{x}$ might be $(1, 1, 0, \ldots, 1)$.) We wish to ask two related questions about P:

1. Marginalization: The probability of observing a state where one or more of the components attain certain values.

2. Conditioning: The probability of observing a state given that the value(s) attained by one or more component is known.

Example 8.4.1. Let X_i be the random variable for the number of heads observed after flipping a fair coin on the i^{th} toss. So $\Sigma = \{0, 1\}$, and $P(X_i = 0) = 1/2$ and $P(X_i = 1) = 1/2$. Let $\boldsymbol{X} = (X_1, X_2)$ be the random vector showing the outcome of two successive coin flips. So the probability of each possible outcome is

$$P(\boldsymbol{X} = (0,0)) = 1/4, \quad P(\boldsymbol{X} = (1,0)) = 1/4, \quad P(\boldsymbol{X} = (0,1)) = 1/4, \quad P(\boldsymbol{X} = (1,1)) = 1/4,$$

See Fig. (8.7) for illustration. The following questions are examples of marginalization and conditioning:

1. Marginalization: What is the probability of observing a state where the outcome of the first toss was a "1"?

2. Conditioning: It is known that the outcome of the first toss is a "1", what is the set of possible outcomes and their respective probabilities?

Solution. 1. From the list of all possible outcomes, we see that there are two outcomes with a "1" in the first entry (namely, $(1,0)$ and $(1,1)$) each with probability $1/4$.

$$P(X_1 = 1) = P(\boldsymbol{X} = (1,0)) + P(\boldsymbol{X} = (1,1)) = 1/4 + 1/4 = 1/2.$$

2. Under the condition that the outcome of the first toss is "1", all other outcomes must have zero probability. The only two possible outcomes are $(1,0)$ and $(1,1)$ which each occur with equal probability. Therefore,

$$P(X_2 = 0 \,|\, X_1 = 1) = 1/2, \quad P(X_2 = 1 \,|\, X_1 = 1) = 1/2.$$

Example 8.4.1 is relatively straightforward because X_1 and X_2 are independent. The next example will be more interesting.

Consider the statistical version of the Ising model which was discussed in Section 7.5 in the context of discrete optimization. (We have used it back then to illustrate application of the Dynamic Programming in the combinatorial optimization.) Historically, the Ising model was first used to describe the polarization of a magnet at different temperatures in the context of physics. In this present context, it might be used to model predictions for outbreaks of an epidemic in different neighborhoods.

We introduce the following probability distribution over the 2^n-dimensional space Σ (space of cardinality 2^n):

$$\boldsymbol{x} = (x_i = \pm 1 | i = 1, \cdots, n) : P(\boldsymbol{x}) = Z^{-1} \exp\left(\sum_{i=1}^{n-1} J x_i x_{i+1}\right), \tag{8.42}$$

$$Z = \sum_{\boldsymbol{x}} \exp\left(\sum_{i=1}^{n-1} J x_i x_{i+1}\right), \tag{8.43}$$

where J is a constant that determines the coupling strength between adjacent components and Z is the normalization constant (See figure 8.8.). The normalization constant, also called the partition function, is introduced to guarantee that the sum of probabilities over all the states is unity. (See also Example 8.2.7.)

For $n = 2$ one gets the example of a bi-variate probability distribution

$$P(\boldsymbol{x}) = P(x_1, x_2) = \frac{\exp(J x_1 x_2)}{4\cosh(J)}. \tag{8.44}$$

$P(\boldsymbol{x})$ is called a **joint** or **multivariate** probability distribution function of $\boldsymbol{x} = (x_1, \cdots, x_n)$, because it shows probability of all the components, $x_1, \cdots, x_n$, together.

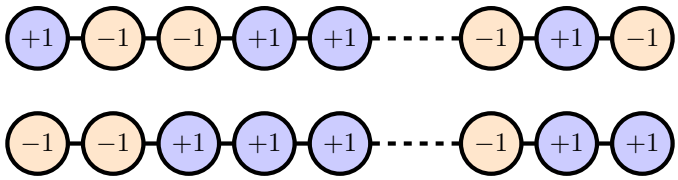


Figure 8.8: Two possible realizations of the n-component Ising model on a line.

$P(+1,+1) \propto \exp(+J)$ $P(+1,-1) \propto \exp(-J)$ $P(-1,+1) \propto \exp(-J)$ $P(-1,-1) \propto \exp(+J)$

Figure 8.9: Set of all possible outcomes for the 2-component Ising model and their relative probabilities. The normalization constant, Z is found by summing the probabilities over all states.

It is useful to consider **conditional** probability distribution. For the example above with $n = 2$,

$$P(x_1|x_2) = \frac{P(x_1, x_2)}{\sum\limits_{x_1} P(x_1, x_2)} = \frac{\exp(Jx_1x_2)}{2\cosh(Jx_2)} \tag{8.45}$$

is the probability to observe x_1 under condition that x_2 is known. Notice that, $\sum_{x_1} P(x_1|x_2) = 1,\ \forall x_2$.

We can marginalize the multivariate (joint) distribution over a subset of variables. For example,

$$P(x_1) = \sum_{\boldsymbol{x}\setminus x_1} P(\boldsymbol{x}) = \sum_{x_2,\cdots,x_n} P(x_1, \cdots, x_n). \tag{8.46}$$

Multivariate Gaussian (Normal) distribution

Now let us consider n zero-mean random variables $X_1, X_2, \ldots, X_n$ sampled i.i.d. from a generic Gaussian distribution

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \exp\left(-\frac{1}{2} \sum_{i,j=1,\cdots,n} x_i A_{ij} x_j\right), \tag{8.47}$$

where A is the symmetric, $A = A^T$, positive definite, $A \succ 0$, matrix. If the matrix is diagonal then the probability distribution (8.47) is decomposed into a product of terms, each dependent on one of the variables. This is the special case when each of the random

variables, $X_1, \cdots, X_n$, is statistically independent of others. Z in Eq. (8.47) is the normalization factor (remind that we call it, consistently with examples above, the partition function) which is

$$Z = \frac{(2\pi)^{n/2}}{\sqrt{\det A}}. \tag{8.48}$$

Moments of the Gaussian distribution are

$$\forall i: \quad \mathbb{E}\left[X_i\right] = \mu_i; \quad \forall i,j: \quad \mathbb{E}\left[(X_i-\mu_i)(X_j-\mu_j)\right] = (A^{-1})_{ij} := \Sigma_{ij}, \tag{8.49}$$

where $A^{-1}_{ij} = \Sigma_{ij}$ denotes i,j component of the inverse of the matrix A. The Σ matrix (which is also symmetric and positive definite, as its inverse is by construction) is called the co-variance matrix. Standard notation for the multi-variate statistics with mean vector, $\mu = (\mu_i | i = 1, \cdots, n)$ and co-variance matrix, Σ, is $\mathcal{N}(\mu, \Sigma)$ or $\mathcal{N}_n(\mu, \Sigma)$.

The Gaussian distribution is remarkable because of its "invariance" properties.

Theorem 8.4.2 (Invariance of Normal/Gaussian distribution under conditioning and marginalization)**.** Consider $X \sim \mathcal{N}_n\left(\mu, \Sigma\right)$ and split the n dimensional random vector into two components, $X = (X_1, X_2)$, where X_1 is a p-component sub-vector of X and X_2 is a q-component sub-vector of X, $p+q=n$. Assume also that the mean vector, μ, and the covariance matrix, Σ, are split into components as follows

$$\mu = (\mu_1, \mu_2); \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{8.50}$$

where thus μ_1 and μ_2 are p and q dimensional vectors and Σ_{11}, Σ_{12}, Σ_{21} and Σ_{22} are $(p \times p)$, $(p \times q)$, $(q \times p)$ and $(q \times q)$ matrices. Then, the following two statements hold:

- Marginalization: $p(x_1) := \int dx_2 p(x_1, x_2)$ is the following Normal/Gaussian distribution, $\mathcal{N}\left(\mu_1, \Sigma_{11}\right)$.
- Conditioning: $p(x_1|x_2) := \frac{p(x_1,x_2)}{p(x_2)}$ is the Normal/Gaussian distribution, $\mathcal{N}\left(\mu_{1|2}, \Sigma_{1|2}\right)$, where

$$\mu_{1|2} := \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \quad \Xi_{1|2} := \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}. \tag{8.51}$$

Proof of the theorem is recommended as a useful technical exercise (not graded) which requires direct use of some basic linear algebra. (You will need to use or derive explicit formula for the inverse of a positive definite matrix split into four quadrangles, as in Eq. (8.50).)

8.4.2 Central Limit Theorem

Take n random instances, also called samples, $X_1, \cdots, X_n$ generated i.i.d. from a distribution with mean μ and variance, $\sigma > 0$, and compute $Y_n = \sum_{i=1}^n X_i/n$. What is Prob(Y_n)?

Theorem 8.4.3 (Weak Version of the Central Limit Theorem)**.** $\sqrt{n}(Y_n - \mu)$, converges in distribution to a Gaussian with mean, μ, and variance, σ^2, i.e.

$$n \to \infty : \quad \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \sim \mathcal{N}(0,\ \sigma^2). \tag{8.52}$$

Let us sketch the prove of the weak-CLT (8.52) in a simple case $\mu = 0$, $\sigma = 1$. Obviously, $\mu_1(Y_n\sqrt{n}) = 0$. Compute

$$\mu_2(Y_n\sqrt{n}) = \mathbb{E}\left[\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right)^2\right] = \frac{\sum_i \mathbb{E}\left[X_i^2\right]}{n} + \frac{\sum_{i\neq j}\mathbb{E}\left[X_i X_j\right]}{n} = 1.$$

Now the third moment:

$$\mu_3(Y_n\sqrt{n}) = \mathbb{E}\left[\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right)^3\right] = \frac{\sum\limits_i \mathbb{E}\left[X_i^3\right]}{n^{3/2}} \to 0,$$

at $n \to \infty$, assuming $\mathbb{E}\left[X_i^3\right] = O(1)$. Can you guess what will happen with the fourth moment? It is, $\mu_4(Y_n\sqrt{n}) = 3 = 3m_2(Y_n)$.

Example 8.4.4 (Sum of Gaussian variables)**.** Compute the probability density, $p_n(y_n)$, of the random variable $Y_n = n^{-1}\sum_{i=1}^n X_i$, where $X_1, X_2, \ldots, X_n$ are sampled i.i.d. from the normal distribution

$$p(x) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

exactly.

Solution. First, recall that the characteristic function of a Gaussian distribution is a Gaussian

$$G(t) = \int\limits_{\mathbb{R}} e^{itx}p(x)dx = \exp\left(i\mu t - \frac{\sigma^2 t^2}{2}\right).$$

Let us now evaluate the characteristic function for $p_n(y_n)$

$$G_n(t) = \int\limits_{\mathbb{R}^n} dx_1 \cdots dx_n \exp\left(i\frac{t}{n}\sum_{i=1}^n x_i\right) p(x_1)\cdots p(x_n)$$

$$\Rightarrow \quad (G(t/n))^n = \exp\left(i\mu t - \frac{\sigma^2 t^2}{2n}\right).$$

The inverse Fourier transform of $G_n(t)$ results in

$$p_n(y_n) = \int_{-\infty}^{+\infty} \frac{dt}{2\pi} G_n(t) e^{-ity_n} = \int_{-\infty}^{+\infty} \frac{dt}{2\pi} \exp\left(-it(y_n - \mu) - n\frac{\sigma^2 t^2}{2}\right)$$
$$= \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{n(y_n - \mu)^2}{2\sigma^2}\right).$$

Example 8.4.5 (Failure of the central limit theorem)**.** Calculate the probability density distribution of the random variable $Y_n = n^{-1}\sum_{i=1}^{n} X_i$, where $X_1, X_2, \ldots, X_n$ are independently chosen from the Cauchy distribution with the following probability density

$$p(x) = \frac{\gamma}{\pi} \frac{1}{x^2 + \gamma^2}, \tag{8.53}$$

and show that the CLT does not hold in this case. Explain why.

Solution. The characteristic function of the Cauchy distribution is

$$G(k) = \frac{\gamma}{\pi} \int_{-\infty}^{+\infty} \frac{dx}{x^2 + \gamma^2} e^{ikx} = e^{-\gamma k}. \tag{8.54}$$

The resulting expressions for the characteristic functional is

$$G_n(k) = (G(k/n))^n = G(k). \tag{8.55}$$

This expression shows that for any n the random variable Y_n is Cauchy-distributed with exactly the same width parameter as the individual samples. The CLT "fails" in this case because we have ignored an important requirement/condition for the CLT to hold – existence of the variance. (See Example 8.2.4.) □

Exercise 8.2. Assume that you play a dice game 100 times. Awards for the game are as follows: \$0.00 for $1, 3$ or 5, \$2.00 for 2 or 4 and \$26.00 for 6.

(1) What is the expected value of your winnings?

(2) What is the standard deviation of your winnings?

(3) Estimate the probability that you win at least 400\$.

Exercise (not graded)**.** Experiment with the CLT for different distributions mentioned in the lecture.

The CLT holds for independent but not necessarily identically distributed variables too. (That is one can use different distributions generating different variables in the summed up sequence.)

We may be interested in not only deviations on the order of the standard deviation but would also like to discuss arbitrary deviations.

Theorem 8.4.6 (Cramér theorem (strong version of the CLT))**.** The normalized sum, $Y_n = \sum_{i=1}^n X_i/n$, of the i.i.d. variables, $X_i \sim p_X(x)$, satisfies

$$\forall x > \mu : \quad \lim_{n\to\infty} \frac{1}{n} \log \text{Prob}\left(Y_n \geq x\right) = -\Phi^*(x), \tag{8.56}$$

$$\Phi^*(x) := \sup_{t\in\mathbb{R}} \left(tx - \Phi(t)\right), \tag{8.57}$$

$$\Phi(t) := \log\left(\mathbb{E}\exp\left(tX\right)\right), \tag{8.58}$$

where, $\Phi(t)$, is the cumulant generating function of $p_X(x)$ and $\Phi^*(x)$ is the Legendre-Fenchel transform of the cumulant generating function, also called the Cramér function.

Three comments are in order. First, an informal ("physical") version of Eq. (8.56) is

$$n \to \infty : \quad \text{Prob}\left(Y_n\right) \propto \exp\left(-n\Phi^*(x)\right). \tag{8.59}$$

Second, the cumulant generating function, $\Phi(t)$, is equal to the characteristic function (8.26) of the minus imaginary argument, i.e., $\Phi(t) = G(-it)$. Also, and third, the weak version of the CLT (8.52) is equivalent to approximating the Cramér function (asymptotically exact) by a Gaussian distribution centered around its minimum.

Exercise (not graded)**.** Prove the strong-CLT (8.56,8.57). [Hint: use saddle point/stationary point method to evaluate the integrals.] Give an example of an expectation for which not only vicinity of the minimum but also other details of $\Phi^*(x)$ are significant at $n \to \infty$? More specifically give an example of the object which behavior is controlled solely by left/right tail of $\Phi^*(x)$? $\Phi^*(0)$ and its vicinity?

Example 8.4.7. Compute the Crámer function for the Bernoulli process, i.e. (generally unfair) coin toss

$$X = \begin{cases} 0, & \text{with probability } 1-\beta; \\ 1, & \text{with probability } \beta. \end{cases} \tag{8.60}$$

Solution.

$$\Phi(t) = \log(\beta e^t + 1 - \beta), \tag{8.61}$$

$$0 < x < 1 : \quad \Phi^*(x) = x \log\frac{x}{\beta} + (1-x)\log\frac{1-x}{1-\beta}. \tag{8.62}$$

Eqs. (8.61,8.62) are noticeable for two reasons. First of all, they lead (after some algebraic manipulations) to the famous Stirling formula for the asymptotic of a factorial

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \left(1 + O(1/n)\right).$$

(Do you see how?) Second, the $x \log x$ structure is an "entropy" which will appear a number of times in following lectures - stay tuned.

The Crámer theorem (8.4.6) gives a powerful asymptotic result, however it says nothing about the rate of convergence, i.e., the behavior at large but finite n. Quite remarkably the deficiency can be cured by the following application of the Chernoff bound (8.3.1).

Theorem 8.4.8 (Chernoff Bound Version of the Central Limit Theorem, adapted from [17])**.** Let $X_1, \cdots, X_n$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and a well-defined (bounded) Crámer function $\Phi^*(x) = \sup_t(tx - \log(\mathbb{E}[\exp(tX_i)])$. Then,

$$P\left(\sum_{i=1}^{n} X_i \geq nx\right) \leq \exp\left(-n\Phi^*(x)\right), \ \forall x > \mu,$$
$$P\left(\sum_{i=1}^{n} X_i \leq nx\right) \leq \exp\left(-n\Phi^*(x)\right), \ \forall x < \mu.$$

Proof. Consider the case of $x > \mu$:

$$\begin{aligned}
P\left(\sum_{i=1}^{n} X_i \geq nx\right) &= P\left(e^{t\sum_{i=1}^{n} X_i} \geq e^{tnx}\right) \quad \text{for any } t > 0 \text{ (to be chosen later)} \\
&\leq e^{-tnx}\mathbb{E}\left[e^{t\sum_{i=1}^{n} X_i}\right] \quad \text{by Markov's inequality (its Chernoff bound's version)} \\
&= e^{-tnx}\prod_{i=1}^{n}\mathbb{E}[e^{tX_i}] = e^{-tnx+n\Phi(t)} \quad \text{since } X_i \text{ are i.i.d.} \\
&\leq \exp\left(-n\sup_{t>0}(tx - \Phi(t))\right) = \exp\left(-n\sup_{t\in\mathbb{R}}(tx - \Phi(t))\right) \\
&= \exp\left(-n\Phi^*(x)\right).
\end{aligned}$$

Here, $\Phi(t) = \log(\mathbb{E}[\exp(tX_i)])$ is convex in t; $\Phi^*(x)$ is convex in x and it achieves its minimum at $x = \mu$. Therefore, for $x > \mu$ the sup in, $\Phi^*(x) = \sup_{t\in\mathbb{R}}(tx - \Phi(t))$, is achieved at $t > 0$ (positive slope) and we can thus replace, $\sup_{t>0}$ by $\sup_{t\in\mathbb{R}}$. This completes the proof for $x > \mu$.

In the case of $x < \mu$, one needs to pick $t < 0$, and otherwise the proof is fully equivalent. □

Exercise 8.3. Let $X = \sum_{i=1}^{n} X_i$, where the X_i are independent (not necessarily identically distributed) Poisson random variables. That is, each X_i is independently drawn from a Poisson distribution with parameter λ_i, i.e. $X_i \sim \text{Pois}(\lambda_i)$. Denote the characteristic function of X by $G_X(t)$ and the characteristic function of X_i by $G_{X_i}(t)$. Show that
(1) $G_X(t) = \prod_{i=1}^{n} G_{X_i}(t)$;
(2) $X \sim \text{Pois}(\lambda)$, where $\lambda = \sum_{i=1}^{n} \lambda_i$.

8.4.3 Bayes Theorem

We already saw how to get conditional probability distribution and marginal probability distribution from the joint probability distribution

$$P(x|y) = \frac{P(x,y)}{P(y)}, \quad P(y|x) = \frac{P(x,y)}{P(x)}. \tag{8.63}$$

Combining the two formulas to exclude the joint probability distribution we arrive at the famous Bayes formula

$$P(x|y)P(y) = P(y|x)P(x). \tag{8.64}$$

Here, in Eqs. (8.63,8.64) both x and y may be multivariate. Rewriting Eq. (8.64) as

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}, \tag{8.65}$$

one often refers (in the field of the so-called Bayesian inference/reconstruction) to $P(x)$ as the "prior" probability distribution which measures the degree of the initial "belief" in X. Then, $P(x|y)$, called the "posterior", measured the degree of the (statistical) dependence of x on y, and the quotient $\frac{P(y|x)}{P(y)}$ represents the "support/knowledge" y provides about x.

A good visual illustration of the notion of the conditional probability can be found at `http://setosa.io/ev/conditional-probability/`.

Example 8.4.9. Consider the three component Ising model. (a) Compute the normalization constant. (b) Compute the marginal probability, $P(x_1)$. (c) Compute the conditional probability, $P(x_3|x_1)$.

Solution. The set of all possible outcomes is shown in Figure 8.10. (a) The normalization constant, Z, is found by summing the probabilities over all the states:

$$Z = e^{2J} + e^0 + e^0 + e^{-2J} + e^{2J} + e^0 + e^0 + e^{-2J} = 4 + 4\cosh(2J).$$

Therefore, the probabilities of the states are, $P(1,1,1) = e^{2J}/(4+4\cosh(2J))$, etc.
(b) The marginal probability, $P(x_1)$, is given by summing all the probabilities corresponding to $P(x_1 = +1)$ and all the probabilities corresponding to $P(x_1 = -1)$:

$$\begin{aligned} P(x_1 = +1) &= P(1,1,1) + P(1,1,-1) + P(1,-1,-1) + P(1,-1,1) \\ &= \frac{e^{2J} + e^0 + e^0 + e^{-2J}}{4 + 4\cosh(2J)} = \frac{1}{2} \\ P(x_1 = -1) &= P(-1,-1,-1) + P(-1,-1,1) + P(-1,1,1) + P(-1,1,-1) \\ &= \frac{e^{2J} + e^0 + e^0 + e^{-2J}}{4 + 4\cosh(2J)} = \frac{1}{2}. \end{aligned}$$

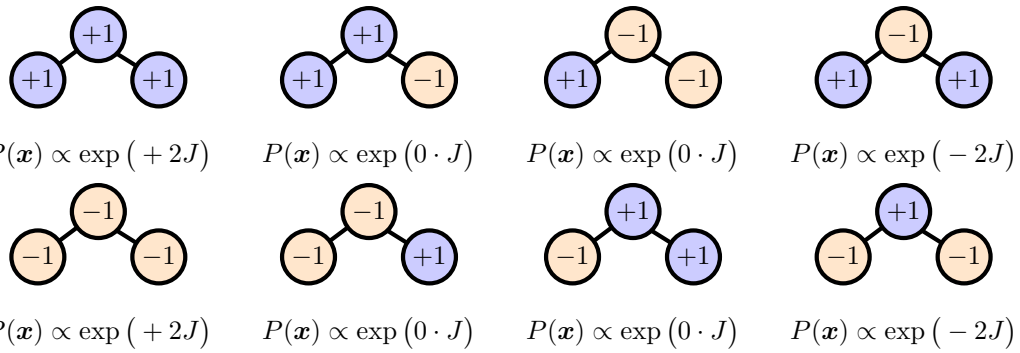


Figure 8.10: Set of all possible outcomes for the 3-component Ising model and their relative probabilities. The normalization constant, Z, is found by summing the probabilities over all states.

The conditional probability is found by

$$P(x_3 = +1|x_1 = +1) = \frac{P(x_3 = +1, x_1 = +1)}{P(x_1 = +1)} = \frac{Z^{-1}(e^{2J} + e^{-2J})}{Z^{-1}(e^{2J} + 1 + 1 + e^{-2J})} = \frac{\cosh(2J)}{1 + \cosh(2J)},$$
$$P(x_3 = +1|x_1 = -1) = \frac{P(x_3 = +1, x_1 = -1)}{P(x_1 = -1)} = \frac{Z^{-1}(e^{0} + e^{0})}{Z^{-1}(e^{2J} + 1 + 1 + e^{-2J})} = \frac{1}{1 + \cosh(2J)},$$
$$P(x_3 = -1|x_1 = +1) = \frac{P(x_3 = -1, x_1 = +1)}{P(x_1 = +1)} = \frac{Z^{-1}(e^{0} + e^{0})}{Z^{-1}(e^{2J} + 1 + 1 + e^{-2J})} = \frac{1}{1 + \cosh(2J)},$$
$$P(x_3 = -1|x_1 = -1) = \frac{P(x_3 = -1, x_1 = -1)}{P(x_1 = -1)} = \frac{Z^{-1}(e^{2J} + e^{-2J})}{Z^{-1}(e^{2J} + 1 + 1 + e^{-2J})} = \frac{\cosh(2J)}{1 + \cosh(2J)}.$$

Exercise 8.4. The joint probability density of two real random variables X_1 and X_2 is

$$\forall x_1, x_2 \in \mathbb{R}: \quad p(x_1, x_2) = \frac{1}{Z} \exp(-x_1^2 - x_1 x_2 - x_2^2).$$

(1) Calculate the normalization constant Z.

(2) Calculate the marginal probability density, $p(x_1)$.

(3) Calculate the conditional probability density, $p(x_1|x_2)$.

8.5 Information-Theoretic View on Randomness

8.5.1 Entropy.

Consider a random variable X that takes outcomes $x \in \mathcal{X}$. The goal is to develop a systematic and meaningful way to quantify the amount of information gained when we learn

that a particular outcome actually occurred. We suppose that the information content of an outcome, x, which we denote $h(x)$ and will also be calling *surprise*, depends only on the probability of the outcome.

The question becomes: how to quantify the information content? Let us start formulating a list of requirements that the information content (surprise) must satisfy:

1. Deterministic outcomes provide no information. If an outcome is certain to occur, then its information content, $h(x)$, must be zero. That is,

$$h(x) = 0 \quad \text{if} \quad P(x) = 1.$$

2. Learning that an unlikely outcome has occurred provides more information than learning that a likely outcome has occurred. The information content of an outcome must be a strictly decreasing function of its probability. That is,

$$h(x_1) > h(x_2) \quad \text{for} \quad P(x_1) < P(x_2).$$

3. Independent events provide original information. If two independent events occur, then the information content of the pair of outcomes must be the sum of the information content of each individual outcome. That is,

$$h(x, y) = h(x) + h(y) \quad \text{provided that} \quad P(x, y) = P(x)P(y).$$

With a little work, it can be shown that only one family of continuous functions satisfies this modest list of requirements. We are forced to define the information content of an outcome to be the log of the probability:

$$h(x) = -\log\big(P(x)\big). \tag{8.66}$$

The base of the logarithm, or equivalently, the multiplicative scaling constant, can be chosen arbitrarily. Convention, which is standard in the information theory, is to use a unit scaling and log base 2, i.e. $\log \to \log_2$ in Eq. (8.66).

Terminology. Standard scientific term used for the information gained by learning the outcome x, which we also called the surprise of x, is the *configurational entropy.*

Consistently with all of the above, but now introducing the *entropy* of all possible outcomes, i.e. entropy of a random variable X, is defined as the expectation of the configurational entropy over the outcomes

$$H(X) = -\mathbb{E}_{P(X)}\big[\log\big(P(X)\big)\big] = \sum_{x\in\mathcal{X}} P(x)h(x) = -\sum_{x\in\mathcal{X}} P(x)\log\big(P(x)\big), \tag{8.67}$$

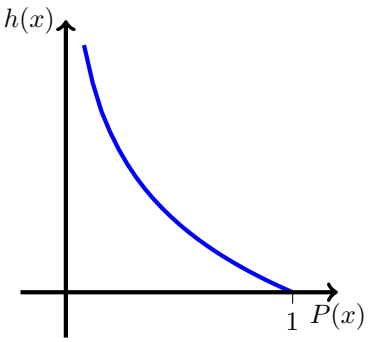


Figure 8.11: The information content $h(x)$ of an outcome x plotted against the probability of x. Negative-logs are the only family of functions that satisfy the requirements of $h(x)$. The base of the logarithm (or equivalently, the multiplicative scaling constant) can be chosen arbitrarily.

where x is drawn from the space $\mathcal{X}$. We can also say that the entropy is a measure of uncertainty. In the case of a deterministic process, i.e. when there is only one outcome with thhe probability 1, the configurational entropy becomes equal to the entropy and according to Eq. (8.67) both are zero, $0 \log 0 = 0$.

Terminology. Yet another term associated with the entropy of a random variable, X, is the *measure of uncertainty*. Following the tradition of information theory, we use the symbol H for entropy. However, be aware that an alternative notation, S, is customary in Statistical Mechanics/Physics.

Let us familiarize ourselves with the concept of entropy on example of the Bernoulli(β) process (8.60). In this case, there are only two states, $P(X = 1) = \beta$ and $P(X = 0) = 1-\beta$, and therefore

$$H = -\beta \log \beta - (1 - \beta) \log(1 - \beta). \tag{8.68}$$

Notice that H, considered as a function of β, is concave and has a global maximum at $\beta = 1/2$ (Fig. 8.12. Therefore, $\beta = 1/2$, corresponding to the fair coin in the process of coin flipping, is the least uncertain case (maximum entropy). If we plot the entropy as the function of β. The entropy is zero at $\beta = 0$ and $\beta = 1$ as both of these cases are deterministic, i.e. fully certain and thus least uncertain. (See accompanied ijulia file.)

The expression for entropy (8.67), has the following properties (some of these can be interpreted as alternative definitions):

- $H \geq 0$

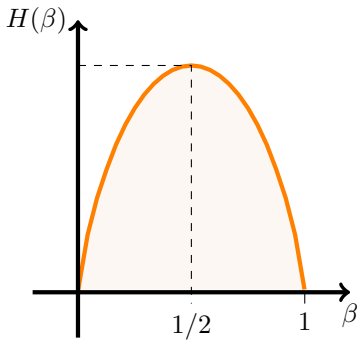


Figure 8.12: The entropy of a Bernoulli random variable as a function of β. The entropy is zero when $\beta = 0$ or $\beta = 1$; this is precisely when one of the outcomes is certain and the random variable is actually deterministic. The entropy is maximized at $\beta = 1/2$; this is precisely when the two outcomes are equi-probable

- $H = 0$ iff the process is deterministic, i.e. $\exists x$ s.t. $P(x) = 1$.
- $H \leq \log(|\mathcal{X}|)$ and $H = \log(|\mathcal{X}|)$ iff x is distributed uniformly over the set $\mathcal{X}$.
- Entropy is the measure of average uncertainty.
- Entropy is less than the average number of bits needed to describe the random variable (the equality is achieved for uniform distribution) [a].
- Entropy is the lower bound on the average length of the shortest description of a random variable

Example 8.5.1. The so called Zipf's law states that the frequency of the n-th most frequent word in randomly chosen English document can be approximated by

$$p_n = \begin{cases} \frac{0.1}{n}, & \text{for } n \in 1, \ldots, 12367 \\ 0, & \text{for } n > 12367 \end{cases} \tag{8.69}$$

Under an assumption that English documents are generated by picking words at random according to Eq. (8.69) compute the entropy of the made-up English per word.

[a] Take integers which are smaller or equal than n, and represent them in the binary system. We will need $\log_2(n)$ binary variables (bits) to represent any of the integers. If all the integers are equally probable then $\log_2(n)$ is exactly the entropy of the distribution. If the random variable is distributed non-uniformly than the entropy is less than the estimate.

Solution. Substituting the distribution (8.69) into the definition of entropy one derives

$$H = -\sum_{i=1}^{12367} \frac{0.1}{n} \log_2 \frac{0.1}{n} \approx \frac{0.1}{\ln 2} \int_{10}^{123670} \frac{\ln x}{x} dx = = \frac{1}{20 \ln 2} (\ln^2 123670 - \ln^2 10) \approx 9.9 \text{ bits.}$$

It is known, from the famous work of Shannon [18], that entropy of English alphabet per character is fairly low, ~ 1 bit. Therefore, the character-based entropy of a typical English text is much smaller than its entropy per word. This result is intuitively clear: after the first few letters one can often guess the rest of the word, but prediction of the next word in the sentence is a less trivial task.

8.5.2 Comparing Probability Distributions: Kullback-Leibler Divergence

The concepts of information content (surprise) and of entropy provide a number of useful tools in probability. One of the most important tools is a method of comparing two probability distributions. For illustration, let X be a random variable taking values $x \in \mathcal{X}$, and let P_1 be the probability distribution of X, which we consider as the ground truth. Assume that P_1 is approximated or modelled by the probability distribution $P_2(x)$, then the difference in the information content of x, as measured by the two probability distributions, is

$$\log\big(P_1(x)\big) - \log\big(P_2(x)\big) \equiv \log\left(\frac{P_1(x)}{P_2(x)}\right)$$

The *Kullback-Leibler (KL) divergence* is defined as the expectation of the difference in the information context between the ground truth and its proxy (approximation) with respect to the probability distribution of the former, P_1,

$$D(P_1 \| P_2) := \sum_{x \in \mathcal{X}} P_1(x) \log \frac{P_1(x)}{P_2(x)}. \tag{8.70}$$

Note that the KL divergence is not symmetric, i.e. $D(P_1\|P_2) \neq D(P_2\|P_1)$. Moreover it is not a proper metric of comparison as it does not satisfy the so-called triangle inequality. A metric, $d(a, b)$, is a function mapping two elements a and b from the same space to $\mathbb{R}$ that satisfies (i) non-negativity, i.e. $d(a, b) \geq 0$, and zero if and only if $a = b$, i.e. $d(a, a) = 0$; (ii) symmetric, i.e. $d(a, b) = d(b, a)$, and (iii) the triangle inequality, $d(a, b) \leq d(a, c) + d(b, c)$.

The last two conditions do not hold in the case of the KL divergence. However, an infinitesimal version of the KL divergence, the Hessian of the KL divergence around its minimum, also called the *Fisher information*, satisfies all the requirements of a metric.

Example 8.5.2. An illusionist has a biased coin that comes up heads 70% of the time. Use the KL divergence to quantify the amount information that would be lost if the biased coin were modeled as a fair coin.

Solution. We regard the biased probability distribution as the 'ground truth', P_1, and the fair probability distribution, P_2, as our approximation. The KL divergence between the two becomes

$$\begin{aligned} D(P_1||P_2) = \mathbb{E}_{P_1}\left[\log\left(P_1/P_2\right)\right] &= \sum_x P_1(x)\log\left(P_1(x)/P_2(x)\right) \\ &= 0.3\log_2(0.3/0.5) + 0.7\log_2(0.7/0.5) \\ &= 0.118. \end{aligned}$$

We lose approximately 0.118 bits of information by modeling the biased coin as a fair coin.

Exercise 8.5. Assume that a random variable X_2 is generated by the known probability distribution $P_2(x)$, where $x \in \mathcal{X}$ and $\mathcal{X}$ is finite. Consider the vector $(P_1(x)|x \in \mathcal{X})$ that satisfies $P_1(x) \geq 0$ for all $x \in \mathcal{X}$ and $\sum_{x\in\mathcal{X}} P_1(x) = 1$. Show that $D(P_1\|P_2)$, as a function of $P_1(x)$, is non-negative and that it achieves its minimum when $P_1(x) = P_2(x)\ \forall x \in \mathcal{X}$, i.e.

$$\arg\min_{(P_1(x)|x\in\mathcal{X})} D(P_1\|P_2)\Big|_{\substack{\sum_{x\in\mathcal{X}} P_1(x) = 1 \\ \forall x\in\mathcal{X}:\quad P_1(x)\geq 0}} = (P_2(x)|x\in\mathcal{X}). \tag{8.71}$$

8.5.3 Joint and Conditional Entropy

The notion of entropy naturally extends to the multivariate statistics. If we have a pair of discrete random variables, X and Y, taken values $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ respectively, their joint entropy is

$$H(X,Y) := -\mathbb{E}\big[\log\big(P(X,Y)\big)\big] = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(x,y)\log\big(P(x,y)\big). \tag{8.72}$$

One must ask whether $H(X,Y) \stackrel{?}{=} H(X)+H(Y)$, that is, one asks whether the expected information in the entire system is equal to the sum of the expected information of X and Y individually. To answer this question, we examine the expected amount information in the system beyond that which can be gained from X, that is, we examine the quantity $H(X,Y) - H(X)$,

$$\begin{aligned} H(X,Y) - H(X) &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(x,y)\log\big(P(x,y)\big) + \sum_{x\in\mathcal{X}} P(x)\log\big(P(x)\big) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(x,y)\log\big(P(x,y)\big) + \sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(x,y)\log\big(P(x)\big) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(x,y)\log\left(\frac{P(x,y)}{P(x)}\right) \end{aligned}$$

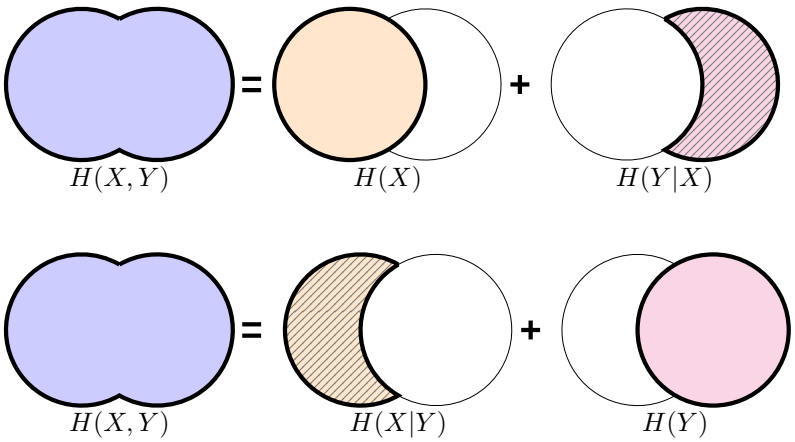


Figure 8.13: Venn diagram illustrating the relationship between entropy, joint entropy and conditional entropy. (It is customary in information theory to use venn diagrams to illustrate entropies, conditional entropies and mutual information. Be advised that the shapes in the diagram do not actually represent sets of objects. See e.g. pp 141 of [19] for a detailed discussion with examples.).

If X and Y are independent, then $P(x, y) = P(x)P(y)$ and the result is $H(Y)$. However, if X and Y are not independent, then $P(x, y)/P(x) = P(y|x)$ by Bayes theorem. We define the conditional entropy $H(Y|X) := H(X, Y) - H(X)$ and observe that it can be computed by

$$H(Y|X) = -\mathbb{E}\big[\log\big(P(Y|X)\big)\big] = -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} P(x, y)\log\big(P(y|x)\big). \tag{8.73}$$

The definitions of the joint and conditional entropies naturally lead to the following relation between the two

$$H(X, Y) = H(X) + H(Y|X), \tag{8.74}$$

called the chain rule (Fig. 8.13).

One can naturally extend the chain rule from the bi-variate to the multi-variate case $(X_1, \cdots, X_n) \sim P(x_1, \cdots, x_n)$ as follows

$$H(X_n, \cdots, X_1) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1). \tag{8.75}$$

Notice, that the choice of the order in the chain is arbitrary.

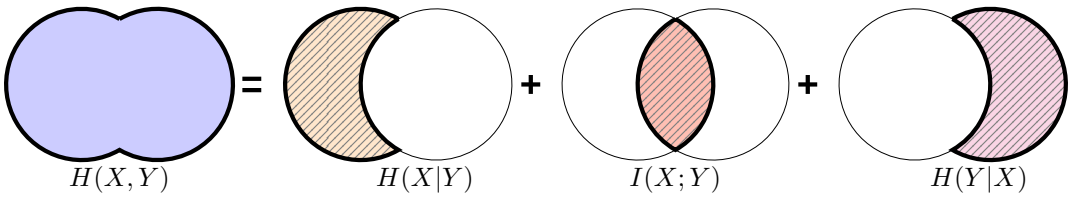


Figure 8.14: Venn diagram explaining relations between the mutual information and respective entropies for two random variables. (It is customary in information theory to use venn diagrams to illustrate entropies, conditional entropies and mutual information. Be advised that the shapes in the diagram do not actually represent sets of objects. See e.g. pp 141 of [19] for a detailed discussion with examples.)

8.5.4 Independence, Dependence, and Mutual Information.

Comparing the two information sources, say tracking events x and y, one assumption, which is rather dramatic, may be that the probabilities are independent, i.e. $P(x,y) = P(x)P(y)$ and then, $P(x|y) = P(x)$ and $P(y|x) = P(y)$. Mutual information, which we are about to discuss, will be zero in this case. Thus, naturally, the mutual information is introduced as the measure of dependence

$$I(X;Y) = \mathbb{E}_{P(x,y)}\left[\log \frac{P(x,y)}{P(x)P(y)}\right] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}. \tag{8.76}$$

Intuitively the mutual information measures the information that X and Y share. In other words, it measures how much knowing one of these random variables reduces uncertainty about the other. For example, if X and Y are independent, then knowing X does not give any information about Y and vice versa - the mutual information is zero. In the other extreme, if X is a deterministic function of Y then all information conveyed by X is shared with Y. In this case the mutual information is the same as the uncertainty contained in X itself (or Y itself), namely the entropy of X (or Y).

The mutual information is obviously related to respective entropies,

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y). \tag{8.77}$$

The relation is illustrated in Fig. (8.14). Mutual Information also possesses the following properties

$$I(X;Y) = I(Y;X) \text{ (symmetry)} \tag{8.78}$$

$$I(X;X) = S(X) \text{ (self-information)} \tag{8.79}$$

The conditional mutual information between two random variables (two sources of information), X and Y, given another random variable, Z, is

$$I(X;Y|Z) := H(X|Z) - H(X|Y,Z) = \mathbb{E}_{P(x,y,z)}\left[\log \frac{P(x,y|z)}{P(x|z)P(y|z)}\right] \tag{8.80}$$

The entropy chain rule (8.74) when applied to the mutual information of $(X_1, \cdots, X_n) \sim P(x_1, \cdots, x_n)$ results in

$$I(X_n, \cdots, X_1; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \cdots, X_1) \tag{8.81}$$

See Fig. (8.14) for the Venn diagram illustration of Eq. (8.81).

We recommend the reader to check [19] for extended discussions on entropy, mutual information and related.

The notions of joint, conditional entropy and mutual information in the context of two random variables is illustrated in the following three examples.

Example 8.5.3. Consider two Bernoulli random variables X and Y with a joint probability mass function $P(X,Y)$ given by

	$y=0$	$y=1$
$x=0$	0	0.2
$x=1$	0.8	0

Compute the entropy of X, the joint entropy of X and Y, and the conditional entropy of Y given X. Discuss the results.

Solution. The joint probability mass function indicates that the outcome of Y is completely determined by the outcome of X, and vice versa. Intuitively, we should expect that all the information in the entire system is fully contained in X, and that once X is known, no additional information can be gained from Y. Lets do the calculations to verify that our intuition is correct. The entropy of X is

$$H(X) = -\sum_{x\in\mathcal{X}} P(x)\, \log_2\big(P(x)\big) = -0.2\log_2\big(0.2\big) - 0.8\log_2\big(0.8\big) = 0.722.$$

The joint entropy of X and Y is

$$\begin{aligned} H(X,Y) &= -\sum_{x,y\in\mathcal{X},\mathcal{Y}} P(x,y)\, \log_2\big(P(x,y)\big) \\ &= -0\log_2(0) - 0.8\log_2\big(0.8\big) - 0.2\log_2\big(0.2\big) - 0\log_2(0) = 0.722. \end{aligned}$$

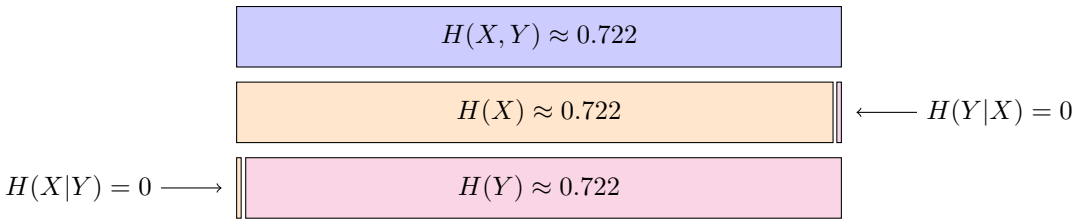


Figure 8.15: Schematic for example 8.5.3. The entropy of the whole system (blue) is the same as the entropy of X (orange). The conditional entropy of Y given X is zero (illustrated by the bar of 'zero' width at the end of the second row). The bottom row shows that the entropy of Y (pink) also coincides with that of the entire system (and, incidentally, is fully shared with that of X).

For this situation, the expected information content of the entire system is exactly the same as the expected information content of X alone. No additional information can be expected from (X, Y) that cannot be expected from X. We anticipate (and verify) that there is no additional information that can be expected from Y once X is known.

$$\begin{aligned} H(Y|X) &= -\sum_{x,y\in\mathcal{X},\mathcal{Y}} P(x,y)\,\log_2\big(P(y|x)\big) \\ &= -0\log_2(0) - 0.8\log_2\big(1\big) - 0.2\log_2\big(1\big) - 0\log_2(0) = 0. \end{aligned}$$

See Fig. 8.15 for illustration. *Comment:* Similar calculations for $H(Y)$ and $H(X|Y)$ would show that Y also contains all the expected information in the system, and that no additional information can be expected from X once Y is known.

Example 8.5.4. Consider two Bernoulli random variables X and Y with a joint probability mass function $P(X, Y)$ given by

	$y=0$	$y=1$
$x=0$	0.45	0.45
$x=1$	0.05	0.05

Compute the entropies of X and of Y. Compute the joint entropy of X and Y. Compute the conditional entropy of Y given X. Discuss the results.

Solution. The marginal distributions are:

$$P(X=x) = \begin{cases} 0.9, & x=0 \\ 0.1, & x=1 \end{cases}; \qquad \text{and} \qquad P(Y=y) = \begin{cases} 0.5, & y=0 \\ 0.5, & y=1 \end{cases}.$$

We observe that X and Y are independent, so we should intuitively expect that there is no 'overlap' in the information in X and Y. Let's do the calculations to verify that our intuition is correct. The entropy of X is

$$H(X) = -\sum_{x\in\mathcal{X}} P(x)\,\log_2\big(P(x)\big) = -0.9\log_2\big(0.9\big) - 0.1\log_2\big(0.1\big) = 0.469.$$

The entropy of Y is

$$H(Y) = \sum_{y\in\mathcal{Y}} P(y)\,\log_2\big(P(y)\big) = 0.5\log_2\big(0.5\big) + 0.5\log_2\big(0.5\big) = 1.0.$$

The joint entropy of X and Y is

$$\begin{aligned} H(X,Y) &= -\sum_{x,y\in\mathcal{X},\mathcal{Y}} P(x,y)\,\log_2\big(P(x,y)\big) \\ &= -0.45\log_2(0.45) - 0.45\log_2\big(0.45\big) - 0.05\log_2\big(0.05\big) - 0.05\log_2(0.05) = 1.469. \end{aligned}$$

For this situation, the expected information content of the entire system is more than the expected information content of X alone. The additional expected information of the system (X,Y) beyond that of X must be information expected from Y that is not contained in X. We anticipate (and verify) that the additional information that can be expected from Y when X is known is non-zero.

$$\begin{aligned} H(Y|X) &= -\sum_{x,y\in\mathcal{X},\mathcal{Y}} P(x,y)\,\log_2\big(P(y|x)\big) \\ &= -0.45\log_2(0.5) - 0.45\log_2\big(0.5\big) - 0.05\log_2(0.5) - 0.05\log_2\big(0.5\big) = 1.0. \end{aligned}$$

Furthermore, X and Y are independent so X actually contains no information about Y. We anticipate (and verify) that all the expected information content of Y will contribute to the expected information of the entire system.

$$\begin{aligned} H(Y) &= \sum_{y\in\mathcal{Y}} P(y)\,\log_2\big(P(y)\big) \\ &= 0.5\log_2\big(0.5\big) + 0.5\log_2\big(0.5\big) = 1.0. \end{aligned}$$

Performing similar calculations for $H(Y)$ and $H(X|Y)$ would show that Y also contains all the information content in the system, and that no additional information is gained by learning X once Y is known. See Fig. 8.15 for illustration.

Example 8.5.5. Consider two Bernoulli random variables X and Y with a joint probability mass function $P(X,Y)$ given by

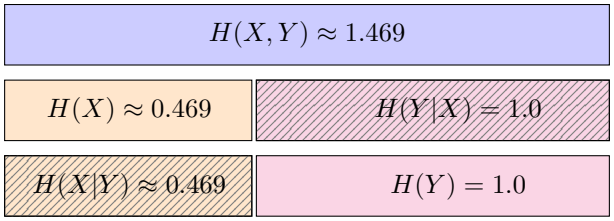


Figure 8.16: Schematic for example 8.5.4. The entropy of the whole system (blue) is equal to the entropy of X (orange) plus the entropy of Y conditioned on X (pink, shaded). Observe that in this example, the entropy of Y conditioned on X is equal to the entropy of Y(pink) because the there is no overlap in information between X and Y.

	$y=0$	$y=1$
$x=0$	0.2	0.3
$x=1$	0	0.5

Compute the entropies of X and of Y. Compute the joint entropy of X and Y. Compute the conditional entropy of Y given X. Discuss the results.

Solution. The entropy of X is

$$\begin{aligned} H(X) &= \sum_x P(x) \log\big(P(x)\big) \\ &= -0.5 \log_2(0.5) - 0.5 \log_2(0.5) = 1.0. \end{aligned}$$

The conditional entropy of Y given X is

$$\begin{aligned} H(Y|X) &= -\sum_{x,y} P(x,y) \log\big(P(y|x)\big) \\ &= -0.2 \log_2(0.4) - 0.4 \log_2(0.6) - 0 \cdot \log_2(0) - 0.4 \log_2(1) = 0.485. \end{aligned}$$

The entropy of Y is

$$H(Y) = \sum_y P(y) \log\big(P(y)\big) = -0.2 \log_2(0.2) - 0.8 \log_2(0.8) = 0.722.$$

The conditional entropy of X given Y is

$$\begin{aligned} H(X|Y) &= -\sum_{x,y} P(x,y) \log\big(P(x|y)\big) \\ &= -0.2 \log_2(1) - 0.4 \log_2(0.375) - 0 \cdot \log_2(0) - 0.4 \log_2(0.625) = 0.764. \end{aligned}$$

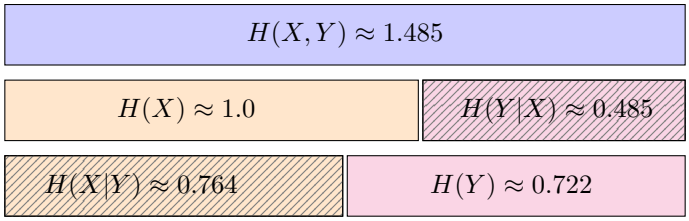


Figure 8.17: Schematic for example 8.5.5. The entropy of the whole system (blue)is equal to the entropy of X (orange) plus the entropy of Y conditioned on X (pink, shaded). Observe that in this example, the entropy of Y conditioned on X is less than the entropy of Y (pink) because the some of the information content in Y overlaps with that of X.

$P(x,y)$		X				$P(y)$
		x_1	x_2	x_3	x_4	
	y_1	1/8	1/16	1/32	1/32	1/4
Y	y_2	1/16	1/8	1/32	1/32	1/4
	y_3	1/16	1/16	1/16	1/16	1/4
	y_4	1/4	0	0	0	1/4
$P(x)$		1/2	1/4	1/8	1/8	

Table 8.1: Exemplary joint probability distribution function $P(x,y)$ and the marginal probability distributions, $P(x)$, $P(y)$, of the random variables x and y.

The joint entropy is

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} P(x,y) \log\big(P(x,y)\big) \\ &= -0.2\log_2(0.2) - 0.4\log_2(0.4) - 0\cdot\log(0) - 0.4\log_2(0.4) = 1.522. \end{aligned}$$

See Fig. 8.17 for illustration. *Comment:* In this example, X and Y are not independent, and therefore some information is shared between the two. This explains why the joint entropy is less than the the sum of the individual entropies, i.e. $H(X,Y) < H(X)+H(Y)$. This also explains why the information content of X conditioned on Y is less than the information content of X alone, i.e. $H(X|Y) < H(X)$. (Similarly, it explains why $H(Y|X) < H(Y)$.)

Exercise 8.6. The joint probability distribution $P(x,y)$ of two random variables X and Y is described in Table 8.1. Calculate the conditional probabilities $P(x|y)$ and $P(y|x)$, marginal entropies $H(X)$ and $H(Y)$, as well as the mutual information $I(X;Y)$.

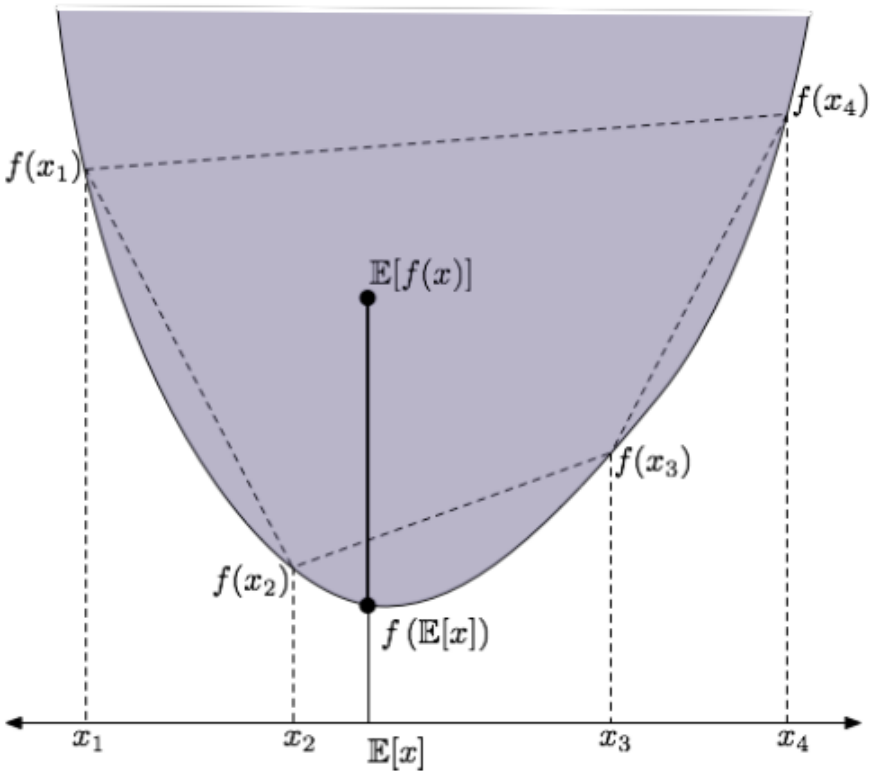


Figure 8.18

8.5.5 Probabilistic Inequalities for Entropy and Mutual Information

Let us now discuss the case when a random one dimensional variable, X, is drawn from the space of reals, $x \in \mathbb{R}$, with the probability density, $p(x)$. Now consider averaging a convex function of X, $f(X)$. One observes that the following statement, called Jensen inequality, holds

$$\mathbb{E}\left[f(X)\right] \geq f\left(\mathbb{E}\left[X\right]\right). \tag{8.82}$$

Obviously the statement becomes equality when $p(x) = \delta(x)$. To gain a bit more of intuition consider the case of the Bernoulli-like distribution, $p(x) = \beta\delta(x - x_1) + (1 - \beta)\delta(x - x_0)$. We derive

$$f\left(\mathbb{E}[X]\right) = f\left(x_1\beta + x_0(1-\beta)\right) \leq \beta f(x_1) + (1-\beta)f(x_0) = \mathbb{E}\left[f(X)\right], \tag{8.83}$$

where the critical inequality in the middle is simply expression of the function $f(x)$ convexity (taken verbatim from the definition).

See also Fig. (8.18) with another (graphical) hint on the proof of the Jensen inequality.

In fact, the Jensen inequality holds over any spaces.

Notice that the entropy, considered as a function (or functional in the continuous case) of probabilities at a particular state is convex. This observation gives rise to multiple consequences of the Jensen inequality (for the entropy and the mutual information):

- (Information Inequality)

$$D(p\|q) \geq 0, \quad \text{with equality iff} \quad p = q$$

- (conditioning reduces entropy)

$$H(X|Y) \leq H(X) \quad \text{with equality iff } X \text{ and } Y \text{ are independent}$$

- (Independence Bound on Entropy)

$$H(X_1, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_i) \quad \text{with equality iff } X_i \text{are independent}$$

Another useful inequality [Log-Sum Theorem]

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}, \tag{8.84}$$

with equality iff a_i/b_i is constant. Convention: $0 \log 0 = 0$, $a \log(a/0) = \infty$ if $a > 0$ and $0 \log 0/0 = 0$. Consequences of the Log-Sum theorem

- (Convexity of Relative Entropy) $D(p\|q)$ is convex in the pair p and q
- (Concavity of Entropy) For $X \sim p(x)$ we have $H(P) := H_P(X)$ (notations are extended) is a concave function of $P(x)$.
- (Concavity of the mutual information in $P(x)$) Let $(X, Y) \sim P(x, y) = P(x)P(y|x)$. Then $I(X;Y)$ is a concave function of $P(x)$ for fixed $P(y|x)$.
- (Concavity of the mutual information in $P(y|x)$) Let $(X, Y) \sim P(x, y) = P(x)P(y|x)$. Then $I(X;Y)$ is a concave function of $P(y|x)$ for fixed $P(x)$.

We will see later (discussing Graphical Models) why the convexity/concavity properties of the entropy-related objects are useful.

Example 8.5.6. Prove that $H(X) \leq \log_2 n$, where n is the number of possible values of the random variable $x \in X$.

Solution. The simplest proof is via the Jensen's inequality. It states that if f is a convex function and U is a random variable then

$$\mathbf{E}[f(U)] \geq f(\mathbf{E}[U]). \tag{8.85}$$

Let us define

$$f(u) = -\log_2 u, \quad u = 1/P(x)$$

Obviously, $f(u)$ is convex. In accordance with (8.85) one obtains

$$\mathbf{E}[\log_2 P(X)] \geq -\log_2 \mathbf{E}[1/P(X)],$$

where $\mathbf{E}[\log_2 P(X)] = -H(X)$ and $\mathbf{E}[1/P(X)] = n$, so $H(X) \leq \log_2 n$.

Note, in passing, that the Jensen's inequality leads to a number of other useful expressions for entropy, e.g. $H(X|Y) \leq H(X)$ with equality iff X and Y are independent, and more generally, $H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$ with equality iff all X_i are independent.

Chapter 9

Stochastic Processes

In Chapter 8, we discussed random vectors and their probability distributions (for example $\boldsymbol{X} = (X_1, \ldots, X_N)$ with $\boldsymbol{X} \sim P_{\boldsymbol{X}}$). In Chapter 9, we will discuss stochastic processes, which are a natural extension of our inquiry into random vectors. A *stochastic process* is a collection of random variables, often written as a path, or sequence, $\{X_t | t = 1, \cdots, T\}$, or simply $(X_t)_{t=1}^{T}$, whose components X_t take values from the same state space Σ. Typically we think of t as time, and we consider the cases where T is finite, $T < \infty$, and where it is infinite, $T \to \infty$. Both the state space, Σ, and the index set, t, may be either discrete or continuous.

We will discuss three basic examples: (a) of Bernoulli process, which is space-time discrete, (b) Poisson process, which is space discrete but time continuous, and (c) continuous space and continuous time process described by Langevien equation, which is an example of the so-called Stochastic Differential Equation (SDE), in Section 9.1, Section 9.2 and Section 9.3, respectively.

The three basic examples of the stochastic processes can also be classified in terms of the amount of memory needed to generate them.

In our first two examples (Bernoulli processes and Poisson processes), the random variables X_t are independent, i.e. memory-less, meaning that the outcome of each X_t does not influence, and is not influenced by, the outcomes of any of the other X_t. In general, however, the components X_t within the path $\{X(t)\}$, need not be independent. Thus, stochastic process described by the Langevien equation results in dependent, i.e. correlated in time X_t.

Time-correlations within a random path, $\{X(t)\}$, described by an SDE may be complicated and difficult to analyze. Consequently, one often considers a discrete-time simplification, called a *Markov process*, discussed in Section 9.4, where the memory holds only for a

single time step, i.e. X_t depends only on the outcome of the previous step, X_{t-1}.

We conclude this Chapter with a brief discussion in Section 9.5 of the Markov Decision Process (MDP), which is controlled formulation involving (conditioned to) Markov process. (Queuing theory, discussed in Section 9.6, is a bonus material.)

9.1 Bernoulli Process (Discrete Space, Discrete Time)

A *Bernoulli process* is a sequence of independent Bernoulli random variables that are often called events or trials. For the case where each event can take only one of two outcomes, say "success" or "failure", then a typical sample path of a Bernoulli process may look like $**S*S*S***S$, where S here stands for "success", or equivalently 00101010001. We will discuss only stationary Bernoulli processes, meaning that the probability of success is the same for each X_t, that is $P(\text{success}) = P(X_t = 1) = \beta$ and $P(\text{failure}) = P(X_t = 0) = 1 - \beta$ for each t.

Examples of processes that can be modeled by a Bernoulli process include the number "arrivals" when checked at fixed intervals, such as the "arrival" of a monsoon on each day of a Tucson summer, or any sequence of discrete updates, such as the (random) ups and downs of the stock market.

9.1.1 Probability distribution of the total number of successes

As we discussed in Eq. 8.5, the number of successes k, in n trials follows the binomial distribution

$$\forall k = 0, \cdots, n: \quad P(S = k|n, \beta) = \left(\begin{array}{c} n \\ k \end{array} \right) \beta^k (1-\beta)^{n-k}, \tag{9.1}$$

The mean and variance are found by computing $\mathbb{E}[S]$ and $\mathbb{E}[(S - \mathbb{E}[S])^2]$ respectively:

$$\text{mean}: \qquad \mathbb{E}[S] = n\beta, \tag{9.2}$$

$$\text{variance}: \qquad var(S) = \mathbb{E}[(S - \mathbb{E}[S])^2] = n\beta(1-\beta). \tag{9.3}$$

9.1.2 Probability distribution of the 1^{st} success

Let T_1 be the number of trials until the first success (including the success event too). The Probability Mass Function (PMF) for the time of the first success is the product of the probabilities of $(t-1)$ failures and one success:

$$t = 1, 2, \cdots: \quad P(T_1 = t|\beta) = \beta(1-\beta)^{t-1} \quad \text{[Geometric PMF]} \tag{9.4}$$

The distribution in Eq. 9.4 is called a geometric distribution because the calculation to verify that the probability distribution is normalized involves summing up the geometric sequence: $\sum_{t=1}^{\infty} \beta(1-\beta)^{t-1} = \beta(1-(1-\beta))^{-1} = 1$. The mean and variance of the geometric distribution are

$$\text{mean :} \qquad \mathbb{E}[T_1] = \frac{1}{\beta}, \tag{9.5}$$

$$\text{variance :} \qquad var(T_1) = \mathbb{E}[(T_1 - \mathbb{E}[T_1])^2] = \frac{1-\beta}{\beta^2}. \tag{9.6}$$

The Bernoulli process is memoryless, meaning that each outcome is independent of the past. If n trials have already occurred, the future sequence $x_{n+1}, x_{n+2}, \cdots$ is also a Bernoulli process and it is independent of the first n trials. Moreover, suppose we have observed the process for n times and no success has occurred. Then the PMF for the remaining arrival times is also geometric,

$$P(T - n = k | T > n, \beta) = \beta(1-\beta)^{k-1}. \tag{9.7}$$

9.1.3 Probability distribution of the k^{th} success

What about the k^{th} arrival? Let T_k be the number of trials until k^{th} success (inclusive), then we write

$$t = k, k+1, \cdots : \quad P(T_k = t|\beta) = \begin{pmatrix} t-1, \\ k-1 \end{pmatrix} \beta^k (1-\beta)^{t-k} \quad \text{[Pascal PMF]}, \tag{9.8}$$

The mean and variance are found by computing $\mathbb{E}[T_k]$ and $\mathbb{E}[(T_k - \mathbb{E}[T_k])^2]$ respectively:

$$\text{mean :} \qquad \mathbb{E}[T_k] = \frac{k}{\beta}, \tag{9.9}$$

$$\text{variance :} \qquad var(T_k) = \mathbb{E}[(T_k - \mathbb{E}[T_k])^2] = \frac{k(1-\beta)}{\beta^2}. \tag{9.10}$$

The combinatorial factor accounts for the number of configurations of k arrivals in T_k trials.

Exercise 9.1. Define $\tau_k = T_k - T_{k-1}, \quad k = 2, 3, \cdots$, so that τ_k is the inter-arrival time between the $(k-1)^{\text{st}}$ and k^{th} arrivals. Write down the probability density distribution function for the k^{th} inter-arrival time, τ_k.

The following two sections discuss two different continuous-time limits of Bernoulli processes, namely Poisson Processes and Brownian motion.

9.2 Poisson Process (Discrete Space, Continuous Time)

We will build on the material of the previous Section, where a discrete-time, discrete-space Bernoulli process was discussed, and extend it to continuous time, thus arriving at *Poisson process*.

Formally, a Poisson process is defined in terms of a collection of random variables, $\{N_t\}$, indexed by time, that count the number of independent 'arrivals' on the interval $[0, t]$. Recall the Poisson distribution which was defined in Section 8.1.1:

$$\forall k \in \{0, 1, 2, \dots\} : \quad \mathrm{P}(N = k; \tilde{\lambda}) = \frac{\tilde{\lambda}^k e^{-\tilde{\lambda}}}{k!}. \tag{9.11}$$

In the following derivation, we will show that the outcomes of a Poisson process follow a Poisson distribution with parameter λt.

The distribution for the Poisson Process can be derived by subdividing the interval $[0, t]$ into n subintervals of length $\Delta t := t/n$. For sufficiently small Δt, the probability of two or more arrivals on any subinterval is negligible and the occurrence of arrivals on any two subintervals are independent. Under these conditions, the probability of k arrivals in n sub-intervals can be modeled by a binomial distribution. The probability of an arrival, β, is proportional to the length of the sub-interval: $\beta \propto t/n$, or equivalently $\beta = \lambda t/n$ where λ is the constant of proportionality. In the limit as $n \to \infty$, we get

$$P(N_t = k; \lambda) = \lim_{n\to\infty} \binom{k}{n} \beta^k (1-\beta)^{n-k} = \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} \tag{9.12}$$

$$= \lim_{n\to\infty} \frac{n^k + O(n^{k-1})}{k!} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} \tag{9.13}$$

$$= \frac{(\lambda t)^k e^{-\lambda t}}{k!}. \tag{9.14}$$

where we have used the facts that $(1 - \frac{\lambda t}{n})^n \to e^{-\lambda t}$ and $(1 - \frac{\lambda t}{n})^k \to 1$. Notice that the dimensionless parameter $\tilde{\lambda}$ in Eq. (9.11) is replaced by λt in Eq. (9.14). Hence λ has the dimension of inverse time: $[\lambda] = [1/t]$.

Common examples of Poisson processes:

- E-mail arrivals with infrequent check.
- Collision of high-energy beams a high frequency (10 MHz) where there is a small chance of actual collision.
- Radioactive decay of a nucleus with the trial being to observe a decay within a small time interval.

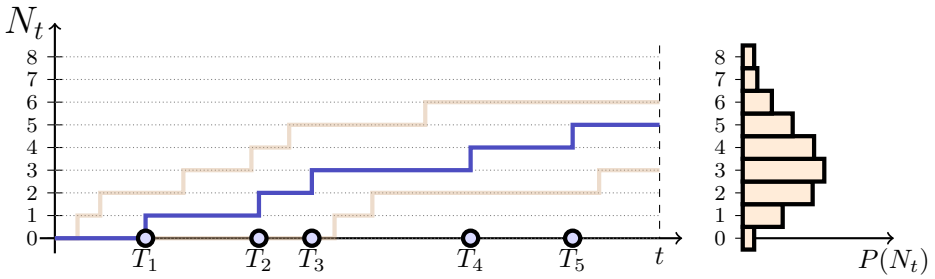


Figure 9.1: One realization of a Poisson process over the interval $[0, t]$ (blue) in which five "arrivals" occurred at the random times $T_1, \ldots, T_5$, therefore $N(t) = 5$. Two other realizations are shown (grey) in which 6 arrivals and 3 arrivals occurred on $[0, t]$. The random variable N_t follows a Poisson distribution (right) with parameter λt where λ is a fixed parameter.

- Spin flips in a magnetic field.

Properties of Poisson Processes

The Poisson process has the following key properties:

- **Initialization**: No arrivals have occurred at $t = 0$, that is, $N(0) = 0$.
- **Independence**: The number of arrivals that occur on two time intervals are independent if on only if the two time intervals are disjoint.
- **Distribution**: The number of arrivals that occur on an interval depends only on the length of the interval and not its location. In particular, in the limit as $\Delta t \to 0$, $P(N(\Delta t) = 1) \to \lambda \Delta t$ and $P(N(\Delta t) \geq 2) = 0$.

With a small amount of effort, it can be shown that these three properties are both necessary and sufficient conditions to define a stochastic process whose components follow a Poisson distribution with parameter λt.

A summary of the relationship between the Bernoulli process and the Poisson process is given in table 9.1.

Probability Distributions of the 1^{st} and the k^{th} Arrival Times

Let T_1 be the (random) time of the first arrival. The probability density of T_1 per unit time can be found from Eq. (9.11) by (1) recalling that a PDF is the derivative of the CDF, and (2) recognizing the equivalency $P(T_1 < t) = P(N_t \geq 1)$ since the event that first arrival

	Bernoulli	Poisson
Times of Arrival	Discrete	Continuous
Arrival Rate	p/per trial	λ/unit time
PMF of Number of arrivals	Binomial	Poisson
PMF of Interarrival Time	Geometric	Exponential
PMF of k^{th} Arrival Time	Pascal	Erlang

Table 9.1: Comparison between the Bernoulli process and the Poisson process

occurs before time t is equivalent to the event that the number of arrivals by time t is greater than or equal to 1. Therefore,

$$\begin{aligned} P(T_1 = t) &= \lim_{\Delta t \to 0} \tfrac{1}{\Delta t}\big(P(T_1 < t + \Delta t) - P(T_1 < t)\big) = \tfrac{d}{dt} P(T_1 < t) \\ &= \tfrac{d}{dt} P(N_t \geq 1) = \tfrac{d}{dt}\big(1 - P(N_t = 0)\big) \\ &= \tfrac{d}{dt}\big(1 - e^{-\lambda t}\big) = \lambda e^{-\lambda t} \end{aligned}$$

Hence the time of the first arrival follows an exponential distribution with parameter λ.

The three properties above imply that, like the Bernoulli process, the Poisson process is *Memoryless* and that it has *Fresh Starts*.

- Memoryless: if we observe the process for t seconds and no arrival has occurred, then the density of the remaining time of arrival is exponential.
- Fresh Starts: the time of the next arrival is independent of the past, and hence is also exponentially distributed with parameter λ.

The probability density that the first arrival occurs before time t can therefore be found by integration:

$$P(T_1 \leq t) = \int_0^t dt' p_{T_1}(t') = \int_0^t dt' \lambda e^{-\lambda t} = 1 - \exp(-\lambda t).$$

By extension, the probability density of the time of the k^{th} arrival, one derives

$$p(T_k = t; \lambda) = \frac{\lambda^k t^{k-1} \exp(-\lambda t)}{(k-1)!}, \quad t > 0 \quad \text{(Erlang "of order" k)}.$$

Merging and Splitting Processes

One of the most important features shared by the Bernoulli and Poisson processes is their invariance with respect to mixing and splitting. We will show it on the example of the Poisson process but the same applies to Bernoulli process.

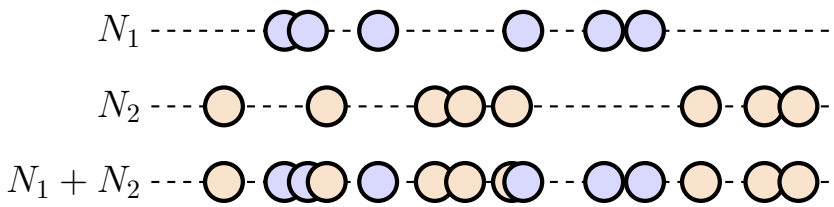


Figure 9.2: Merging and Splitting Poisson Processes

Merging: Let $N_1(t)$ and $N_2(t)$ be two independent Poisson processes with rates λ_1 and λ_2 respectively. Let us define $N(t) = N_1(t) + N_2(t)$. This random process is derived combining the arrivals as shown in Fig. (9.2). The claim is that $N(t)$ is the Poisson process with the rate $\lambda_1 + \lambda_2$. To see it we first note that $N(0) = N_1(0) + N_2(0) = 0$. Next, since $N_1(t)$ and $N_2(t)$ are independent and have independent increments their sum also has independent increments. Finally, consider an interval of length τ, $(t, t + \tau]$. Then the number of arrivals in the interval are Poisson($\lambda_1\tau$) and Poisson($\lambda_2\tau$) and the two numbers are independent. Therefore the number of arrivals in the interval associated with $N(t)$ is Poisson($(\lambda_1 + \lambda_2)\tau$) - as sum of two independent Poisson random variables. We can obviously generalize the statement to a sum of many Poisson processes. Note that in the case of the Bernoulli process the story is identical provided that collision is counted as one arrival.

Splitting: Let $N(t)$ be a Poisson process with rate λ. Here, we split $N(t)$ into $N_1(t)$ and $N_2(t)$ where the splitting is decided by coin tossing (Bernoulli process) - when an arrival occur we toss a coin and with probability β and $1-\beta$ add arrival to N_1 and N_2 respectively. The coin tosses are independent of each other and are independent of $N(t)$. Then, the following statements can be made

- N_1 is a Poisson process with rate $\lambda\beta$.
- N_2 is a Poisson process with rate $\lambda(1 - \beta)$.
- N_1 and N_2 are independent, thus Poisson.

Example 9.2.1. Astronomers estimate that the meteors above a certain size hit the earth on average once every 1000 years, and that the number of meteor hits follows a Poisson distribution.

(a) What is the probability to observe at least one large meteor next year?

(b) What is the probability of observing no meteor hits within the next 1000 years?

(c) Calculate the probability density $p(T_k)$, where the random variable T_k represents the appearance time of the k^{th} meteor.

Solution. The probability of observing k meteors in a time interval $[0, t]$ is given by

$$P(k|t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \tag{9.15}$$

where $\lambda = 0.001$ (events per year) is the average hitting rate and we simplify notations, $P(k|t, \lambda) \to P(k|t)$.

(a) $P(k > 0 \text{ meteors next year}) = 1 - P(0|1) = 1 - e^{-0.001} \approx 0.001$.

(b) $P(k = 0 \text{ meteors next 1000 years}) = P(0|1000) = e^{-1} \approx 0.37$.

(c) It is intuitively clear that

(probability that $T_k > t$) = (probability to get $k - 1$ arrivals in interval $[0, t]$),

Therefore

$$\int_t^\infty p(T_k) dT_k = P(k-1|t),$$
$$p(T_k) = \frac{\lambda^k T_k^{k-1}}{(k-1)!} e^{-\lambda T_k}.$$

□

Exercise 9.2. Customers arrive at a store with the Poisson rate of 5 per hour. 40%/60% of arrivals are men/women.

(a) Compute probability that at least 10 customers have entered between 10 and 11 am.

(b) Compute probability that exactly 5 women entered between 10 and 11 am.

(c) Compute the expected inter-arrival time of men.

(d) Compute probability that no men arrive between 2 and 4 pm.

9.3 Stochastic Processes that are Continuous in Space-time

The stochastic processes discussed so far were memory-less

In this Section we discuss the stochastic dynamics of continuous variables governed by the Langevin equation. We discuss how to derive the Fokker-Planck equations which describe the temporal evolution of the probability of a state. We then go into some additional detail for the foundational example of stochastic dynamics in free space (no drift) that describe Brownian motion where the Fokker-Planck equations simplify to the diffusion equation.

9.3.1 Random Walks on the Integers

For a binomial process $\boldsymbol{Y} = (Y_1, \dots Y_n)$ that takes outcomes ± 1 with equal probability, the stochastic process $\boldsymbol{X}$ where $X_j := \sum_{i=1}^{j} Y_i$ is a called a *random walk* on the integers. The PMF of this random walk is a binomial distribution. The PMF converges to a Gaussian distribution with mean zero and variance $\sqrt{n}$, which is a direct result of the central limit theorem.

Example 9.3.1. A grasshopper is dropped on a number-line and proceeds to take a random walk by making either a unit jump to the right with probability β or a unit jump to the left with probability $1 - \beta$. The starting location of the grasshopper is random, and is $X_0 = 0$ with probability 0.7, and $X_0 = 5$ with probability 0.3. Find the probability distribution of the grasshopper at time t.

Solution. Begin by examining the case where $X_0 = 0$. The possible outcomes after one jump either $X_1 = -1$ (by jumping left) or $X_1 = +1$ (by jumping right). The possible outcomes after two jumps are $X_2 = -2$ (by making two jumps to the left), $X_2 = 0$ (by either jumping left then right, or right then left), or $X_2 = +2$ (by making two jumps to the right). Define the function $F(n, k)$ to be the number of possible combinations of n left and right jumps ending at $X_n = k$. Specifically $F(n, k) = \binom{n}{(n+k)/2}$ if $(n + k)/2$ is an integer between 0 and n, and $F(n, k) = 0$ otherwise. Hence, the probability of jumping to k in n steps from $k = 0$ is $F(n, k)\beta^k(1 - \beta)^{n-k}$. Repeating for $X_0 = 5$ and applying Bayes theorem gives:

$$\begin{aligned} P(X_n = k) &= P(X_n = k|X_0 = 0)P(X_0 = 0) + P(X_n = k|X_0 = 5)P(X_0 = 5) \\ &= 0.7F(n, k)\beta^k(1 - \beta)^{n-k} + 0.3F(n, k - 5)\beta^k(1 - \beta)^{n-k} \end{aligned}$$

Exercise. Modify example 9.3.1 for the random starting location $P(X_0 = x) = 2^{-x}/3$.

9.3.2 From Random Walks to Brownian Motion

Example 9.3.2. Consider a random walk on the line with n jumps that occur at times $t = \Delta, 2\Delta, \dots, n\Delta$, where $\Delta t = 1/n$. Find the PMF of the random walk if the jumps are $\pm\sqrt{\Delta}$ with equal probability, so $P(X_{j+1} = X_j + \sqrt{\Delta}) = P(X_{j+1} = X_j - \sqrt{\Delta}) = 1/2$, and use your result to show the stochastic process has mean zero and unit variance when $t = 1$.

Solution. $P(X_n = (n-2k)\sqrt{\Delta}) = \binom{n}{k}\left(\frac{1}{2}\right)^n$. The mean and variance are found by recognizing that $X_n = \sqrt{\Delta}(2Y - n)$, where $Y \sim \mathrm{B}(n, 1/2)$. Therefore, $E[X_n] = \sqrt{\Delta}(2E[Y] - n) = 0$ and $\mathrm{Var}[X_n] = 4\Delta\mathrm{Var}[Y] = 1$.

Example 9.3.2 could be further extended by using the Central Limit Theorem to show that the CDF of the random walk converges to the CDF of a standard normal distribution in

the limit as $n \to \infty$. This result informs when and how to approximate a high-dimensional discrete-time stochastic processes by a continuous-time stochastic process and vice versa. (Solutions to continuous-time are often easier to analyze and solutions to discrete stochastic processes are often easier to compute numerically.)

The central limit theorem further implies that the jumps need not be Bernoulli processes—every random walk with steps i.i.d. random variables will have the same limit, provided that the variance of the random variables scales as $1/\sqrt{n}$ as $n \to \infty$. This limit is Brownian Motion.

9.3.3 Langevin equation in continuous time and discrete time

Many stochastic processes in 1d can be described in the continuous-time and discrete-time forms as follows

$$\dot{x} = v(x) + \sqrt{2D}\xi(t), \quad \langle \xi(t) \rangle = 0, \quad \langle \xi(t_1)\xi(t_2) \rangle = \delta(t_1 - t_2) \tag{9.16}$$

$$x_{n+1} - x_n = \Delta v(x_n) + \sqrt{2D\Delta}\xi(t_n), \quad \langle \xi(t_n) \rangle = 0, \quad \langle \xi(t_n)\xi(t_k) \rangle = \delta_{kn}. \tag{9.17}$$

The first on the rhs of the Stochastic Differential Equation (SDE) (9.16) determines the (deterministic) drift and the second term (called the *Langevin term*) determines the random "noise", which has mean zero and variance determined by D. The noise is considered independent at each time step. These equations, also called the *Langevin equations*, describe the evolution of a "particle" positioned at $x \in \mathbb{R}$. The two terms on the rhs of Eq. (9.16) correspond to deterministic drift/advancement of the particle (also dependent on its position at the previous time step) and, respectively, on a random correction/increment. The random correction models uncertainty of the environment the particles moves through. (We can also think of it as representing random kicks by other "invisible" particles). The uncertainty is represented in a probabilistic way – therefore we will be talking about the probability distribution function of paths, i.e. trajectories of the particle.

The square root on the rhs of Eq. (9.17) may seem mysterious, let us clarify its origin on the basic "no (deterministic) drift" example of $v(x) = 0$. (This will be the running example through out this lecture.) In this case the Langevin equation describes the Brownian motion. Direct integration of the linear equation with the inhomogeneous source results in

$$\forall t \geq 0 : \quad x(t) = \int_0^t dt' \xi(t'), \tag{9.18}$$

$$\forall t \geq 0 \quad \langle x^2(t) \rangle = \int_0^t dt_1 \int_0^t dt_2 2D\delta(t_1 - t_2) = 2D \int_0^t dt_1 = 2Dt, \tag{9.19}$$

where we also set $x(0) = 0$. Infinitesimal version of Eq. (9.19) is

$$\langle (x_{n+1} - x_n)^2 \rangle = 2D\Delta, \tag{9.20}$$

which can thus be derived from the Brownian (no drift) version of Eq. (9.17).

9.3.4 The Wiener Process: A Rigorous Defintion of Brownian Motion

Note that the notations introduced above for the SDE, in its continuous time form (9.16), and in the discrete time form (9.17), are originally from physics and custom in (at least some part of) applied mathematics. The notations are intuitive and simple, however they are formally ambiguous as dependent on the notion of the δ-function, i.e. a generalized function. This is similar to the ambiguity associated with the δ-function use, for example, as a source term in a linear ODE describing the Green function. Recall that to resolve the ambiguity associated with the δ-function we should either regularize the δ function or use it only under an integral. Therefore in theoretical mathematics, statistics and engineering, we more often see the SDE (9.16) restated in the differential form

$$dx(t) = v(x(t))dt + \sqrt{2D}dW(t), \tag{9.21}$$

where $W(t)$ denotes the so-called "standard Brownian motion", also called the Wiener process/term. (In some of the mathematics literature, $dB(t)$, is used instead of $dW(t)$.) Formally:

Definition 9.3.3 (Wiener Process)**.** The *Wiener process*, $\{W(t)\}_{t\geq 0^+}$, is a continuous time stochastic process in $\mathbb{R}$ that is characterized by the following three properties:

1. $W(0) = 0$;
2. $W(t)$ is almost surely continuous (i.e. with probability 1, $W(t)$ is continuous in t);
3. $W(t)$ has stationary independent increments, that are normally distributed with mean zero and variance equal to the length of the increment, that is $W_t - W_{t'} \sim \mathcal{N}(0, t-t')$ (for $0 \leq t' \leq t$).

The differential form (9.21) of the Langevin equation is advantageous because it naturally leads to the following integral version of the Langevin equation resolving the aforementioned ambiguity

$$x(t+\Delta) - x(t) = \int_t^{t+\Delta} dx(t') = \int_t^{t+\Delta} v(x(t'))dt' + \sqrt{2D}\int_t^{t+\Delta} dW(t'), \tag{9.22}$$

where the first term on the rhs is the standard (Lebesgue) integral, while the second term is the so-called Ito integral. According to Eqs. (9.18,9.19), and consistently with the formal definition of the Wienner process above, the heuristic intepretation of the Ito integral is that when $\Delta \to 0$, the increment, $x(t+\Delta) - x(t)$, becomes Gaussian, zero mean normally distributed with the variance, $2D\Delta$.

9.3.5 From the Langevin Equation to the Path Integral

The Langevin equation can also be viewed as relating the change in $x(t)$, i.e. its dynamic increment, to stochastic dynamics of the δ-correlated source $\xi(t_n) = \xi_n$ characterized by the Probability Density Function (PDF)

$$p(\xi_1, \cdots, \xi_N) = (2\pi)^{-N/2} \exp\left(-\sum_{n=1}^{N} \frac{\xi_n^2}{2}\right). \tag{9.23}$$

Note that Eqs. (9.16,9.17,9.23) are the starting points for our further derivations, but they should also be viewed as a way to simulate the Langevin equation on computer by generating many paths at once, i.e. simultaneously. Notice, for completeness, that there are also other ways to simulate the Langevin equation through the so-called telegraph process.

Let us now formally express ξ_n via x_n from Eq. (9.17) and substitute it into Eq. (9.23)

$$p(\xi_1, \cdots, \xi_{N-1}) \to p(x_1, \cdots, x_N) = (4\pi D)^{-(N-1)/2} \exp\left(-\frac{1}{4D\Delta} \sum_{n=1}^{N-1} (x_{n+1} - x_n - \Delta v(x))^2\right). \tag{9.24}$$

One gets an explicit expression for the measure over a path written in the discretized way. And here is a typical way of how we state it in the continuous form (e.g. as a notational shortcut)

$$p\{x(t)\} \propto \exp\left(-\frac{1}{4D} \int_0^T dt\, (\dot{x} - v(x))^2\right) \tag{9.25}$$

This object is called (in physics and math) "path integral" and/or Feynmann-Kac integral.

9.3.6 From the Path Integral to the Fokker-Plank (through sequential Gaussian integrations)

The Probability Density Function (PDF) of a path is a useful general object. However we may also want to marginalize it thus extracting the marginal PDF, for being at the position x_N at the (temporal) step N, from the joint PDF (of the path) conditioned to being at the initial position, x_1, at the moment of time t_0, $p(x_1, \cdots, x_N|x_0)$, and also from the prior/initial (distribution) $p_0(x_0)$ – both assumed known:

$$p_N(x_N) = \int dx_0 \cdots dx_{N-1} p(x_0, \cdots, x_N) = \int dx_1 \cdots dx_{N-1} p(x_1, \cdots, x_N|x_0) p_0(x_0). \tag{9.26}$$

It is convenient to derive relation between $p_N(\cdot)$ and $p_0(\cdot)$ in steps, i.e. through an induction/recurrence, integrating over $dx_0, \cdots, dx_{N-1}$ sequentially. Let us proceed analyzing the

case of the Brownian motion where, $F = 0$. Then the first step of the induction becomes

$$p_1(x_1) = (4\pi D)^{-1/2} \int dx_0 \exp\left(-\frac{1}{2D\Delta}(x_1 - x_0)^2\right) p_0(x_0) \tag{9.27}$$

$$= (4\pi D)^{-1/2} \int d\epsilon \exp\left(-\frac{\epsilon^2}{4D\Delta}\right) p_0(x_1 + \epsilon) \tag{9.28}$$

$$\approx (4\pi D)^{-1/2} \int d\epsilon \exp\left(-\frac{\epsilon^2}{4D\Delta}\right) \left(p_0(x_1) + \epsilon\partial_{x_1} p_0(x_1) + \frac{\epsilon^2}{2}\partial^2_{x_1} p_0(x_1)\right) \tag{9.29}$$

$$= p_0(x_1) + \Delta D \partial^2_{x_1} p_0(x_1), \tag{9.30}$$

where transitioning from Eq. (9.28) to Eq. (9.29) one makes Taylor expansion in ϵ, also assuming that $\epsilon \sim \sqrt{\Delta}$ and keeping only the leading terms in Δ. The resulting Gaussian integrations are straightforward. We arrive at the discretized (in time) version of the diffusion equation

$$\partial_t p(x|t) = D\partial^2_x p(x|t), \tag{9.31}$$

where we write, $p(x|t)$, to emphasize that this is the probability of being at the position x at the moment of time t, i.e. the expression is conditioned to t and thus: $\forall t : \int dx p(x|t) = 1$. Of course it is not surprising that the case of the Brownian motion has resulted in the diffusion equation for the marginal PDF. Restoring the deterministic drift term, $v(x)$, (derivation is straightforward) one arrives at the Fokker-Planck equation, generalizing the zero-drift diffusion equation

$$\partial_t p(x|t) + \partial_x(v(x)p(x|t)) = D\partial^2_x p(x|t). \tag{9.32}$$

9.3.7 Analysis of the Kolmogorov-Fokker-Planck Equation: General Features and Examples

Here we only give a very brief and incomplete description on the properties of the distribution which analysis is of a fundamental importance for Statistical Mechanics. See e.g. [20].

The Fokker-Planck equation (9.32) is a linear and deterministic Partial Differential Equation (PDE). It describes continuous in phase space, x, and time, t, evolution/flow of the probability density distribution.

Derivation was for a particle moving in 1d, $\mathbb{R}$, but the same ideology and logic extends to higher dimensions, $\mathbb{R}^d$, $d = 1, 2, \cdots$. There are also extension of this consideration to compact continuous spaces. Thus one can analyze dynamics on a circle, sphere or torus.

Analogs of the Fokker-Planck can be derived and analyzed for more complicated probabilities than just the marginal probability of the state (path integral marginalized to given time). An example here is of the so-called first-passage, or "first-hitting" problem.

The temporal evolution is driven by two terms, often called "diffusion" and "advecton". The terminology originates from fluid mechanics and describes how probabilities can "flow" in the phase space. The diffusion originates from the stochastic source, while the advection is associated with a deterministic (possibly nonlinear) force.

Linearity of the Fokker-Planck does not imply that it is simpler than the original nonlinear problem. Deriving the Fokker-Planck we made a transition from nonlinear, stochastic but ODE to linear PDE. This type of transition from nonlinear representation of many trajectories to linear probabilistic representation is typical in math/statistics/physics. The linear Fokker-Planck equation can be viewed as the continuous-time, continuous-space version of the discrete-time/discrete space Master equation describing evolution of a (finite dimensional) probability vector in the case of a Markov Chain.

The Fokker-Planck Eq. (9.32) can be represented in the 'flux' form:

$$\partial_t p(x|t) + \partial_x J(t;x) = 0 \tag{9.33}$$

where $J(t;x)$ is the flux of probability through the space-state point x at the moment of time t. The fact that the second (flux) term in Eq. (9.33) has a gradient form, corresponds to the global conservation of probability. Indeed, integrating Eq. (9.33) over the whole continuous domain of achievable x, and assuming that if the domain is bounded there is no injection (or dissipation) of probability on the boundary, one finds that the integral of the second term is zero (according to the standard Gauss theorem of calculus) and thus, $\partial_t \int dx p(x|t) = 0$. In the steady state, when $\partial_t p(x|t) = 0$ for all x (and not only in the result of integration over the entire domain) the flux is constant - does not depend on x. The case of zero-flux is the special case of the so-called 'equilibrium' statistical mechanics. (See some further comments below on the latter.)

If the initial probability distribution, $p(x|0)$ is known, $(x|t)$ for any consecutive t is well defined, in the sense that the Fokker-Planck is the Cauchy (initial value) problem with unique solution.

Remarks about simulations. One can solve PDE but can also analyze stochastic ODE approaching the problem in two complementary ways - correspondent to Eulerian and Lagrangian analysis in Fluid Mechanics describing "incompressible" flows in the probability space.

Main and simplest (already mentioned) example of the Langevin dynamic is the Brownian motion, i.e. the case of zero drift, $v = 0$. Another example, principal for the so-called 'equilibrium statistical physics', is where the velocity (of the deterministic drift) is a spatial gradient of a potential, $U(x)$: $v(x) = -\partial_x U(x)$. Think, for example about x representing a over-damped particle connected to the origin by a spring. $U(x)$ is the poten-

tial/energy stored within the spring. In this case of the gradient drift the stationary (i.e. time-independent) solution of the Fokker-Planck Eq. (9.32) can be found explicitly,

$$p(x|t)\Big|_{t\infty} \to p_{st}(x) = Z^{-1} \exp\left(-\frac{U(x)}{D}\right). \tag{9.34}$$

This solution is called Gibbs distribution, or equilibrium distribution.

9.3.8 Examples and Exercises

Example 9.3.4. Consider the motion of a Brownian particle in the parabolic potential, $U(x) = \gamma x^2/2$. (The situation is typical for the particle, which is located near minimum or maximum of a potential.) The Langevin equation (9.16) in this case becomes

$$\frac{dx}{dt} + \gamma x = \sqrt{2D}\xi(t), \quad \langle \xi(t) \rangle = 0, \quad \langle \xi(t_1)\xi(t_2) \rangle = \delta(t_1 - t_2) \tag{9.35}$$

Write a formal solution of Eq. (9.35) for $x(t)$ as a functional of $\xi(t)$. Compute $\langle x^2(t) \rangle$ as a function of t and interpret the results. Write the Kolmogorov-Fokker-Planck (KFP) equation for $p(x|t)$, and solve it for the initial condition, $p(x|0) = \delta(x)$.

Solution. Multiply Eq. (9.35) by the integrating factor $e^{\gamma t}$ to get

$$\frac{d}{dt}\Big(x(t)e^{\gamma t}\Big) = \sqrt{2D}\xi(t)e^{\gamma t},$$

which has the formal solution

$$x(t)e^{\gamma t} = x(0) + \sqrt{2D}\int_0^t \xi(t')\, e^{\gamma t'}\, dt'.$$

The formal solution simplifies to

$$x(t) = x(0)e^{-\gamma t} + \sqrt{2D}\int_0^t \xi(t')e^{-\gamma(t-t')}dt'$$

We wish to find the mean and the variance of $x(t)$. The first two moments, $\langle x(t) \rangle$ and $\langle x^2(t) \rangle$, become

$$\begin{aligned}
\langle x(t) \rangle &= \left\langle x(0)e^{-\gamma t} + \sqrt{2D}\int_0^t \xi(t')\, e^{-\gamma(t-t')}\, dt' \right\rangle \\
&= \langle x(0) \rangle e^{-\gamma t} + \sqrt{2D}\int_0^t \langle \xi(t') \rangle\, e^{-\gamma(t-t')}\, dt' \\
&= x(0)e^{-\gamma t}
\end{aligned}$$

where we have used that $\langle(\xi(t)\rangle = 0$,

$$\begin{aligned}\langle x^2(t)\rangle &= \left\langle \left(x(0)e^{-\gamma t} + \sqrt{2D}\int_0^t \xi(t')\, e^{-\gamma(t-t')}\, dt' \right) \left(x(0)e^{-\gamma t} + \sqrt{2D}\int_0^t \xi(t'')\, e^{-\gamma(t-t'')}\, dt'' \right) \right\rangle \\ &= \left\langle x(0)^2 \right\rangle e^{-2\gamma t} + 2D\int_0^t \int_0^t \left\langle \xi(t')\xi(t'') \right\rangle e^{-\gamma(t-t')}\, e^{-\gamma(t-t'')}\, dt'\, dt'' \\ &= x(0)^2 e^{-2\gamma t} + 2De^{-2\gamma t}\int_0^t \int_0^t \delta(t'-t'')e^{\gamma(t'+t'')} dt'\, dt'' \\ &= x(0)^2 e^{-2\gamma t} + \frac{D}{\gamma}\left(1 - e^{-2\gamma t}\right).\end{aligned}$$

The interpretation of the solution is as follows: (i) The contribution to $\langle(x^2(t)\rangle$ from the initial condition decays at a rate of $2\gamma t$. (ii) At the smallest times, $t \ll 1/\gamma$, we expand the term $(1-e^{-2\gamma t})$ as a first order Taylor polynomial to find the usual diffusion $\langle x^2(t)\rangle \simeq 2Dt$, since the particle does not feel the potential. (iii) At larger time scale $t \gg 1/\gamma$ the dispersion saturates, $\langle x^2(t)\rangle \simeq D/\gamma$).

The Kolmogorov-Fokker-Planck equation, $\partial_t p(x|t) = (\gamma\partial_x x + D\partial_x^2)p(x|t)$, should be supplemented by the initial condition $p(x|t) = \delta(x)$. Then, the solution (the Green function) is

$$p(x|t) = \frac{1}{\sqrt{2\pi\langle x^2(t)\rangle}} \exp\left(-\frac{x^2}{2\langle x^2(t)\rangle}\right). \tag{9.36}$$

The meaning of the expression is clear: the probability function $p(x|t)$ is Normal/Gaussian with the variance which is time-dependent. □

Example 9.3.5. Prove that the moments $\langle x^{2k}(t)\rangle$ for the Brownian motion in $\mathbb{R}$ obey the following recurrent equation

$$\partial_t\langle x^{2k}\rangle = 2k(2k-1)D\langle x^{2(k-1)}\rangle. \tag{9.37}$$

Solve this equation for a particle starting from $x = 0$ at $t = 0$.

Solution. Recall the definition of the k^{th} moment or a random variable and recall Eq.(9.31).

$$\begin{aligned}\partial_t\langle x^{2k}\rangle &= \partial_t \int_{-\infty}^{+\infty} x^{2k} p(x|t)\, dt = \int_{-\infty}^{+\infty} x^{2k}\partial_t p(x|t)\, dt = \int_{-\infty}^{+\infty} x^{2k} D\partial_{xx} p(x|t)\, dt \\ &= D\int_{-\infty}^{+\infty} \partial_{xx} x^{2k} p(x|t)\, dt = D\int_{-\infty}^{+\infty} 2k(2k-1)x^{2k-2}p(x|t)\, dt = 2k(2k-1)D\langle x^{2(k-1)}\rangle.\end{aligned}$$

Example 9.3.6 (Brownian motion in parabolic potential)**.** The conditional probability distribution for a Brownian particle in a parabolic potential, $U(x) = \alpha x^2/$ is described by the advection-diffusion equation

$$D\partial_x^2 p + \alpha\partial_x(xp) = \partial_t p. \tag{9.38}$$

Write down stochastic ODE for the underlying stochastic process, $x(t)$, and, given the initial condition for, $p(x|t) = \delta(x)$, compute respective statistical moments $\langle x^k(t)\rangle$.

Solution. We propose that the corresponding Langevin equation is

$$\dot{x} = -\alpha x + \sqrt{2D}\xi(t),$$

and verify our proposal by going through derivations similar to these shown in Example 9.3.4.

Let $\mu_k(t)$ be the k^{th} moment of the random process, so $\mu_k(t) := \langle (x(t)\rangle = \int x^k n(x,t)\,dx$. A differential equation for $\mu_k(t)$ can be derived from the KFP equation:

$$\begin{aligned}
\partial_t\big(\mu_k(t)\big) &= \partial_t \int_{-\infty}^{\infty} x^k\, p(x|t)\,dx \\
&= \int_{-\infty}^{\infty} x^k\,\alpha\,\partial_x\big(xp(x|t)\big) + x^k\,\partial_{xx}\big(p(x|t)\big)\,dx \\
&= -k\int_{-\infty}^{\infty} \alpha x^k\, p(x|t)\,dx + k(k-1)\int_{-\infty}^{\infty} x^{k-2}\,p(x|t)\,dx \\
&= -\alpha k\mu_k(t) + k(k-1)\mu_{k-2}(t). \qquad (9.39)
\end{aligned}$$

where the boundary terms from integration by parts vanish at $\pm\infty$ because n and its derivatives are smaller than any polynomial as $x \to \pm\infty$.

In the case where $p(x|0) = \delta(x)$, then $\mu_0(t) = 1$ and $\mu_1(t) = 0$. Applying the recursive relationship, we get $\mu_{2k+1}(t) = 0$ for all odd-order moments. For even order moments, the solutions are subjected to solving Eq. (9.39). For example, the second moment can be found by solving the differential equation

$$\partial_t\mu_2(t) = -2\alpha\mu_2(t) + 2,$$

hence $\mu_2(t) = \frac{1}{\alpha}(1 - e^{-2\alpha t})$.

Example 9.3.7. Consider the following expectation over the Langevin term, $\xi(t)$,

$$\Psi(t;x) = \left\langle \exp\left(\int_0^t d\tau Q(x(\tau))\right)\right\rangle_{x(t)=x,\ x(0)=0} \qquad (9.40)$$

$$\leftarrow \Big|_{N\to\infty} \Psi_N(x) = \int dx_0\cdots dx_{N-1} p\left(x, x_{N-1}, \cdots, x_1|x_0\right)\delta\left(x_0\right) e^{\Delta(Q(x_{N-1})+\cdots+Q(x_0))} \qquad (9.41)$$

$$= \int \frac{dx_1}{\sqrt{2\pi D\Delta}}\cdots\frac{dx_{N-1}}{\sqrt{4\pi D\Delta}}\exp\left(-\frac{(x-x_{N-1})^2+\cdots+x_1^2}{4D\Delta}\right)e^{\Delta(Q(x_{N-1})+\cdots+Q(0))},$$

where $x(t)$ is the Brownian motion, thus satisfying, $\dot{x}(t) = \xi(t)$, with $x(t)$ set to zero initially, $x(0) = 0$; $Q(x(t))$ is a given function of x, which is finite everywhere in $\mathbb{R}$; and Eq. (9.40) and Eq. (9.41) show, respectively, continuous time and discrete time versions of the same expectation of interest.

- (a) Derive partial differential equation governing "evolution" of $\Psi(t;x)$ in t and x.
- (b) Suggest a scheme allowing to compute the first and second moments of $\int\limits_0^t d\tau Q(x(\tau))$,

$$\phi^{(1)}(t) := \left\langle \int\limits_0^t d\tau Q(x(\tau)) \right\rangle_{x(0)=0}, \quad \phi^{(2)}(t) := \left\langle \left(\int\limits_0^t d\tau Q(x(\tau)) \right)^2 \right\rangle_{x(0)=0}, \quad (9.42)$$

explicitly, i.e. bypassing solving the PDE (derived in (a)), which does not allow explicit solutions for a general $Q(x(t))$.

Solution. (a) First of all notice that $\Psi(0;x) = \delta(x)$. Further, observe that when $Q(x) = 0$, the resulting PDE becomes the diffusion Eq. (9.32) with $p(t;x)$ substituted by $\Psi(t;x)$, respectively. We can rewrite Eq. (9.41) as a recurrence

$$\forall k = 1, \cdots, N: \quad \Psi_k(x) = \int \frac{dx_{k-1}}{\sqrt{4\pi D\Delta}} \exp\left(-\frac{(x - x_{k-1})^2}{4D\Delta} + \Delta Q(x_{k-1}) \right) \Psi_{k-1}(x_{k-1}), \quad (9.43)$$

where $\Psi_0(x) = \delta(x)$. Changing the integration variable, $x_{k-1} \to \epsilon = x_{k-1} - x$, keeping the Gaussian expression in the integrand intact and expanding all other terms in the Taylor series in ϵ, then evaluating the resulting Gaussian integrals, we arrive at the following differential version of the recurrence:

$$\forall k = 1, \cdots, N: \quad \Psi_k(x) = \left(1 + \Delta Q(x) + \Delta D\partial_x^2 + O(\Delta^2)\right) \Psi_{k-1}(x).$$

The continuous time version of the Eq. (9.43), and therefore the desirable Cauchy (initial value) problem stated as a PDE supplemented with the initial condition, becomes:

$$\partial_t \Psi(t;x) = Q(x)\Psi(t;x) + D\partial_x^2 \Psi(t;x), \ \Psi(0;x) = \delta(x). \quad (9.44)$$

(b) Let us substitute $Q(x)$ in the expressions above, e.g. in the PDE (9.44), by $\delta * Q(x)$, expand the PDE (9.44) in the Taylor series in δ and write down relations for the zero, first and second terms in δ. We derive the following set of diffusion equations (first homogeneous and two other inhomogeneous)

$$\partial_t \psi^{(0)}(t;x) = D\partial_x^2 \psi^{(0)}(t;x), \ \psi^{(0)}(0;x) = \delta(x), \quad (9.45)$$

$$\partial_t \psi^{(1)}(t;x) = Q(x)\psi^{(0)}(t;x) + D\partial_x^2 \psi^{(1)}(t;x), \ \psi^{(1)}(0;x) = 0, \quad (9.46)$$

$$\partial_t \psi^{(2)}(t;x) = Q(x)\psi^{(1)}(t;x) + D\partial_x^2 \psi^{(2)}(t;x), \ \psi^{(2)}(0;x) = 0, \quad (9.47)$$

where

$$n = 0, 1, \cdots : \ \psi^{(n)}(t; x) := \left\langle \left(\int_0^t d\tau Q(x(\tau) \right)^n \right\rangle_{x(t)=x, \ x(0)=0} .$$

Eqs. (9.45,9.46,9.47) can be resolved explicitly

$$\psi^{(0)}(t; x) = \frac{1}{\sqrt{4\pi D t}} \exp\left(-\frac{x^2}{4Dt}\right),$$
$$\psi^{(1)}(t; x) = \int_0^t \int_{-\infty}^{\infty} d\xi \, d\tau \frac{1}{\sqrt{4\pi D(t-\tau)}} \exp\left(-\frac{(x-\xi)^2}{4D(t-\tau)}\right) \psi^{(0)}(\tau; \xi)$$
$$\psi^{(2)}(t; x) = \int_0^t \int_{-\infty}^{\infty} d\xi \, d\tau \frac{1}{\sqrt{4\pi D(t-\tau)}} \exp\left(-\frac{(x-\xi)^2}{4D(t-\tau)}\right) \psi^{(1)}(\tau; \xi)$$

Finally, observe that

$$k = 0, \cdots : \quad \phi^{(k)}(t) = \int dx \psi^{(k)}(t; x).$$

□

Exercise 9.3 (Self-propelled particle)**.** The term "self-propelled particle" refers to an object capable to move actively by gaining energy from the environment. Examples of such objects range from the Brownian motors and motile cells to macroscopic animals and mobile robots. In the simplest two-dimensional model, the self-propelled particle moves in the xy-plane with fixed speed v_0. The Cartesian components of the particle velocity, $\dot{x}(t)$ and $\dot{y}(t)$, in the polar coordinates are

$$\dot{x} = v_0 \cos\varphi, \qquad \dot{y} = v_0 \sin\varphi,$$

where the polar angle φ defines the direction of motion. Assume that φ evolves according to the stochastic equation

$$\frac{d\varphi}{dt} = \sqrt{2D}\xi, \tag{9.48}$$

where $\xi(t)$ is the Gaussian white noise with zero mean and the following pair correlation function, $\langle \xi(t_1)\xi(t_2)\rangle = \delta(t_1 - t_2)$. The initial condition are chosen to be $\varphi(0) = 0$, $x(0) = 0$ and $y(0) = 0$.

(a) Calculate $\langle x(t)\rangle$, $\langle y(t)\rangle$.

(b) Calculate $\langle r^2(t)\rangle = \langle x^2(t)\rangle + \langle y^2(t)\rangle$.

Hint: Derive equation for probability density of observing φ at the moment of time t, solve the equation and use the result. You may also consider using first and second derivatives of the object of interest over t, as well as evaluations fro the Example 9.3.7.

9.4 Markov Process [discrete space, discrete time]

It may be tempting when studying stochastic processes to assume that the random events are independent and identically distributed (i.i.d). However, complex systems in the real world often "jump" from one random state to another in such a way that previous states influence future states. In general, the memory may last for more than one jump, but there is also a large family of interesting random processes which do not have long memory and the jumps are only directly influenced by the current state, and not by any previous states. More precisely, the *Markovian simplification* is $P(X_{t+1}|X_t, X_{t-1}, X_{t-2}, \dots) = P(X_{t+1}|X_t)$. These random processes are called *Markov Processes* (MPs) or, equivalently, *Markov Chains* (MCs).

9.4.1 Transition Probabilities

The Markovian simplification allows us to think of a Markov chain as a random walk on a directed graph. The vertices of the graph correspond to the various states and the edges correspond to transitions between the states and are associated with the probability of transitioning between the corresponding states[a].

The graphical representation of a MC is as follows: introduce a directed graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the set of vertices, $\mathcal{V} = \{i\}$, is associated with the set of states, and the set of directed edges, $\mathcal{E} = \{j \leftarrow i\}$, which correspond to possible transitions between the states. Note that we may also have "self-loops", $\{i \leftarrow i\}$, included in the set of edges. To complete the description we need to associate to each vertex the probability $P(X_t = j|X_{t-1} = i)$ of transitioning from the state i to the state j, which we write as $p_{j\leftarrow i}$ or p_{ji}. Since p_{ji} is a probability, it must statisfy

$$\forall (j \leftarrow i) \in \mathcal{E} : \quad p_{ji} \geq 0 \tag{9.49}$$

and

$$\forall i : \quad \sum_{j:(j\leftarrow i)\in\mathcal{E}} p_{ji} = 1. \tag{9.50}$$

The combination of $\mathcal{G}$ and $p := (p_{ji}|(j \leftarrow i) \in \mathcal{E})$ defines a MC. Mathematically we also say that the tuple (finite ordered set of elements), $(\mathcal{V}, \mathcal{E}, p)$, defines the Markov chain.

We will mainly consider stationary Markov chains, where p_{ji} does not change in time. However, for many of the following statements/considerations generalization to the time-dependent processes is straightforward.

[a] A useful interactive playground can be found here http://setosa.io/ev/markov-chains/

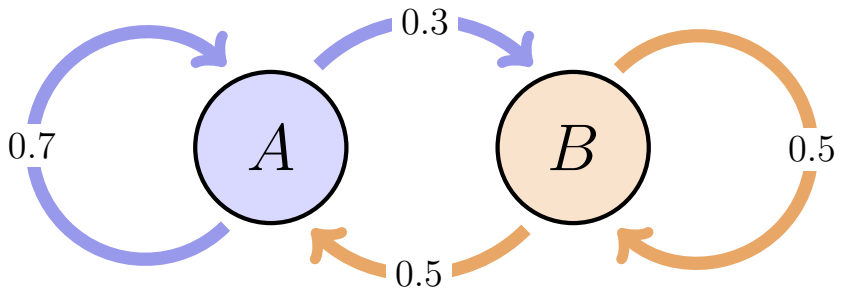


Figure 9.3: An example of a two-state Markov chain.

9.4.2 Sample Trajectories and Analysis by Simulation

One way to analyze a Markov chain is by generating sample trajectories. How does one relate the weighted, directed graph to samples? The relation, actually, has two sides. The direct side is about generating samples, which is done by first initializing the trajectory at a particular state, and then advancing the trajectory from the current state to a randomly selected adjacent state according to the transition probability p_{ij}. The inverse side is about verifying whether given samples where indeed generated according to the rather restrictive MC rules, and even reconstructing the characteristics of the underlying Markov chain.

Example 9.4.1. Describe how a sample trajectory may be generated for the Markov chain illustrated in Fig. 9.3. Assume the system begins in state A at time 0.

Solution. Since the system is in state A at time 0 set $X_0 = A$. At time 1, the system may either remain in state A with probability 0.7 or transition to state B with probability 0.3. Formally we write, $P(X_1 = A|X_0 = A) = 0.7$ and $P(X_1 = B|X_0 = B) = 0.3$. One can generate a sample trajectory by first drawing a random number on the interval $[0, 1)$ and then setting $X_1 = A$ if the random number lies on $[0, 0.7)$ and setting $X_1 = B$ if it lies on $[0.7, 1)$. By the Markov property, the state of the system at time 2 depends only on the state at time 1, so one can then generate a second random number on $[0, 1)$ and define X_2 accordingly. Samples of the trajectory $(x_t)_{t=0}^{\infty}$ can be generated efficiently in this manner and may look something like $AABABBAA$....

The Markov chain defines a random (stochastic) dynamic process. Although time may flow continuously, Markov chains consider time to be discrete (which is sometimes a matter of convenient abstraction and sometimes, actually quite often, events do happen discretely). One uses $t = 0, 1, 2, \cdots$ for the times when jumps occur. Then a particular random trajectory/path/sample of the system will look like

$$i_1(0), i_2(1), \cdots, i_k(t_k), \quad \text{where} \quad i_1, \cdots, i_k \in \mathcal{V}$$

We can also generate many samples (many trajectories)

$$n = 1, \cdots, N: \quad i_1^{(n)}(0), i_2^{(n)}(1), \cdots, i_k^{(n)}(t_k), \quad \text{where} \quad i_1, \cdots, i_k \in \mathcal{V}$$

where N is the number of trajectories.

It can be interesting to ask about various statistics of a trajectory, for example, (i) the proportion of time spent in a particular state, or (ii) the probability that the system takes longer than k steps to return to a particular state once leaving it. Although it may be tempting to assume that the statistics measured from a particular trajectory will be representative of those that would be measured from any other trajectory, this is *not* true for all Markov chains. In the following section, we will demonstrate the necessary property, which will be called ergodicity, that guarantees whether individual trajectories are actually representative of the Markov chain.

Basic Properties of Markov Chains

Definition 9.4.2 (Irreducible)**.** A Markov chain is said to be *irreducible* if one can access any state from any state, formally

$$\forall i, j \in \mathcal{V}: \quad \exists n > 1, \quad \text{s.t.} \quad P(X_n = j | X_0 = i) > 0. \tag{9.51}$$

The Markov chain in Fig. (9.3) is obviously irreducible. However, if we replace $0.3 \to 0$ and $0.7 \to 1$ it becomes reducible because state 1 would no longer be accessible from 2.

Definition 9.4.3 (Aperiodicity)**.** We say that *state i has period k* if every return to the state must occur at times that are multiples of k. Formally the period of state is k where

$$k = \text{greatest common divisor} \left\{ n > 0 : P(X_n = i | X_0 = i) > 0 \right\},$$

provided that the set is not empty (otherwise the period is not defined). If $n = 1$ than the state is said to be *aperiodic*. If all the states of a Markov Chain are aperiodic, then we say that the Markov chain is *aperiodic*.

An irreducible MC only needs one aperiodic state to imply all states are aperiodic. Any MC with at least one self-loop is aperiodic. The Markov chain in Fig. (9.3) is obviously aperiodic. However, it becomes periodic with period two if the two self-loops are removed.

Example 9.4.4. Consider the Markov chain shown in Fig. 9.4. Is this Markov chain reducible or irreducble? Periodic or aperiodic?

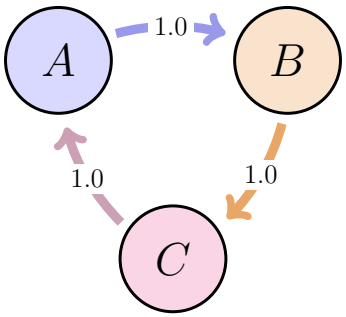


Figure 9.4: An example of a three-state periodic Markov chain.

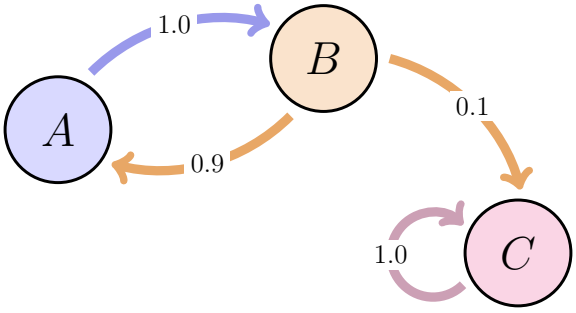


Figure 9.5: An example of a three-state reducible Markov chain.

Solution. The Markov chain shown in Fig. (9.4) is irreducible and periodic. The Markov chain is irreducible because each state is accessible from each other state. The Markov chain is periodic because if we start in state C, we can return to it only after $3, 6, 9, \ldots$ steps (the system will never forget its initial state). We say that state "C" has period 3. We say that a Markov chain is aperiodic if and only if each state has period 1. One can make the this Markov chain aperiodic by adding a self-loop to any of the three state. □

Example 9.4.5. Consider the Markov chain shown in Fig. 9.5. Is this Markov chain reducible or irreducble? Periodic or aperiodic?

Solution. The Markov chain shown in Fig. (9.5) is reducible and periodic. The Markov chain is reducible because states "A" and "B" cannot be accessed from state "C". Notice that once the system enters state "C", it will remain there forever. It cannot escape from state "C". This Markov chain is periodic because state "A" has period 2. □

It is often of interest whether the system is guaranteed to return to a given state upon leaving it, and if so, how many steps we should expect this to take. Define the time of the

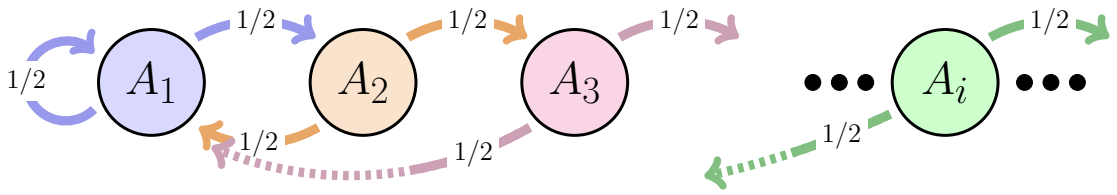


Figure 9.6: An example of a Markov chain with countably many (thus infinite number of) states. In this example, $P(A_{n+1} \leftarrow A_n) = 1/2$ and $P(A_1 \leftarrow A_n) = 1/2$. This Markov chain is positive recurrent. See Example 9.4.8.

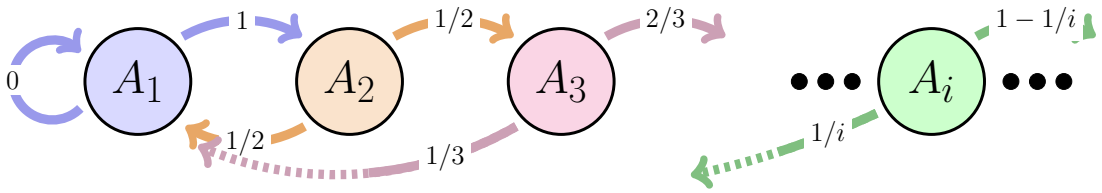


Figure 9.7: An example of a Markov chain with countably many states. In this example, $P(A_{n+1} \leftarrow A_n) = 1 - 1/n$ and $P(A_1 \leftarrow A_n) = 1/n$. This Markov chain is recurrent, but not positive recurrent. See Example 9.4.9.

first return to a state by

$$\tau_1 = \inf_{n \geq 1} \{n : X_n = X_0\}. \tag{9.52}$$

The expected time of return to a particular state is

$$\mathbb{E}[\tau_1 | X_0] = \sum_{n=1}^{\infty} nP(\tau_1 = n). \tag{9.53}$$

Example 9.4.6. Consider the Markov chain illustrated in Fig. 9.3. Compute the probability that the first return to state A is in exactly n steps. (b) Compute the expected return time to state A.

Solution.

(a) Observe that if the first return to state A is in exactly 1 step, then the state must have transitioned from A back to A (a self-loop). If the first return to state A is in exactly n steps (where $n \geq 2$), then the state must have transitioned from A to B on step 1, and then remained at B $n-2$ times, and then returned from B to A on step n.

$$\begin{aligned} P(\tau_1 = n | X_0 = A) &= P(X_1 = B | X_0 = A) \cdot \Big(P(X_k = B | X_{k-1} = B) \Big)^{n-2} \\ &\quad \cdot P(X_n = A | X_{n-1} = B) \\ &= \begin{cases} 0.7, & \text{if } n = 1, \\ 0.3 \cdot (0.5)^{n-2} \cdot 0.5, & \text{otherwise.} \end{cases} \end{aligned}$$

(b) The expected return time to state A is

$$\begin{aligned} \mathbb{E}[\tau_1 | X_0 = A] &= 1 \cdot P(\tau_1 = 1) + 2 \cdot P(\tau_1 = 2) + 3 \cdot P(\tau_1 = 3) + \cdots \\ &= 1\,(0.7) + 2\,(0.3)(0.5) + 3\,(0.3)(0.5)(0.5) + \cdots = 1.6 \end{aligned}$$

□

Example 9.4.7. Consider the Markov chain illustrated in Fig. 9.4. Compute the probability that the first return to the state A occurs in exactly n steps. (b) Compute the expected return time to the state A.

Solution.

(a) Observe that when the state transitions out of A it is guaranteed to transition to B, and from there to C, and from there back to A.

$$P(\tau_1 = n | X_0 = A) = \begin{cases} 1, & \text{if } n = 3, \\ 0, & \text{otherwise.} \end{cases}$$

(b) The expected return time to state A is

$$\begin{aligned} \mathbb{E}[\tau_1 | X_0 = A] &= 1 \cdot P(\tau_1 = 1) + 2 \cdot P(\tau_1 = 2) + 3 \cdot P(\tau_1 = 3) + \cdots \\ &= 1\,(0) + 2\,(0) + 3\,(1) + 4\,(0) \cdots = 3. \end{aligned}$$

□

Example 9.4.8. Consider the Markov chain illustrated in Fig. 9.6. Compute the probability that the first return to state A_1 is in exactly n steps. (b) Compute the expected return time to state A_1.

Solution.

(a) Notice that if the first return time is n, then the state must have transitioned from A_1 to A_2 to A_3 and so on up to A_{n-1}, and then returned to A_1 in the n^{th} step.

$$P(\tau_1 = n | X_0 = A_1), = \frac{1}{2} \cdot \frac{1}{2} \cdot \cdots \frac{1}{2} = \frac{1}{2^n}.$$

(b) The expected return time to state A is

$$\begin{aligned}\mathbb{E}[\tau_1 | X_0 = A_1] &= 1 \cdot P(\tau_1 = 1) + 2 \cdot P(\tau_1 = 2) + 3 \cdot P(\tau_1 = 3) + \cdots \\ &= 1\,(1/2) + 2\,(1/4) + 3\,(1/8) + 4\,(1/16) \cdots = 2.\end{aligned}$$

□

Example 9.4.9. Consider the Markov chain illustrated in Fig. 9.7. Compute the probability that the first return to state A_1 is in exactly n steps. (b) Compute the expected return time to the state A_1.

Solution.

(a) Notice that if the first return time is n, then the state must have transitioned from A_1 to A_2 to A_3 and so on up to A_{n-1}, and then returned to A_1 in the n^{th} step.

$$P(\tau_1 = n | X_0 = A_1) = \frac{1}{1} \cdot \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \cdot \cdots \cdot \frac{n-2}{n-1} \cdot \frac{1}{n} = \frac{1}{n-1}\frac{1}{n}.$$

(b) The expected return time to state A is

$$\begin{aligned}\mathbb{E}[\tau_1 | X_0 = A_1] &= 1 \cdot P(\tau_1 = 1) + 2 \cdot P(\tau_1 = 2) + 3 \cdot P(\tau_1 = 3) + \cdots \\ &= 1\,(0) + 2\,(1)(1/2) + 3\,(1/2)(1/3) + 4\,(1/3)(1/4) \cdots \to \infty.\end{aligned}$$

□

Definition 9.4.10 (Transient, Recurrent, Positive Recurrent)**.** A state i is said to be *transient* if, given that we start in state i, there is a non-zero probability that we never return to i. A state i is said to be *recurrent* if the probability that we never return to i is zero (even if the expected return time is infinite). A state i is said to be *positive-recurrent* if the expected return time is finite.

The distinction between recurrent and positive recurrent becomes important when analyzing Markov chains on countable state spaces.

Exercise 9.4. Give an example of a Markov chain with an infinite number of states, which is irreducible and aperiodic (prove it), but which does not converge to an equilibrium probability distribution.

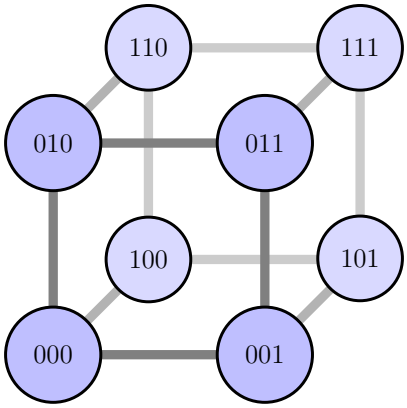


Figure 9.8: Sampling on a Hypercube

Definition 9.4.11 (Ergodic)**.** A state is said to be *ergodic* if the state is aperiodic and positive-recurrent. A Markov chain is said to be *ergodic* if it is irreducible and if every state is ergodic.

Ergodicity can be restated as follows: A MC is ergodic if it is aperiodic and if there is a finite number k_* such that any state can be reached from any other state in exactly k_* steps. (For the example of Eq. (9.58) $k_* = 2$.)

There are other (alternative) descriptions of ergodicity. A particularly intuitive one is: **the MC is ergodic if it is aperiodic and irreducible**. (Notice that ergodicity still holds if we replace positive-recurrence by irreducibility, but the combination of irreducibility and positive-recurrence without aperiodicity does not guarantee ergodicity.) In this course we will not go into related mathematical formalities and details, largely considering generic, i.e. ergodic, MC.

Ergodicity and Sampling

Markov chains are often used to generate samples of a desired distribution. One can imagine a particle that travels over a graph according to the weights of the edges. If the Markov chain is ergodic, then after some time the probability distribution of the particle becomes stationary (one say that the chain is mixed) and then the trajectory of the particle will represent the sample of a distribution. Important information about the distribution, such as its moments or the expectation values of functions, is provided by analyzing the trajectory of a particle.

Imagine that you need to generate a random string of n bits. There are 2^n possible configurations. You can organize these configurations on a hypercube graph with 2^n vertices

where each vertex has n neighbors, corresponding to the strings that differ from it by a single bit as in Fig. 9.8. Our Markov chain will walk along these edges and flip one bit at a time. The trajectory after a long time will correspond to the series of random strings. The important question is how long should we wait before our Markov chain becomes mixed (loses a memory about initial condition)? To answer this question we should look at the MC from a more mathematical point of view.

9.4.3 Evolution of the Probability State Vector

We began the section by defining a Markov chain in terms of the weighted, directed graph $(\mathcal{V}, \mathcal{E}, p)$ and attempting to analyze by generating sample trajectories. A rigorous analysis involves the *probability state vector*, or simply the *state vector*, which is the vector where the i^{th} component represents the probability that the system is in the state i at the moment of time t:

$$\pi(t) := (\pi_i(t))_{i\in\mathcal{V}} \quad \text{where} \quad \pi_i(t) := P(X_t = i). \tag{9.54}$$

Thus, $\pi_i \geq 0$ and $\sum_{i\in\mathcal{V}} \pi_i = 1$.

The probability state vector evolves according to

$$\forall i \in \mathcal{V}, \quad \forall t = 0, \cdots : \quad \pi_i(t+1) = \sum_{j:(i\leftarrow j)\in\mathcal{E}} p_{ij}\pi_j(t). \tag{9.55}$$

We can also rewrite Eq. (9.55) in the vector/matrix form

$$\pi(t+1) = p\pi(t), \tag{9.56}$$

where $p := \{p_{ji}\}$, called the *transition probability matrix*, is the matrix whose (i, j) component is the probability of transitioning from state j to state i and is therefore matrix. It is a stochastic matrix (defined below).

Definition 9.4.12. A matrix is called *stochastic* if all of its components are nonnegative and each column sums to 1.

To analyze the Markov chain acting after k sequential step, we consider repeated application of Eq. (9.56) which results in

$$\pi(t+k) = p^k\pi(t), \tag{9.57}$$

We therefore are interested in analyzing the properties of the matrix p^k,.

Example 9.4.13. Find the stochastic matrix associated with Fig. 9.3? Is the corresponding Markov chain reducible?

Solution. For Fig. (9.3), the stochastic matrix is:

$$p = \begin{pmatrix} 0.7 & 0.5 \\ 0.3 & 0.5 \end{pmatrix}$$

To determine whether the Markov chain is irreducible, we estimate p^k for large k:

$$p^2 = \begin{pmatrix} 0.64 & 0.6 \\ 0.36 & 0.4 \end{pmatrix}, \quad p^{10} \approx p^{100} \approx \begin{pmatrix} 0.625 & 0.625 \\ 0.375 & 0.375 \end{pmatrix}. \tag{9.58}$$

A Markov chain is irredcuible if each state is accessible from every other state. If $\pi(k)$ is the vector of probabilities for each state at time k, given initial probabilities $\pi(0)$, then we observe that for the initial conditions corresponding to each of the two states, $\pi(0) = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$, and $\pi(0) = \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$, we find that every entry of $\pi(k) = p^k\pi(0)$ is non-zero. Therefore every state has non-zero probability at time k and we conclude that the Markov chain is irreducible. □

Example 9.4.14. Find the stochastic matrix associated with Fig. 9.5? Is the corresponding Markov chain reducible?

Solution. For Fig. (9.5), the stochastic matrix is:

$$p = \begin{pmatrix} 0.8 & 0.9 & 0.0 \\ 0.2 & 0.0 & 0.0 \\ 0.0 & 0.1 & 1.0 \end{pmatrix}.$$

Subsequent powers of p are:

$$p^2 = \begin{pmatrix} 0.82 & 0.72 & 0.00 \\ 0.16 & 0.18 & 0.00 \\ 0.02 & 0.10 & 1.00 \end{pmatrix}, \quad \text{and} \quad p^{10} \approx \begin{pmatrix} 0.71 & 0.65 & 0.00 \\ 0.14 & 0.133 & 0.00 \\ 0.14 & 0.21 & 1.00 \end{pmatrix}.$$

Repeated matrix multiplication shows that the the first two entries of the third column are zero for all k. This means that states "A" and "B" are inaccessible from state "C". Therefore the Markov chain is reducible. □

Steady State Analysis

Definition 9.4.15 (Stationary Distribution)**.** The probability state vector (if it exists) that satisfies

$$\pi^* = p\pi^* \tag{9.59}$$

is called the *stationary distribution* or *invariant measure.* (Recall, to be a state vector, each component must be positive, and the components must sum to unity).

Theorem 9.4.16 (Existence of a Stationary Distribution)**.** A Markov chain has a stationary distribution iff it is ergodic. (Equivalently, a MC has a stationary distribution iff it is aperiodic and all of its states are positive recurrent.

Solving Eq. (9.59) for the example of Eq. (9.58) one finds

$$\pi^* = \begin{pmatrix} 0.625 \\ 0.375 \end{pmatrix}, \tag{9.60}$$

which is naturally consistent with Eq. (9.58).

In general, the stochastic matrix for an ergodic Markov chain has one eigenvalue satisfying $\lambda^* = 1$. The stationary distribution π^* is the ℓ^1 normalized eigenvector associated with the unit eigenvalue

$$\pi^* = \frac{e}{\sum_i e_i}, \tag{9.61}$$

And how about other eigenvalues of the transition matrix?

An important practical consequence of the ergodicity is that the steady state is unique and it is universal. Universality means that the steady state does not depend on the initial condition. It may now be timely to ask: why do we care about uniqueness of the steady state, invariance with respect to the initial condition and ergodicity? The most straightforward answer is because it allows us to design powerful techniques to explore complicated phase space. Markov Chain Monte Carlo (MCMC), which will be discussed in chapter 10, is one such technique. At the moment, it is important to appreciate that understanding different properties of MC (and latter MCMC algorithms) allows us to use Markov chains to solve complicated inference and (machine) learning problems in Data Science and related disiplines efficiently. When we turn from analysis of a particular MC to the design of a desirable MC we will be stating the "desires" in terms of uniqueness, invariance and ergodicity. Ergodicity will ensure convergence to a unique desirable probability distribution. Moreover, MCMC allows us to generate samples which are drawn independently from ANY probability distribution. Generating independent samples is generally difficult, but Markov chains helps us to solve the problem bypassing independence and creating much easier to generate dependent samples. We will eventually make the samples independent if we repeat MC many times and show that the sample/state we have started with is forgotten after sufficiently many steps.

Spectrum of the Transition Matrix & Speed of Convergence to the Stationary Distribution

Assume that p is diagonalizable (has $n = |p|$ linearly independent eigenvectors) then we can decompose p according to the following eigen-decomposition

$$p = U\Sigma U^{-1}, \tag{9.62}$$

where $\Sigma = \text{diag}(\lambda_1, \cdots, \lambda_n)$, $1 = \lambda_1 \geq |\lambda_2| \geq |\cdots|\lambda_n|$ and U is the matrix of eigenvectors (each normalized to having an l_2 norm equal to 1) where each row is a right eigenvector of p. Then, the evolution of an initial stochastic vector, $\pi(0)$, in the discrete time $t = 1, \cdots$, is given by

$$\pi(t) = p^t\pi(0) = (U\Sigma U^{-1})^t\pi(0) = U\Sigma^t U^{-1}\pi(0). \tag{9.63}$$

Let us represent the initial $\pi(0)$ as an expansion over the normalized eigenvectors, $u_i, \cdots i = 1, \cdots, n$, of p:

$$\pi(0) = \sum_{i=1}^{n} a_i u_i. \tag{9.64}$$

Taking into account orthonormality of the eigenvectors one derives

$$\pi(t) = \lambda_1 \left(a_1 u_1 + a_2 \left(\frac{\lambda_2}{\lambda_1}\right)^t u_2 + \cdots + a_n \left(\frac{\lambda_n}{\lambda_1}\right)^t u_n \right). \tag{9.65}$$

Since $\lim_{t\to\infty} \pi(t) = \pi^* = u_1$, we get that $a_1 = 1$ and the second term on the rhs of Eq. (9.65) describes the rate of convergence of $\pi(t)$ to the steady state at $t \to \infty$. The convergence is exponential in t with the rate, $\log(\lambda_1/\lambda_2)$.

Example 9.4.17. Find eigenvalues for the MC shown in Fig. (9.9) with the transition matrix

$$p = \begin{pmatrix} 0 & 5/6 & 1/3 \\ 5/6 & 0 & 1/3 \\ 1/6 & 1/6 & 1/3 \end{pmatrix}. \tag{9.66}$$

What does define the speed of the MC convergence to a steady state?

Solution. Let us start by noticing that p is stochastic. If the initial probability distribution is $\pi(0)$, then the distribution after t steps is

$$\pi(t) = p^t\pi(0). \tag{9.67}$$

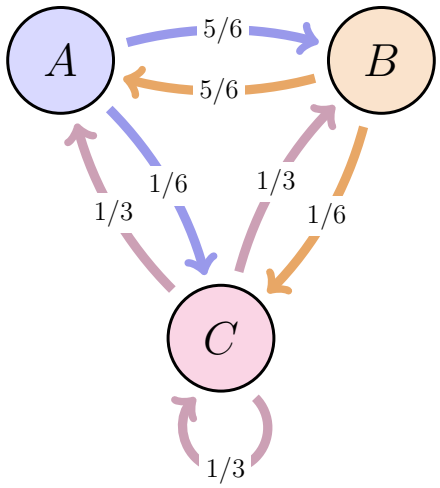


Figure 9.9: Illustration of the Detailed Balance (DB)

As t increases, $\pi(t)$ approaches a stationary distribution π^* (since the Markov chain is ergodic - this property is easy to check for the MC), such that

$$p\pi^* = \pi^*. \tag{9.68}$$

Thus, π^* is an eigenvector of p with eigenvalue 1 with all components positive and normalized. The matrix (9.66) has three eigenvalues $\lambda_1 = 1, \lambda_2 = 1/6, \lambda_3 = -5/6$ and corresponding eigenvectors are

$$\pi^* = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right)^T, \; u_2 = \left(-\frac{1}{2}, -\frac{1}{2}, 1\right)^T, \; u_3 = (-1, 1, 0)^T. \tag{9.69}$$

Suppose that we start in the state "A", i.e. $\pi(0) = (1, 0, 0)^T$. We can write the initial state as a linear combination of the eigenvectors

$$\pi(0) = \pi^* - \frac{u_2}{5} - \frac{u_3}{2}, \tag{9.70}$$

and then

$$\pi(t) = p^t \pi(0) = \pi^* - \frac{\lambda_2^t}{5} u_2 - \frac{\lambda_3^t}{2} u_3. \tag{9.71}$$

Since $|\lambda_2| < 1$ and $|\lambda_3| < 1$, then in the limit $t \to \infty$ we obtain $\pi(t) = \pi^*$. The speed of convergence is defined by the eigenvalue (λ_2 or λ_3), which has the greatest absolute value. □

Note, that the considered situation generalizes to the following powerful statement (see [16] for details):

Theorem 9.4.18 (Perron-Frobenius Theorem)**.** Ergodic Markov chain with transition matrix p has a unique eigenvector π^* with eigenvalue 1, and all its other eigenvectors have eigenvalues with absolute value less than 1.

Some Additional Properties of Markov Chains

Definition 9.4.19 (Reversible)**.** MC is called *reversible* if there exists π^* s.t.

$$\forall\{i,j\} \in \mathcal{E}: \quad p_{ji}\pi_i^* = p_{ij}\pi_j^*, \tag{9.72}$$

where $\{i,j\}$ is our notation for the undirected edge, assuming that both directed edges $(i \leftarrow j)$ and $(j \leftarrow i)$ are elements of the set $\mathcal{E}$.

In physics this property is also called *Detailed Balance* (DB). If one introduces the so-called ergodicity matrix

$$Q := (Q_{ji} = p_{ji}\pi_i^* | (j \leftarrow i) \in \mathcal{E}), \tag{9.73}$$

then DB translates into the statement that Q is symmetric, $Q = Q^T$. The MC for which the property does not hold is called *irreversible*. $Q - Q^T$ is nonzero, i.e. Q is asymmetric for reversible MC. An asymmetric component of Q is the matrix built from currents/flows (of probability). Thus for the case shown in Fig. (9.3)

$$Q = \begin{pmatrix} 0.7 * 0.625 & 0.5 * 0.375 \\ 0.3 * 0.625 & 0.5 * 0.375 \end{pmatrix} = \begin{pmatrix} 0.4375 & 0.1875 \\ 0.1875 & 0.1875 \end{pmatrix} \tag{9.74}$$

Q is symmetric, i.e. even though $p_{12} \neq p_{21}$, there is still no flow of probability from 1 to 2 as the "population" of the two states, π_1^* and π_2^* respectively are different, $Q_{12} - Q_{21} = 0$. In fact, one observes that in the two node situation the steady state of the MC is always in DB.

Note that if a steady distribution, π^*, satisfies the DB condition (9.72) for a MC, $(\mathcal{V}, \mathcal{E}, p)$, it will also be a steady state of another MC, $(\mathcal{V}, \mathcal{E}, \tilde{p})$, satisfying the more general Balance (or global balance) B-condition

$$\sum_{j:(j\leftarrow i)\in\mathcal{E}} \tilde{p}_{ji}\pi_i^* = \sum_{j:(i\leftarrow j)\in\mathcal{E}} \tilde{p}_{ij}\pi_j^*. \tag{9.75}$$

This suggests that many different MC (many different dynamics) may result in the same steady state. Obviously DB is a particular case of the B-condition (9.75).

The difference between DB- and B- can be nicely interpreted in terms of flows (think water) in the state space. From the hydrodynamic point of view reversible MCMC corresponds

to irrotational probability flows, while irreversibility relates to nonzero rotational part, e.g. correspondent to vortices contained in the flow. Putting it formally, in the irreversible case skew-symmetric part of the ergodic flow matrix, $Q = (\tilde{p}_{ij}\pi^*_j|(i \leftarrow j))$, is nonzero and it actually allows the following cycle decomposition,

$$Q_{ij} - Q_{ji} = \sum_\alpha J_\alpha \left(C^\alpha_{ij} - C^\alpha_{ji}\right), \tag{9.76}$$

where index α enumerates cycles on the graph of states with the adjacency matrices C^α. Then, J_α, stands for the magnitude of the probability flux flowing over the cycle α.

One can use the cycle decomposition to modify MC such that the steady distribution stay the same (invariant). Of course, cycles should be added with care, e.g. to make sure that all the transition probabilities in the resulting $\tilde{p}$, are positive (stochasticity of the matrix will be guaranteed by construction). The procedure of "adding cycles" along with some additional tricks (e.g. the so-called lifting/replication) may help to improve mixing, i.e. speed up convergence to the steady state — which is a very desirable property for sampling π^* efficiently.

Example 9.4.20. Given a stationary solution, $\pi^* = (\pi^*_1, \pi^*_2, \pi^*_3)$, construct a three-state Markov chain, i.e. present a (3×3) transition matrix, p, (a) of a general position (satisfies global balance), (b) satisfies detailed balance. Are the constructions unique? Find the spectrum of the transition matrix in the case (b) and verify that the Perron-Frobenius theorem 9.4.18 holds. In the case (b) formulate and solve example of the fastest mixing MC. Can one generalize solution and find the fastest mixing MC of size n, given $\pi^* = (\pi^*_1, \cdots, \pi^*_n)$? Return back to the three state MC and impose the constraint that all the diagonal elements of the transition probability (correspondent to self-loops on the fully connected three state graph) are zero, $p(1,1) = p(2,2) = p(3,3) = 0$. Is the MC unique in this case? Is it ergodic?

Solution. See Mathematica snippet MC-3nodes.nb (or pdf-printout MC-3nodes.pdf) posted at the D2L site of the course. □

Example 9.4.21. Let $\Sigma = \{x_0, x_1, \cdots, x_{K-1}\}$ be K equidistant points on the circle, i.e., $x_k = e^{2\pi ik/K}$. Let $\alpha, \beta \in (0,1)$ be constants that satisfy $\alpha + \beta + \gamma = 1$, and consider the random walk (X_t) defined by

$$P(X_{t+1} = x_{k+1}|X_t = x_k) = \alpha, \tag{9.77a}$$

$$P(X_{t+1} = x_{k-1}|X_t = x_k) = \beta, \tag{9.77b}$$

$$P(X_{t+1} = x_k \quad |X_t = x_k) = \gamma. \tag{9.77c}$$

Let π be the unique stationary distribution.

(a) For what values of α, β, and γ (and K) is the Markov chain ergodic?

(b) What is the stationary distribution? (*intuitive arguments preferred.*)

(c) For what values of α and β does the Markov chain satisfy detailed balance?

(d) Let p denote the transition matrix. Find exact expressions for the eigenvalues of p. *Hint: the linear transformation represented by p is a convolution operator, i.e., there is a g such that $(pv)_k = (g \star v)_k = \sum_\ell g(x_{k-\ell})v(x_\ell)$ for all k-vectors v (identifying k vectors with functions on Σ), and thus can be diagonalized by the discrete Fourier transform.*

(e) The *spectral gap* of p is $1 - |\lambda'|$, where λ' is the second largest (in absolute value) eigenvalue of p. The size of the spectral gap determines how fast an ergodic chain converges to its stationary distribution: the larger the gap, the faster the convergence. Suppose $\gamma = 0.98$ and $\alpha = \beta$. Use the result of the previous part to find the spectral gap of p to leading order in $1/K$ as $K \to \infty$.

(f) Are there initial distributions that converge to the stationary distribution at a rate faster than the second largest eigenvalue? If so, give an example. If not, explain why not.

Solution.

(a) The Markov chain is irreducible if $\gamma < 1$ and aperiodic if $\gamma > 0$. (If K is odd, then $\alpha, \beta > 0$ is also sufficient for aperiodicity.

(b) First, observe that the stationary distribution satisfies

$$\beta\pi(x_{k+1}) + \alpha\pi(x_{k-1}) + \gamma\pi(x_k) = \pi(x_k). \tag{9.78}$$

If we set $\pi(x) = 1/K$, we see that this solves the equation. Since the chain is irreducible, the uniform distribution is therefore the unique stationary distribution.

(c) To satisfy detailed balance, we must have

$$\beta\pi(x_{k+1}) = \alpha\pi(x_{k-1}). \tag{9.79}$$

This holds if and only if $\beta = \alpha$. So the chain satisfies detailed balance whenever $\beta = \alpha = (1 - \gamma)/2$.

(d) The transition matrix has the form

$$p = \begin{bmatrix} \gamma & \alpha & 0 & 0 & \cdots & \beta \\ \beta & \gamma & \alpha & 0 & \cdots & 0 \\ 0 & \beta & \gamma & \alpha & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \beta & \gamma & \alpha \\ \alpha & \cdots & 0 & 0 & \beta & \gamma \end{bmatrix}, \tag{9.80}$$

i.e., p is "circulant." The matrix acts by convolution on K-vectors, and so can be diagonalized by the discrete Fourier transform. That is, if we let U be the $K \times K$ matrix with elements

$$U_{k\ell} = \exp\left(\frac{2\pi i k\ell}{K}\right), \quad k, \ell \in \{0, \cdots, K-1\}, \tag{9.81}$$

then $p = U\Lambda U^{-1}$, where Λ is diagonal. We can check this directly: let u_ℓ denote the ℓth column of U. Then

$$pu_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad pu_1 = \begin{bmatrix} (\gamma + \alpha e^{2\pi i/K} + \beta e^{-2\pi i/K}) \cdot 1 \\ (\gamma + \alpha e^{2\pi i/K} + \beta e^{-2\pi i/K}) \cdot e^{2\pi i/K} \\ \vdots \\ (\gamma + \alpha e^{2\pi i/K} + \beta e^{-2\pi i/K}) \cdot e^{2\pi i(K-1)/K} \end{bmatrix}, \cdots. \tag{9.82}$$

In general, the eigenvalues are

$$\begin{aligned} \lambda_\ell &= \gamma + \alpha e^{2\pi i\ell/K} + \beta e^{-2\pi i\ell/K} \\ &= \gamma + (\alpha + \beta)\cos(2\pi\ell/K) + i(\alpha - \beta)\sin(2\pi\ell/K). \end{aligned}$$

(e) From the above, we have

$$|\lambda_\ell|^2 = \Big(\gamma + (\alpha + \beta)\cos(2\pi\ell/K)\Big)^2 \tag{9.83}$$

Setting $\alpha = \beta = (1-\gamma)/2$, we have

$$|\lambda_\ell|^2 = \Big(\gamma + (1-\gamma)\cos(2\pi\ell/K)\Big)^2. \tag{9.84}$$

To find the second largest eigenvalue, we observe that $|\lambda_\ell|^2 = f(2\pi\ell/K)$ where

$$f(\theta) = \Big(\gamma + (1-\gamma)\cos(\theta)\Big)^2. \tag{9.85}$$

A simple calculation shows that the only critical points of f are $\theta = 0$ and $\theta = \pi$, with $f(0) = 1$ and $f(\pi) = 2\gamma - 1 \in (0, 1)$. By continuity, when K is large we expect the second largest eigenvalue to occur at $\ell = \pm 1$, i.e.,

$$\lambda_{\pm 1} = \gamma + (1-\gamma)\cos(2\pi/K). \tag{9.86}$$

To leading order in $1/K$, this is $1 - 2\pi^2/K^2$, and thus the gap is $2\pi^2/K^2$.

(f) If we take the eigen-vector u_ℓ for $\ell \notin \{0, 1\}$, then $|\lambda_\ell| < |\lambda_1|$. By orthogonality, the entries of u_ℓ sum to 0, and $Re(u_\ell)$ lies in the eigen-space of λ_ℓ. Therefore any initial distribution of the form $\pi + aRe(u_\ell)$ would converge to the stationary distribution with a rate given by $|\lambda_\ell|$.

Exercise 9.5 (Hardy-Weinberg Law)**.** Consider an experiment of mating rabbits. We follow the inheritance of a particular gene that appears in two types, G or g. A rabbit has a pair of genes, either GG (dominant), Gg (hybrid — the order is irrelevant, so gG is the same as Gg) or gg (recessive). In mating two rabbits, the offspring inherits a gene from each of its parents with equal probability. Thus, if we mate a dominant (GG) with a hybrid (Gg), the offspring is dominant with probability 1/2 or hybrid with probability 1/2. Start with a rabbit of given character (GG, Gg, or gg) and mate it with a hybrid. The offspring produced is again mated with a hybrid, and the process is repeated through a number of generations, always mating with a hybrid.

Note: The first experiment of such kind was conducted in 1858 by Gregor Mendel. He started to breed garden peas in his monastery garden and analyzed the offspring of these matings.

(a) Write down the transition matrix P of the Markov chain thus defined. Is the Markov chain irreducible and aperiodic?

(b) Assume that we start with a hybrid rabbit. Let $\pi(n) = (\pi_{GG}(n), \pi_{Gg}(n), \pi_{gg}(n))$ be the probability distribution vector (state) of the character of the rabbit of the n-th generation. In other words, $\pi_{GG}(n)$, $\pi_{Gg}(n)$, $\pi_{gg}(n)$ are the probabilities that the n-th generation rabbit is GG, Gg, or gg, respectively. Compute $\pi(1), \pi(2), \pi(3)$. Is there some kind of law?

(c) Calculate P^n for general n. What can you say about $\pi(n)$ for general n?

(d) Calculate the stationary distribution of the MC, $\pi^* = (\pi^*_{GG}, \pi^*_{Gg}, \pi^*_{gg})$. Does Detailed Balance hold in this case?

9.5 Stochastic Optimal Control: Markov Decision Process

We have discussed in the preceding sections of this chapter Markov Processes. We have also discussed Optimal Control two chapters ago. Let us now merge the two topics and consider what comes under the name of the *Markov Decision Process* (MDP).

Specifically, a MDP represents a stochastic, discrete space, discrete time version of the stochastic optimal control problem, which is stated as follows [21]:

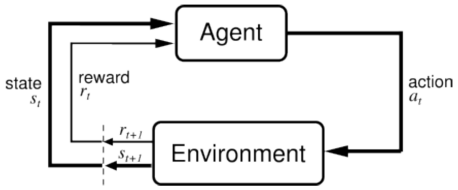


Figure 9.10: General Scheme of Markov Decision Process. Drawing is from [21].

- Given:
  - Set of states, S [nodes of a graph, or squares if we play the example of the 4×4 grid World discussed in the following].
  - Set of actions, A [associated with arrows connecting nodes/squares].
  - $\mathcal{P} : S \times A \times S \to [0, 1]$ - Transition probabilities between states, dependent on actions and sliced (sequentially) in time, $\mathcal{P}(s, a, s') = P(s_{t+1} = s' | s_t = s, a_t = a)$.
  - $R : S \times A \times S \to \mathbb{R}$ – rewards/costs, $R(s, a, s')$, for $s_{t+1} = s'$- next state, $s_t = s$ current state, a_t current action taken.
  - We consider the problem over the infinite time horizon.
  - γ^t discount factor (less reward as time progresses).
- Goal:
  - to maximize the expected sum of rewards over the policy, $\pi : S \to A$, defined as a function mapping s_t to a_t:

$$\pi^* = \arg\max_{\pi(\cdot)} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t), s_{t+1})\right], \tag{9.87}$$

    where averaging is over the random transitions, governed by $\mathcal{P}(s, a, s')$.

Notice that in the present formulation, $R(\cdot, \cdot, \cdot)$, $\mathcal{P}(\cdot, \cdot, \cdot)$ do not depend explicitly on time. (Generalizations are possible but we will not discuss these in the lectures leaving it for an independent studies.)

9.5.1 Bellman Equation & Dynamic Programming

Expectation on the right hand side of Eq. (9.87) is called the global reward, i.e. the reward accumulated over the entire (infinite) time horizon under the given (not necessarily optimal)

policy, π. However, it has a sense to discuss not only the global reward but also respective expected value of the reward, evaluated for the time horizon τ, called the value function:

$$\forall \tau \in [1, \cdots, \infty],\ \forall s_0 :\ V^{\pi}_{\tau}(s_0) \ := \ \mathbb{E}_{s_1,s_2,\cdots}\left[\sum_{t=0}^{\tau-1}\gamma^t R(s_t, \pi(s_t), s_{t+1})\right] \tag{9.88}$$

$$= \sum_{s_1,\cdots,s_\tau}\left(\prod_{t'=0}^{\tau-1}\mathcal{P}(s_{t'}, \pi(s_{t'}), s_{t'+1})\right)\sum_{t=0}^{\tau-1}\gamma^t R(s_t, \pi(s_t), s_{t+1}),$$

which depends on the initial state, s_0, and the policy, $\pi(\cdot)$. Observe that the right hand side of Eq. (9.88) can also be expressed in terms of the value function evaluated at the preceding, $\tau - 1$, step:

$$= \sum_{s_1,\cdots,s_\tau}\left(\prod_{t'=0}^{\tau-1}\mathcal{P}(s_{t'}, \pi(s_{t'}), s_{t'+1})\right)\left(R(s_0, \pi(s_0), s_1) + \gamma\sum_{t=0}^{\tau-2}\gamma^t R(s_{t+1}, \pi(s_{t+1}), s_{t+2})\right)$$

$$= \mathbb{E}_{s'}\left[R(s_0, \pi(s_0), s') + \gamma V^{\pi}_{\tau-1}(s')\right]. \tag{9.89}$$

Let us now introduce the current optimal (over policy) value

$$\forall \tau,\ \forall s :\ V^{*}_{\tau}(s) := \max_{\pi(\cdot)} V^{\pi}_{\tau}(s).$$

Then, optimizing both sides of Eqs. (9.89) over policy we derive

$$\forall \tau,\ \forall s :\ V^{*}_{\tau}(s) = \max_{a}\mathbb{E}_{s'}\left[R(s, a, s') + \gamma V^{*}_{\tau-1}(s')\right]. \tag{9.90}$$

The recursion, suggested by Bellman in his seminal 1948 paper, shows that the optimal solution of Eq. (9.87) can be found by solving Eq. (9.90). Moreover, once the optimal value at the step τ is found we can also find the so-called optimal policy:

$$\forall \tau,\ \forall s :\ \pi^{*}_{\tau}(s) = \arg\max_{a}\sum_{s'}\mathcal{P}(s, a, s')\left[R(s, a, s') + \gamma V^{*}_{\tau-1}(s')\right], \tag{9.91}$$

defined as a function/map from S to A. These, Eqs. (9.90,9.91), called Bellman equations, represent yet another example of what also comes under the (already familiar) name of the Dynamic Programming.

Few important remarks are in order:

- Recurrent Eqs. (9.90) translate into the **value-iteration** Algorithm 5, which we illustrate on the example of the Grid World below.

- Similarly to the state-dependent value function, one may also introduce the value of taking action a_0 in a state s_0 under a policy π:

$$\forall \tau, s_0, a_0 :\ Q^{\pi}_{\tau}(s_0, a_0) = \mathbb{E}_{s_1,s_2,\cdots}\left[R(s_0, a_0, s_1) + \sum_{t=1}^{\tau-1}\gamma^t R(s_t, \pi(s_t), s_{t+1})\right]$$

$$= \mathbb{E}_{s'}\left[R(s_0, a_0, s') + \gamma V^{\pi}_{\tau-1}(s')\right]. \tag{9.92}$$

Therefore, instead of working with $V^{\pi}_{\tau}(s_0)$ we can re-state the entire DP approach and the optimization procedure in terms of $Q^{\pi}_{\tau}(s_0, a_0)$ – the **action-value** function of the state s_0 and action a_0 under policy π – also called the Q-function. In particular, the action-value version of Eq. (9.89) becomes

$$\forall \tau, s_0, a_0 : \; Q^{\pi}_{\tau}(s_0, a_0) = \mathbb{E}_{s'} \left[R(s_0, a_0, s') + \gamma \max_{a'} Q^{\pi}_{\tau-1}(s', a') \right]. \tag{9.93}$$

- There exists an alternative iterative way of solving the optimization problem (9.87) via a **policy-iteration** algorithm. In this case the approach is to alternate between the following two steps till convergence: (a) **policy evaluation**: for a current (not optimal) policy run the value iteration algorithm solving Eqs. (9.89) for either fixed number of steps (in τ) or till a pre-defined tolerance is achieved; (b) **policy improvement**: according to

$$\forall \tau, \; \forall s_0 : \; \pi_{\tau}(s) \leftarrow \arg\max_{a} \sum_{s'} \mathcal{P}(s, a, s') \left[R(s, a, s') + \gamma V^{\pi}_{\tau-1}(s') \right].$$

  This policy iteration approach may be algorithmically advantageous when policy "freezes" faster than the value/cost.

9.5.2 MDP: Grid World Example

MDP may be considered as an interactive probabilistic game one plays (against computer). The game consists in defining transition rates between the states to achieve certain objectives. Once optimal (or sub-optimal) rates are fixed the implementation becomes just a Markov Process.

Let us play this 'Grid World' game with the rules illustrated in Fig. (9.11). An agent lives on the grid (3×4). Walls block the agent's path. The agent actions do not always go as planned: for example 80% of time the action 'North' take the agent 'North' (if there is no wall there), 10% of the time the action 'North' actually takes the agent West; 10% East. If there is a wall the agent would have been taken, she stays put. Big reward, $+1$, or penalty, -1 comes at the end — the game ends after reaching either of the two states. Visiting any other state does not result in a reward.

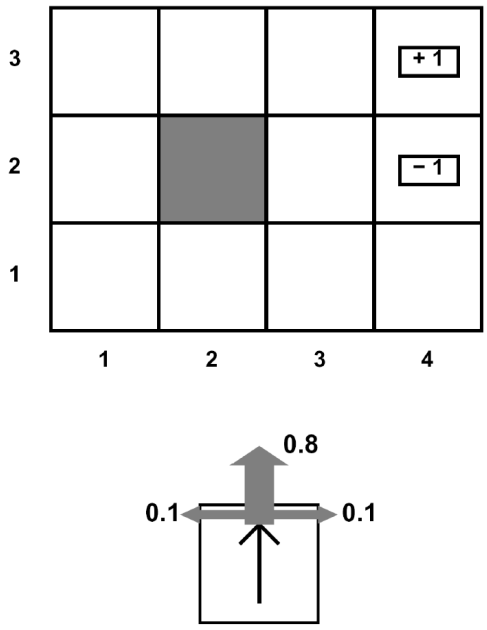


Figure 9.11: Canonical example of MDP from 'Grid World' game.

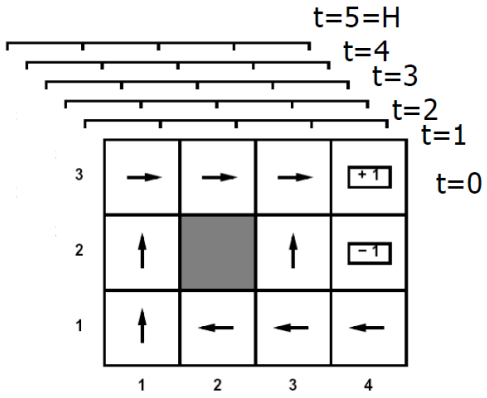


Figure 9.12: Optimal solution set of actions (arrows) for each state, for each time.

Figure 9.13: Value Iteration in Grid World.

Let us now translate these rules into formulas:

$$s \in \{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,4),(3,1),(3,2),(3,3),(3,4)\}$$

$$a \in \{\uparrow,\downarrow,\leftarrow,\rightarrow\}$$

$$P\left((2,1)|(1,1),\uparrow\right)=0.8,\ P\left((1,1)|(1,1),\uparrow\right)=0.1,\ P\left((1,2)|(1,1),\uparrow\right)=0.1,$$

$$P\left((1,2)|(1,1),\rightarrow\right)=0.8,\ P\left((1,1)|(1,1),\rightarrow\right)=0.1,\ P\left((2,1)|(1,1),\rightarrow\right)=0.1,\cdots.$$

$$R(s,a,s')=\begin{cases}+1, & s=(3,4)\\ -1, & s=(2,4)\\ 0, & s\neq(3,4),(2,4).\end{cases}$$

$V_\tau^*(s)$ is the expected sum of rewards accumulated when starting from state s and acting optimally for a horizon of τ steps.

To find the optimal actions (policy), one may consider the *Value Iteration* algorithm (MDP – Value Iteration). One can show that the algorithms outputs solution of the value iteration Eqs. (9.90).

The MDP value iteration algorithm is illustrated for the Grid World example in Fig. (9.13).

Algorithm 5 MDP – Value Iteration (finite horizon version)

Input: Set of states, S; set of actions, A; Transition probabilities between states, $P(s'|s,a)$; rewards/costs, $R(s,a,s')$; γ discount factors

$\forall s: \quad V_0^*(s) = 0$

for $\tau = 0, \cdots, H-1$ **do**

$\forall s: \quad V_{\tau+1}^*(s) \leftarrow \max_a \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V_\tau^*(s')\right]$ – [Bellman update/back-up] – the expected sum of rewards accumulated when starting from state s and acting optimally for a horizon of $\tau+1$ steps

end for

Some of the numbers (for the optimal value function) at the first three iterations, are derived as follows:

$$\begin{aligned}
V_1^*((3,3)) &= P((3,4)|(3,3),\rightarrow) * V_0^*((3,4)) * \gamma = 0.8 * 1 * 0.9 \approx 0.72, \\
V_2^*((3,3)) &= (P((3,4)|(3,3),\rightarrow) * V_1^*((3,4)) + P((3,3)|(3,3),\rightarrow) * V_1^*((3,3))) * \gamma \\
&\approx (0.8 * 1 + 0.1 * 0.72) \approx 0.78, \\
V_3^*((2,3)) &= (P((3,3)|(2,3),\uparrow) * V_2^*((3,3)) + P((2,4)|(2,3),\uparrow) * V_2^*((2,4))) * \gamma \\
&\approx (0.8 * 1 * 0.72 + 0.1 * (-1)) * 0.9 \approx 0.43.
\end{aligned}$$

We also observe (empirically) that as $\tau \rightarrow \infty$ the optimal value functions show a well defined limit, i.e. the expected sum of rewards accumulated when acting optimally freezes (becomes stationary) in the limit.

Exercise 9.6. Consider the following modification of the Grid World example: no discount, $\gamma = 1$; the terminal penalty at $s = (2,4)$ increases by absolute value (from -1) to -2, i.e.

$$R(s,a,s') = \begin{cases} +1, & s = (3,4) \\ -2, & s = (2,4) \\ 0, & s \neq (3,4), (2,4). \end{cases} .$$

(a) Compute optimal values of the expected sum of rewards for the first three iterations of the Algorithm 5, i.e. find $\forall s: \; V_1^*(s), V_2^*(s), V_3^*(s)$;
(b) Find optimal policy for the first three iterations, i.e. find $\forall s: \quad a_1^* = \pi_1^*(s), a_2^* = \pi_2^*(s), a_3^* = \pi_3^*(s)$.
(c) Re-state the original MDP formulation (9.87) as a **linear programming**. [Hint: take advantage of the linearity of the value iteration way of solving the MDP according to Eqs. (9.90).]

(d) What we have discussed so far in this Section is the case of a deterministic policy, where π is a map/function from S to A. It may be advantageous to consider **stochastic policy** where for each state a number of actions can be taken with different probabilities, in which case $\pi(s, a)$ becomes function of two variables (state and action) describing the probability of taking action a when the state is s. Suggest stochastic policy modification of Eq. (9.87).

We will return to the MDP (and the grid world example) in Section **??**, discussing reinforcement learning.

9.6 Queuing Networks *

9.6.1 Queuing: a bit of History & Applications

* There are number of books written on the subject. The book of Frank Kelly and Elena Ydovina [22] is recommended.

Agner Krarup Erlang, a Danish engineer who worked for the Copenhagen Telephone Exchange, published the first paper on what would now be called queueing theory in 1909. He modeled the number of telephone calls arriving at an exchange by a Poisson process and solved the $M/D/1/\infty$ queue in 1917 and $M/D/k/\infty$ queuing model in 1920.

The notations are now standard in the Queuing theory – which is a discipline traditionally considered as a part of Operation Research with deep connection to stochastic processes. In $M/D/k/\infty$, for example,

- **M** stands for Markov or memoryless and it means that arrivals occur according to a Poisson process. Arrivals may also be deterministic, D.
- **D** stands for deterministic and means that the jobs arriving at the queue require a fixed=deterministic amount of service/processing. Processing can also be stochastic, Markovian (or non-Markovian, in which case it is custom to mark it as G - generic sevice; arrival can also be G=generic).
- k describes the number of servers at the queueing node $k = 1, 2, ...$ If there are more jobs at the node than there are servers then jobs will queue and wait for service.
- ∞ stands for the allowed size of the queue (waiting room) - in this case no limit to the waiting room capacity (everybody arriving is admitted to the queue - not denied)

*This is an Auxiliary Section which can be dropped at the first reading. Material from the Section will not contribute midterm and final exams.

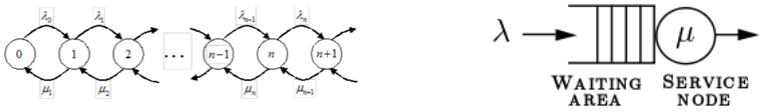


Figure 9.14: On the left: Markov Chain representation of the M/M/1 queue. In the standard situation considered $\forall i: \quad \lambda_i = \lambda, \mu_i = \mu$. On the right: reduced graphical description of a single queue.

We will only be dealing with the case of ∞ waiting room, thus dropping the last argument.

The $M/M/1$ queue is a simple model where a single server handles jobs that arrive according to a Poisson process and have exponentially distributed service requirements.

In an M/G/1 queue the G stands for general and indicates an arbitrary probability distribution.

Many mathematicians and math-engineers contributed the subject since 1930 — Pollaczek, Khinchin, Kendall, Kingman, Jackson, Kelly and others.

Applications: call centers, logistics (at different scales), manufacturing, checkout at the super-market, processing of electric vehicles at the charging stations, etc. In general, any kind of practical systems where arrivals (of whatever coming in units) and processing fits the framework. We are talking about design which would

- Manage queue (controls its size).
- Keep processing units busy (good utilization).
- Have waiting time in the queue under control.

9.6.2 Single Open Queue = Birth/Death process. Markov Chain representation.

Let us discuss in details $M/M/1$. We start by playing with the Java modeling tool – JMT (can upload it from `http://jmt.sourceforge.net/Download.html`)

The process is also called birth-death process - the name is clear from the Markov-Chain representation shown in Fig. (9.14). The MC has infinitely many states, each representing # of customers in the system (waiting room). Arrival of customers is modeled as the Poisson process with the arrival rate, λ. We assume that all customers are identical. The customers are taken from the waiting room based on availability of the servant, and the service is completed with the rate μ of the other Poisson process.

Everything is Poisson in here (recall that mixing and splitting of the Poisson processes is Poisson again).

Let us analyze the (relatively simple) system. Let us start from finding the steady state of the Markov Chain: $\forall i = 0, \cdots, \infty \quad P_i$, where P_i is the probability that the system is in the i-th state, i.e. with i customers in the queue.

The balance equations are

$$\text{\# 0 customers:} \quad \underbrace{\mu P_1}_{arrival} = \underbrace{\lambda P_0}_{departure} \tag{9.94}$$

$$\text{\# 1 customer:} \quad \lambda P_0 + \mu P_2 = (\lambda + \mu) P_1 \tag{9.95}$$

$$\text{\# n customers:} \quad \lambda P_{n-1} + \mu P_{n+1} = (\lambda + \mu) P_n \tag{9.96}$$

Resolving the equations (sequentially), and requiring that the total probability is normalized, $\sum_{i=0}^{\infty} P_i = 1$, we derive

$$P_n = \left(\prod_{i=0}^{n-1} \frac{\lambda}{\mu} \right) P_0 = \left(\frac{\lambda}{\mu} \right)^n P_0 = \rho^n P_0 \tag{9.97}$$

$$1 = \sum_{n=0}^{\infty} P_n = P_0 \sum_{n=0}^{\infty} \rho^n = \frac{P_0}{1-\rho} \tag{9.98}$$

$$P_n = (1-\rho)\,\rho^n. \tag{9.99}$$

where $\rho := \lambda/\mu$ is the traffic intensity.

The average size of the queue is:

$$\sum_{n=0}^{\infty} n P_n = (1-\rho) \sum_{n=0}^{\infty} n \rho^n = \frac{\rho}{1-\rho}. \tag{9.100}$$

We observe that the average queue becomes infinite at $\rho = 1$, i.e. the steady state exists only when $\rho < 1$. This criterium (existence of the steady state) can also be referred to as "stability".

Exercise: Consider a single M/M/m queue, i.e. the system when the number of servers is m. Derive steady state? What is the modified stability criterium? Can a single queue system withe $m = 2$ be unstable?

In this simple queue system we can also study transient time dynamics. The steady state system of Eqs. (9.96) transitions to

$$\forall n: \quad \frac{d}{dt} P_n = \underbrace{\lambda P_{n-1} + \mu P_{n+1}}_{\text{arrival}} - \underbrace{(\lambda + \mu) P_n}_{\text{departure}}. \tag{9.101}$$

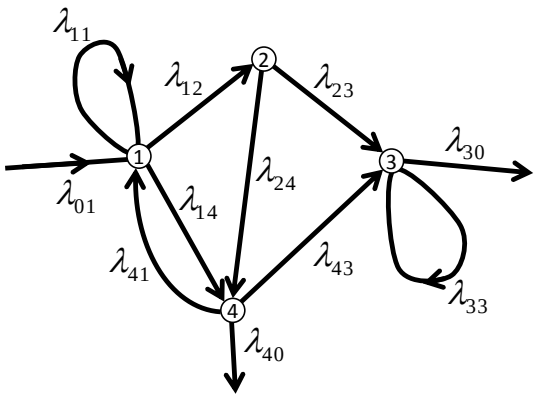


Figure 9.15: Example of a Queuing network.

Solution of this system can be found in an analytic form

$$\begin{aligned} P_k(t) &= e^{-(\lambda+\mu)t}\Bigg(\rho^{(k-i)/2} I_{k-i}(at) \\ &+ \rho^{(k-i-1)/2} I_{k+i+1}(at) + (1-\rho)\rho^k \sum_{j=k+i+2} \rho^{-j/2} I_j(at)\Bigg) \end{aligned} \tag{9.102}$$

where $a := 2\sqrt{\lambda\mu}$, $I_k(x)$ is the modified Bessel function of the first kind, where it is assumed that the system was in the state i at $t = 0$.

Example 9.6.1. Derive Eq. (9.102) from Eq. (9.101). Compute distribution time of the busy period of server. Assuming first come-first served policy, compute distribution of the waiting time and distribution of the total time in the system.

We can also write the dynamical Eq. (9.101) in the following matrix form

$$\frac{d}{dt}P = P^{(\mathrm{tr})}P, \quad P^{(\mathrm{tr})} = \begin{pmatrix} \ddots & & & & & & \\ \cdots & 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \cdots \\ & & & & & & \ddots \end{pmatrix}. \tag{9.103}$$

Notice that in the steady state (achievable at $\rho < 1$) the Detailed Balance (DB) does not hold, $P^{(\mathrm{tr})}_{nm} P^{(\mathrm{st})}_m \neq P^{(\mathrm{tr})}_{mn} P^{(\mathrm{st})}_n$.

9.6.3 Generalization to (Jackson) Networks. Product Solution for the Steady State.

It appears that the description of a single Q- can be extended to a network, e.g. of the type shown in Fig. (9.15). λ are arrivals and processing rates – now denoted in the same way

and indexed by nodes/stations from where the process is coming from and where it is going to (two indexes). We study, $P(n_1, \cdots, n_N; t)$, probability over the entire network and, like in the single queue case, write the balance (also called Master) equation – stated for any state of the network at any time

$$\frac{\partial}{\partial t}P(\boldsymbol{n};t) = \sum_{(i,j)\in\mathcal{E}} \lambda_{ij}\Bigg(\underbrace{(n_i+1)P(\cdots,n_i+1,\cdots,n_j-1,\cdots;t)}_{\text{customers leaving } i \text{ for } j} - \underbrace{n_i P(\cdots,n_i,\cdots,n_j,\cdots;t)}_{\text{customers staying at } i}\Bigg)$$
$$+\sum_{i\in\mathcal{V}} \lambda_{0i}\left(P(\cdots,n_i-1,\cdots;t) - P(\cdots,n_i,\cdots;t)\right)$$
$$+\sum_{i\in\mathcal{V}} \lambda_{i0}\left((n_i+1)P(\cdots,n_i+1,\cdots;t) - n_i P(\cdots,n_i,\cdots;t)\right) \tag{9.104}$$

This equation is written here for the case of $M/M/\infty$ - when the number of servers is infinite - this is the case when jobs are not waiting but are taken for processing (by tellers which are always available) immediately.

Example 9.6.2. Write down a $M/M/m$ version of Eq. (9.104).

Remarkably the (complicated looking) Eq. (9.104) allows an explicit steady state solution (for any graph!)

$$P(\boldsymbol{n}) = Z^{-1}\prod_{i\in\mathcal{E}}\frac{h_i^{n_i}}{\prod_{l_i=1}^{n_i} l_i},$$
$$\forall i\in\mathcal{V}: - \quad h_i\sum_{j\neq 0}^{(i,j)\in\mathcal{E}}\lambda_{ij} + \sum_{j\neq 0}^{(j,i)\in\mathcal{E}}\lambda_{ji}h_j + \lambda_{0i} - \lambda_{i0}h_i = 0$$

which is also called product solution/factorization (name is according to the structure).

Few important things to mention in here

- This is a product form solution which IS NOT a Gibbs (equilibrium) distribution. [Remember our discussions of the Fokker-Planck.]
- The system is stable (solution is finite) if: $\forall i\in\mathcal{V}, \quad h_i < m_i$
- $\boldsymbol{h}$ is a "single-customer" object

Example 9.6.3. Generalize the steady state formula and reformulate the stability for the general $M/M/m$ case.

Example 9.6.4. By analogy with what was discussed for a single queue case, state the skewed detailed balance relation. Show how analysis of the skewed DB leads to the product state solution for the steady state.

9.6.4 Heavy Traffic Limit

Our discussion here is (mainly) based on the material from `http://www.columbia.edu/~ww2040/A1a.html`.

The Heavy traffic limit applies when either of the two cases (or some special combination of the two, which we will not discuss here) takes place (we discuss a single queue case - to make the arguments simpler):

- The number of servers is fixed and the traffic intensity (utilization), λ/μ, approaches unity (from below). The queue length approximation is the so-called "reflected Brownian motion".
- Traffic intensity is fixed and the number of servers and arrival rate are increased to infinity. Here the queue length limit converges to the normal distribution.

Let us give some intuitive picture and then pose a number of technical questions/challenges (some of these with answers not yet fully known).

When the system is congested, i.e. when most of the time it has many customers, an arriving customer will need to wait long. Assuming FIFO (first in first out) protocol, the customer joining queue with L customers will see the queue going down to zero. However, in this time of $L+1$ departures, there will also be many arrivals. If traffic intensity, ρ, is close to 1, the number of arrivals will be on order L as well. Thus, when the customer leaves the queue behind will be comparable to the queue observed when the customer arrived. For the system to go from average to empty will take more like L busy periods.

Example 9.6.5. Assume that $1-\rho \ll 1$ and estimate

- How much time a typical type customer spend in the system ?
- How long does it take for a system to change from an average/typical filling to empty?

Hint (following from the preceding the discussions): The time scale at which the system changes is much longer then the time scale of a single customer.

We therefore have a time scale separation and, therefore, may study a Q-system with many customers on two scales, fluid and diffusive. Let $X(t)$ be some Q-system related process. $\bar{X}_n(t) = nX(nt)$ defines the fluid re-scaling by n. This means that we measure time in the units of n and we measure the state (# of customers) in the units of n. As $n \to \infty$ we shall look for $n^{-1}X(nt) \to \bar{X}(t)$, where $\bar{X}(t)$ is the fluid limit.

At this scale as $n \to \infty$ the arrival process and the service process have fluid limits λt and μt which means that they are deterministic. As we said, queueing is the result of

variability, and so on a fluid scale, **when input and output are not variable, there will be no real queueing behavior in the system**. We may see the queue length grow linearly indefinitely ($\rho > 1$), or go to zero linearly and then stay at 0 ($\rho < 1$), or we may see it constant, ($\rho = 1$). For queueing networks we may observe piecewise linear behavior of queue lengths. This will capture changes in the queue on the fluid scale: The queue changes by n in a time of order n. The stochastic fluctuations of a queue in steady state are scaled down to be identically 0 and uninteresting.

The diffusion scaling looks at the difference between the process and its fluid limit, and measures the time in units of n and the state (counts of customers) in units of $\sqrt{n}$. The diffusion re-scaling of $A(t)$ by n is $\hat{A}_n(t)$ by n is $\hat{A}_n(t) = \sqrt{n}(\bar{A}_n(t) - \bar{A}(t))$. As $n \to \infty$ we shall look (in analogy with the Central Limit Theorem) for $A_n(t)$ converging in the sense of distribution to $\hat{A}(t)$ describing the diffusion limit- it is a diffusion process, such as Brownian motion or reflected Brownian motion. The diffusion limit captures the random fluctuation of the system around its fluid limit.

Here is a (formal) statement on the heavy traffic asymptotic for the waiting time (including both fluid and diffusive limits): Consider $G/G/1$ indexed by j. For queue j let T_j denotes the random inter-arrival time, S_j denote the random service time; $\rho_j = \frac{\lambda_j}{\mu_j}$ denote the traffic intensity with $\frac{1}{\lambda_j} = \mathbb{E}(T_j)$ and $\frac{1}{\mu_j} = \mathbb{E}(S_j)$; $W_{q,j}$ denotes the waiting time in queue for a customer in steady state; $\alpha_j = -\mathbb{E}[S_j - T_j]$ and $\beta_j^2 = \text{var}[S_j - T_j]$. If $T_j \xrightarrow{d} T$, $S_j \xrightarrow{d} S$, and $\rho_j \to 1$, then $\frac{2\alpha_j}{\beta_j^2} W_{q,j} \xrightarrow{d} \exp(1)$ provided that: (a) Var[S-T] > 0, and (b) for some $\delta > 0, \mathbb{E}[S_j^{2+\delta}]$ and $\mathbb{E}[T_j^{2+\delta}]$ are both less than some constant C, $\quad \forall j$.

Chapter 10

Elements of Inference and Learning

Statistical Inference describes set of (inference) tasks/operations over the statistical model of the phenomenon. These tasks, assuming knowledge of the statistical model, include: (a) sampling from the probability distribution; (b) computing marginal probabilities; (c) finding the most likely configuration/state. However, the statistical model may or may not be known. In the latter case one needs to learn the model before posing and resolving the challenge of inference.

In the following we will start discussing statistical inference and then shift our attention to the discussion of learning statistical models we aim to infer.

10.1 Statistical Inference: Sampling and Stochastic Algorithms

10.1.1 Monte-Carlo Algorithms: General Concepts and Direct Sampling

This lecture should be read in parallel with the respective IJulia notebook file. Monte-Carlo (MC) methods refers to a broad class of algorithms that rely on repeated random sampling to obtain results. Named after Monte Carlo -the city- which once was the capital of gambling, i.e. playing with randomness. The MC algorithms can be used for numerical integration, e.g. computing weighted sum of many contributions, expectations, marginals, etc. MC can also be used in optimization.

Sampling is a selection of a subset of individuals/configurations from within a statistical population to estimate characteristics of the whole population.

There are two basic flavors of sampling. Direct Sampling MC - mainly discussed in this lecture and Markov Chain MC. DS-MC focuses on drawing **independent** samples from a distribution, while MCMC draws correlated (according to the underlying Markov Chain)

samples.

Let us illustrate both on the simple example of the 'pebble' game - calculating the value of π by sampling interior of a circle.

Direct-Sampling by Rejection vs MCMC for 'pebble game'

In this simple example we will construct distribution which is uniform within a circle from another distribution which is uniform within a square containing the circle. We will use direct product of two rand() to generate samples within the square and then simply reject samples which are not in the interior of the circle.

In respective MCMC we build a sample (parameterized by a pair of coordinates) by taking previous sample and adding some random independent shifts to both variables, also making sure that when the sample crosses a side of the square it reappears on the opposite side. The sample "walks" the square, but to compute area of the circle we count only for samples which are within the circle (rejection again).

See IJulia notebook associated with this lecture for an illustration.

Direct Sampling by Mapping

Direct Sampling by Mapping consists in application of the deterministic function to samples from a distribution you know how to sample from. The method is exact, i.e. it produces independent random samples distributed according to the new distribution. (We will discuss formal criteria for independence in the next lecture.)

For example, suppose we want to generate exponential samples, $y_i \sim p(y) = \exp(-y)$ – one dimensional exponential distribution over $[0, \infty]$, provided that one-dimensional uniform oracle, which generates independent samples, x_i from $[0, 1]$, is available. Then $y_i = -\log(x_i)$ generates desired (exponentially distributed) samples.

Another example of DS MS by mapping is given by the Box-Miller algorithm which is a smart way to map two-dimensional random variable distributed uniformly within a box to the two-dimensional Gaussian (normal) random variable:

$$\int_{-\infty}^{\infty} \frac{dxdy}{2\pi} e^{-(x^2+y^2)/2} = \int_0^{2\pi} \frac{d\varphi}{2\pi} \int_0^{\infty} r dr\, e^{-r^2/2} = \int_0^{2\pi} \frac{d\varphi}{2\pi} \int_0^{\infty} dz\, e^{-z} = \int_0^1 d\theta \int_0^1 d\psi = 1.$$

Thus, the desired mapping is $(\psi, \theta) \to (x, y)$, where $x = \sqrt{-2\log\psi}\cos(2\pi\Theta)$ and $y = \sqrt{-2\log\psi}\sin(2\pi\theta)$.

See IJulia notebook associated with this lecture for numerical illustrations.

Direct Sampling by Rejection (another example)

Let us now show how to get positive Gaussian (normal) random variable from an exponential random variable through rejection. We do it in two steps

- First, one samples from the exponential distribution:

$$x \sim p_0(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

- Second, aiming to get a sample from the positive half of Gaussian,

$$x \sim p_0(x) = \begin{cases} \sqrt{2/\pi}\exp(-x^2/2), & x > 0, \\ 0, & \text{otherwise} \end{cases},$$

  one accepts the generated sample with the probability

$$p(x) = \frac{1}{M}\sqrt{2/\pi}\exp(x - x^2/2),$$

  where M is a constant which should be larger than, $\max(p(x)/p_0(x)) = \sqrt{2/\pi}\, e^{1/2} \approx 1.32$, to guarantee that $p(x) \leq 1$ for all $x > 0$.

Note that the rejection algorithm has an advantage of being applicable even when the probability densities are known only up to a multiplicative constant. (We will discuss issues related to this constant, also called in the multivariate case the partition function, extensively.)

See IJulia notebook associated with this lecture for numerical illustration.

We also recommend

- Introduction to direct Sampling, Chapter of Monte Carlo Lecture Notes by J. Goodman (NYU)
- Lecture on Monte Carlo Sampling, from Berkley course of M. Jordan on Bayesian Modeling and Inference

for additional reading on DS-MC.

Importance Sampling

One important application of MC is in computing sums, integrals and expectations. Suppose we want to compute an expectation of a function, $f(x)$, over the distribution, $p(x)$, i.e. $\int dx p(x) f(x)$, in the regime where $f(x)$ and $p(x)$ are concentrated around very different x.

In this case the overlap of $f(x)$ and $p(x)$ is small and as a result a lot of MC samples drawn from $p(x)$ will be 'wasted'.

Importance Sampling is the method which aims to fix the small-overlap problem. The method is based on adjusting the distribution function from $p(x)$ to $\tilde{p}(x)$ and then utilizing the following obvious formula

$$\mathbb{E}_p[f(x)] = \int dx p(x) f(x) = \int dx \tilde{p}(x) \frac{f(x)p(x)}{\tilde{p}(x)} = \mathbb{E}_{\tilde{p}} \left[\frac{f(x)p(x)}{\tilde{p}(x)} \right]$$

See the IJulia notebook associated with this lecture contrasting DS example, $p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ and $f(x) = \exp\left(-\frac{(x-4)^2}{2}\right)$, with IS where the choice of the proposal distribution is, $\tilde{p}(x) = \frac{1}{\sqrt{\pi}} \exp\left(-(x-2)^2\right)$. This example shows that we are clearly wasting samples with DS.

Note one big problem with IS. In a realistic multi-dimensional case it is not easy to guess the proposal distribution, $\tilde{p}(x)$, right. One way of fixing this problem is to search for good $\tilde{p}(x)$ adaptively.

A comprehensive review of the history and state of the art in Importance Sampling can be found in multiple lecture notes of A. Owen posted at his web page, for example follow this link. Check also adaptive importance sampling package.

Direct Brute-force Sampling

This algorithm relies on availability of the uniform sampling algorithm from $[0, 1]$, rand(). One splits the $[0, 1]$ interval into pieces according to the weights of all possible states and then use rand() to select the state. The algorithm is impractical as it requires keeping in the memory information about all possible configurations. The use of this construction is in providing the bench-mark case useful for proving independence of samples.

Direct Sampling from a multi-variate distribution with a partition function oracle

Suppose we have an oracle capable of computing the partition function (normalization) for a multivariate probability distribution and also for any of the marginal probabilities. (Notice that we are ignoring for now the issue of the oracle complexity.) Does it give us the power to generate independent samples?

We get affirmative answer to this question through the following **decimation** algorithm generating independent sample $x \sim P(x)$, where $x := (x_i | i = 1, \cdots, N)$:

Validity of the algorithm follows from the exact representation for the joint probability distribution function as a product of ordered conditional distribution function (chain rule

Algorithm 6 Decimation Algorithm

Input: $P(x)$ (expression). Partition function oracle.

1: $x^{(d)} = \varnothing; \quad I = \varnothing$
2: **while** $|I| < N$ **do**
3: Pick i at random from $\{1, \cdots, N\} \setminus I$.
4: $x^{(I)} = (x_j | j \in I)$
5: Compute $P(x_i|x^{(d)}) := \sum_{x \setminus x_i; x^{(I)} = x^{(d)}} P(x)$ with the oracle.
6: Generate random $x_i \sim P(x_i|x^{(d)})$.
7: $I \cup i \leftarrow I$
8: $x^{(d)} \cup x_i \leftarrow x^{(d)}$
9: **end while**

Output: $x^{(\mathrm{dec})}$ is an independent sample from $P(x)$.

for distribution):

$$P(x_1, \cdots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \cdots P(x_n|x_1, \cdots, x_{n-1}). \tag{10.1}$$

(The chain rule follows directly from Bayes rule/formula. Notice also that ordering of variables within the chain rule is arbitrary.) One way of proving that the algorithm produces an independent sample is to show that the algorithm outcome is equivalent to another algorithm for which the independence is already proven. The benchmark algorithm we can use to state that the Decimation algorithm (6) produces independent samples is the brute-force sampling algorithm described in the beginning of the lecture. The crucial point here is that the decimation algorithm can be interpreted in terms of splitting the $[0, 1]$ interval hierarchically, first according to $P(x_1)$, then subdividing pieces for different x_1 according to $P(x_2, x_1)$, etc. This guidanken experiment will result in the desired proof.

Note that in general efforts of the partition function oracle are exponential in the problem size. However in some special cases the partition function can be computed efficiently (polynomially in the number of steps).

In the following exercise we suggest to test performance of the direct sampling algorithm on the example of the Ising model. We remind that in the case of the Ising model probability of a binary vector (spin configuration), x, is given by

$$p(x) = \frac{\exp\left(-\beta E(x)\right)}{Z}, \quad E(x) = -\frac{1}{2} \sum_{\{i,j\} \in \mathcal{E}} x_i J_{ij} x_j + \sum_{i \in \mathcal{V}} h_i x_i, \tag{10.2}$$

$$Z = \sum_x \exp\left(-\beta E(x)\right). \tag{10.3}$$

Exercise 10.1. Consider example of the Ising model with zero singleton term, $h = 0$, and uniform pair-wise term, $J\beta = -1$, i.e. $J_{ij}\beta = -1$, $\forall\{i, j\} \in \mathcal{E}$, over $n \times n$ grid-graph with the nearest-neighbor interaction. Construct (write down pseudo-algorithm and then code) the decimation algorithm (6).
Compare the performance of the direct sampling for $n = 2, 3, 4, 5$ – that is find out how the time required to generate the next i.i.d. sample depends on n – and explain.

10.1.2 Inference via Markov-Chain Monte-Carlo

Markov Chain Monte Carlo (MCMC) methods belong to the class of algorithms for sampling from a probability distribution which is based on constructing a Markov chain that converges to the target steady distribution.

Examples and flavors of MCMC are many (and some are quite similar) – heat bath, Glauber dynamics, Gibbs sampling, Metropolis-Hastings, Cluster algorithm, Warm algorithm, etc – in all these cases we only need to know the transition probability between states while the actual stationary distribution may be not known or, more accurately, known up to the normalization factor, also called the partition function. Below, we will discuss in details two key examples: Gibbs sampling and Metropolis-Hastings.

Gibbs Sampling

Assume that the direct sampling is not feasible, because of the high dimensionality of the state vector (too many components): computation of the joint probability distribution and its marginalizations (over a small number of components) is of the "exponential" complexity (more on this below). The main point of the Gibbs sampling is that given a multivariate distribution is that, even though to sample from the joint probability distribution is not feasible, sampling from the probability distribution of a few components (conditioned on the rest of the state vector) can be done efficiently. We utilize this remarkable feature of the marginal probability distributions and create the Gibbs Sampling algorithm (Algorithm 7). The algorithm starts from a current sample of the vector x, pick a component at random, compute probability for this component/variable conditioned to the rest (other components of the state vector), and sample from this conditional distribution. As mentioned above, the conditional distribution is over a single component, and it is therefore an easy one to compute. We continue the process till convergence, which can be identified empirically, for example, by checking if estimation of the histogram or observable(s) stopped changing.

Example 10.1.1. Describe Gibbs sampling on the example of a general Ising model. Build respective Markov chain. Show that the algorithm obeys Detailed Balance.

Algorithm 7 Gibbs Sampling

Input: Given $p(x_i|x_{\sim i} = x \setminus x_i), \quad \forall i \in \{1, \cdots, N\}$. Start with a sample $x^{(t)}$.

loop Till convergence
  Draw an i.i.d. i from $\{1, \cdots, N\}$.
  Generate a random $x_i \sim p\left(x_i|x^{(t)}_{\sim i}\right)$.
  $x_i^{(t+1)} = x_i$.
  $\forall j \in \{1, \cdots, N\} \setminus i : \quad x_j^{(t+1)} \leftarrow x_j^{(t)}$.
  Output $x^{(t+1)}$ as the next sample.
end loop

Solution. Starting from a state, $x^{(t)}$, we pick a random node, i, and compare two candidate states, ($x_i = 1$ and $x_i = -1$). Then we calculate the corresponding conditional (all spins except i are fixed) probabilities $p(x_i = +1|x^{(t)}_{\sim i}$ and $p(x_i = -|x^{(t)}_{\sim i}$, which we denote, $p_\pm$, respectively. By construction,

$$p_+ + p_- = 1, \quad p_+/p_- = e^{-\beta\Delta E}, \quad \Delta E = E(x_i = +1, x^{(t)}_{\sim i}) - E(x_i = -1, x^{(t)}_{\sim i}), \qquad (10.4)$$

where ΔE, is the energy difference between the two configurations. Next, one accepts the configuration $x_i = 1$ with the probability p_+ or the configuration $x_i = -1$ with the probability p_-.

MC corresponding to the algorithm is defined over the 2^N-dimensional hupercube, when N is the number of spins, and 2^N is the number of states. To check the DB condition, assume that the conditional probabilities, $p_\pm$, corresponds to the steady state and compute the probability fluxes from the state $x^{(t)}$ to the states $(x_i = \pm 1, x^{(t)}_{\sim i})$. We derive

$$Q_{-+} = \frac{1}{Z} e^{-\beta E(x_i=-1, x^{(t)}_{\sim i})} p_+, \quad Q_{+-} = \frac{1}{Z} e^{-\beta E(x_i=+1, x^{(t)}_{\sim i})} p_-. \qquad (10.5)$$

Combining Eqs. (10.4) and Eqs. (10.5), we observe that, $Q_{-+} = Q_{+-}$, i.e. the detailed balance holds. □

Metropolis-Hastings Sampling

Metropolis-Hastings (MH) sampling is an MCMC method which is built, like Gibbs sampling, assuming that the desired stationary distribution, $\tilde{\pi}(x)$, is known explicitly upto the normalization constant (also called the partition function), where thus the normalized distribution is $\pi(x) = \tilde{\pi}(x)/Z$ and $Z := \sum_x \tilde{\pi}(x)$. Let us also introduce the so-called "proposal" distribution, $p(x'|x)$, and assume that drawing a sample proposal x' from the

current sample x is (computationally) easy. The combination of $\tilde{\pi}(x)$ and $p(x'|x)$, as well as an arbitrary initialization of the state vector, $x^{(t)}$, set the MH Algorithm 8. Starting with $x^{(t)}$ the algorithm draws a sample, x', according to the proposal distribution, $p(x'|x^{(t)})$, but then accepts or reject the proposed state, x', according to the probability

$$\min\Big\{1, \frac{p(x_t|x')\tilde{\pi}(x')}{p(x'|x^{(t)})\tilde{\pi}(x^{(t)})}\Big\}. \tag{10.6}$$

The procedure is repeated till (empirical) convergence.

Observe that by construction

$$\forall x, x': \quad \underbrace{\min\Big\{1, \frac{p(x|x')\tilde{\pi}(x')}{p(x'|x)\tilde{\pi}(x)}\Big\}p(x'|x)}_{\text{MH trans. prob.}(x'\leftarrow x)}\,\tilde{\pi}(x) = \underbrace{\min\Big\{1, \frac{p(x'|x)\tilde{\pi}(x)}{p(x|x')\tilde{\pi}(x')}\Big\}p(x|x')}_{\text{MH trans. prob.}(x\leftarrow x')}\,\tilde{\pi}(x'), \tag{10.7}$$

i.e. the algorithm satisfies the Detailed Balance condition.

Algorithm 8 Metropolis-Hastings Sampling

Input: Given $\tilde{\pi}(x)$ and $p(x'|x)$. Start with a sample x_t.

loop Till convergence
  Draw a random $x' \sim p(x'|x^{(t)})$.
  Compute $\alpha = \frac{p(x_t|x')\tilde{\pi}(x')}{p(x'|x^{(t)})\tilde{\pi}(x^{(t)})}$.
  Draw random $\beta \in U([0,1])$, uniform i.i.d. from $[0,1]$.
  if $\beta < \min\{1, \alpha\}$ **then**
    $x^{(t)} \leftarrow x'$ [accept]
  else
    x' is ignored [reject]
  end if
  $x^{(t)}$ is recorded as a new sample
end loop

Note that the Gibbs sampling, introduced before, can be considered as the Metropolis-Hastings without rejection, where the proposal distribution is chosen specifically as the respective conditional probability distribution. (That is Gibbs sampling should be considered as a special case of the more general MH algorithm.)

Example 10.1.2. Consider MC shown in Fig. (10.1). Show that this MC is ergodic and can be viewed as a particular MH Algorithm 8 for the two spin Ising model. What is the proposal distribution in this case? What is the resulting stationary distribution? Does it

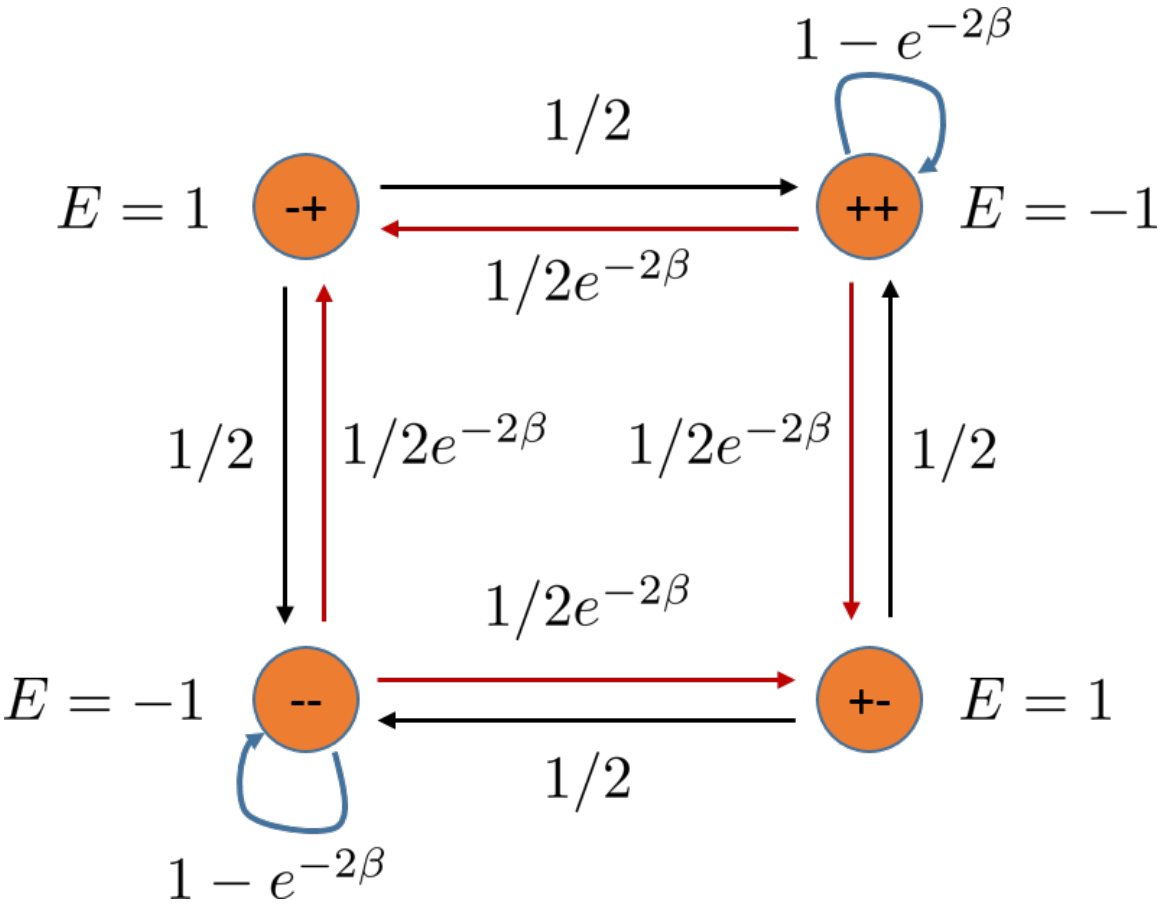


Figure 10.1: Example of the MC induced by a Metropolis-Hastings for a two spin example.

obey Detailed Balance? Will the steady distribution change if the rejection is removed from the consideration? How?

Solution. The cardinality (size) of the state space is $2^2 = 4$: $x = (x_1 = \pm 1, x_2 = \pm 1)$. Inspecting the MC we observe that it is a-periodic and positive-recurrent, thus ergodic. Observe that some of the transition probabilities are exactly $1/2$. Therefore, it is natural to associate these with the cases when Eq. (10.6) returns unity. Next, we link self-loops in the MC to rejections, that are the cases when a proposed state is rejected. The two observations combined suggest that we are dealing with an instance of the MH Algorithm 8 where

$$\tilde{\pi}(x_1, x_2) = \exp(-\beta E(x)),\ E(x = (x_1, x_2)) = -\frac{x_1 x_2}{2},\ p(x'|x) = \exp\left(-\beta\left(E(x') - E(x)\right)\right).$$

That is the stationary distribution is of the attractive (ferromagnetic) type with the unit pair-wise strength and without bias (with zero magnetic field). The proposal distribution, $p(x'|x)$, is such that only a single spin flip is allowed (cannot flip two spins in one step) and

the proposal may be accepted only if the resulting energy gain is positive, i.e. when the algorithm proposes to move from the state when the spins are miss-aligned to the aligned state. Removal of the rejection translates into setting β to zero (temperature becomes infinite). In this case the resulting distribution is uniform (so-called, para-magnetic). □

The freedom in selecting MH proposal distribution translates into strong variations in the resulting algorithm mixing time. Typically we would like to find proposal which mixes fast. Even though evaluating mixing time of the proposal is a challenge, one can estimate it empirically following the following heuristics. If the largest distance between the states (measured in the number of the algorithm's elementary steps) is L, the MH algorithm executes a random walk, i.e. it advances diffusively covering distance L in $T \sim L^2$ steps. This will then be a low bound estimate on the mixing time. Notice that the actual mixing time, i.e. the time to arrive at a sample which is almost independent on the initial sample, may be significantly slower, e.g. due to rejection (if these are frequent). This slow (diffisive) exploration of the phase space by the MH algorithm is linked to DB.

The particular form of the MH proposal, illustrated in the Example 10.1.2 for the two spin Ising model, generalizes to the so-called Glauber (dynamics) Algorithm 9. (Check snippet illustrating performance of the Glauber algorithm on a 128×128 square lattice.)

Algorithm 9 Glauber Sampling

Input: Ising model on a graph. (See Eq. (10.2).) Start with a sample x

loop Till convergence
  Pick a node i at random.
  $-x_i \leftarrow x_i$
  Compute $\alpha = \exp\left(x_i \left(\sum_{j\in\mathcal{V}:\{i,j\}\in\mathcal{E}} J_{ij}x_j - 2h_i\right)\right)$.
  Draw random $\beta \in U([0,1])$, uniform i.i.d. from $[0,1]$.
  if $\alpha < \beta < 1$ **then**
    $-x_i \leftarrow x_i$ [reject]
  end if
  Output: x as a sample
end loop

Exercise 10.2 (Spanning Trees.)**.** Let G be an undirected complete graph. The following MCMC algorithm results in a uniform stationary probability distribution of all the spanning trees of G: Start with some spanning tree; add uniformly-at-random some edge from G (so that a cycle forms); remove uniformly-at-random an edge from this cycle; repeat. Suppose

now that the graph G is positively weighted, i.e., each edge e has some cost $c_e > 0$. Suggest an MCMC algorithm that samples from the set of spanning trees of G, with the stationary probability distribution proportional to the overall weight of the spanning tree for the following cases:
(i) the weight of any spanning tree of G is the sum of costs of its edges;
(ii) the weight of any spanning tree of G is the product of costs of its edges. In addition,
(iii) estimate the average weight of a spanning tree using the algorithm of uniform sampling.
(iv) implement all the algorithms on a (4×4) square lattice with randomly assigned weights. Verify that the algorithm converges to the right value.

For useful additional reading on sampling and computations for the Ising model see `https://www.physik.uni-leipzig.de/~janke/Paper/lnp739_079_2008.pdf`.

Exactness and Convergence

MCMC algorithm is called (casually) exact if one can show that the generated distribution "converges" to the desired stationary distribution. However, "convergence" may mean different things.

The strongest form of convergence – called **exact independence test** (warning - this is our 'custom' term) – states that at each step we generate an independent sample from the target distribution. To prove this statement means to show that empirical correlation of the consecutive samples is zero in the limit when Nnumber of samples $\to \infty$:

$$\lim_{N \to +\infty} \frac{1}{N} \sum_{n=0}^{N} f(x_n) g(x_{n-1}) \to \mathbb{E}\left[f(x)\right] \mathbb{E}\left[g(x)\right], \tag{10.8}$$

where $f(x)$ and $g(x)$ are arbitrary functions (however such that respective expectations on the rhs of Eq. (10.8) are well-defined).

A weaker statement – call it **asymptotic convergence** – suggests that in the limit of $N \to \infty$ we reconstruct the target distribution (and all the respective existing moments):

$$\lim_{N \to +\infty} \frac{1}{N} \sum_{n=0}^{N} f(x_n) \to \mathbb{E}\left[f(x)\right], \tag{10.9}$$

where $f(x)$ is an arbitrary function such that the expectation on the rhs is well defined.

Finally, the weakest statement – call it **parametric convergence** – corresponds to the case when one arrives at the target estimate only in a special limit with respect to a special parameter. It is common, e.g. in statistical/theoretical physics and computer sicence, to study the so-called thermodynamic limit, where the number of degrees of freedom (for

example number of spins/variables in the Ising model) becomes infinite:

$$\lim_{s\to s_*}\lim_{N\to+\infty}\frac{1}{N}\sum_{n=0}^{N} f_s(x_n) \to \mathbb{E}\left[f_{s_*}(x)\right]. \tag{10.10}$$

For additional math (but also intuitive as written for applied mathematicians, engineers and physicists) reading on the MCMC (and in general MC) convergence see "The mathematics of mixing things up" article by Persi Diaconis and also [16].

Exact Monte Carlo Sampling (Did it converge yet?)

(This part of the lecture is a bonus material - we discuss it only if time permits.)

The material follows Chapter 32 of D.J.C. MacKay book [19]. An extensive set of modern references, discussions and codes are also available at the website [23] on perfectly random sampling with Markov Chains.

As mentioned already the main problem with MCMC methods is that one needs to wait (and sometimes for too long) to make sure that the generated samples (from the target distribution) are i.i.d. If one starts to form a histogram (empirical distribution) too early it will deviate from the target distribution. One important question in this regards is: For how long shall one run the Markov Chain before it has 'converged'? To answer this question (prove) it is very difficult, in many cases not possible. However, there is a technique which allows to check the **exact convergence**, for some cases, and do it on the fly - as we run MCMC.

This smart technique is the Propp-Wilson exact sampling method, also called **coupling from the past**. The technique is based on a combination of three ideas:

- The main idea is related to the notion of the **trajectory coalescence**. Let us observe that if starting from different initial conditions the MCMC chains share a single random number generator, then their trajectories in the phase space can coalesce; and having coalesced, will not separate again. This is clearly an indication that the initial conditions are forgotten.

  Will running all the initial conditions forward in time till coalescence generate exact sample? Apparently not. One can show (sufficient to do it for a simple example) that the point of coalescence does not represent an exact sample.

- However, one can still achieve the goal by **sampling from a time T_0 in the past**, up to the present. If the coalescence has occurred the present sample is an unbiased sample; and if not we restart the simulation from the time T_0 further into the past, reusing the same random numbers. The simulation is repeated till a coalescence occur

at a time before the present. One can show that the resulting sample at the present is exact.

- One problem with the scheme is that we need to test it for all the initial conditions - which are too many to track. Is there a way to **reduce the number of necessary trials**. Remarkably, it appears possible for sub-class of probabilistic models the so-called **'attractive'** models. Loosely speaking and using 'physics' jargon - these are **'ferromagnetic'** models - which are the models where for a stand alone pair of variables the preferred configuration is the one with the same values of the two variables. In the case of attractive model monotonicity (sub-modularity) of the underlying model suggests that the paths do not cross. This allows to only study limiting trajectories and deduce interesting properties of all the other trajectories from the limiting cases.

10.2 Statistical Inference: General Relations, Calculus of Variations and Trees

This lecture largely follow material of the mini-course on *Graphical Models of Statistical Inference: Belief Propagation & Beyond*. See links to slides and lecture notes at the following web-site.

10.2.1 From Ising Model to (Factor) Graphical Models

Brief reminder of what we have learned so far about the Ising Model. It is fully described by Eqs. (10.2,10.3). The weight of a "spin" configuration is given by Eq. (10.2). Let us not pay much of attention for now to the normalization factor Z and observe that the weight is nicely factorized. Indeed, it is a product of pair-wise terms. Each term describes "interaction" between spins. Obviously we can represent the factorization through a graph. For example, if our spin system consists only of three spins connected to each other, then the respective graph is a triangle. Spins are associated with nodes of the graphs and "interactions", which may also be called (pair-wise) factors, are associated with edges.

It is useful, for resolving this and other factorized problems, to introduce a bit more general representation — in terms of graphs where both factors and variables are associated with nodes/vertices. Transformation to the factor-graph representation for the three spin example is shown in Fig. (10.2).

Ising Model, as well as other models discussed later in the lectures, can thus be stated

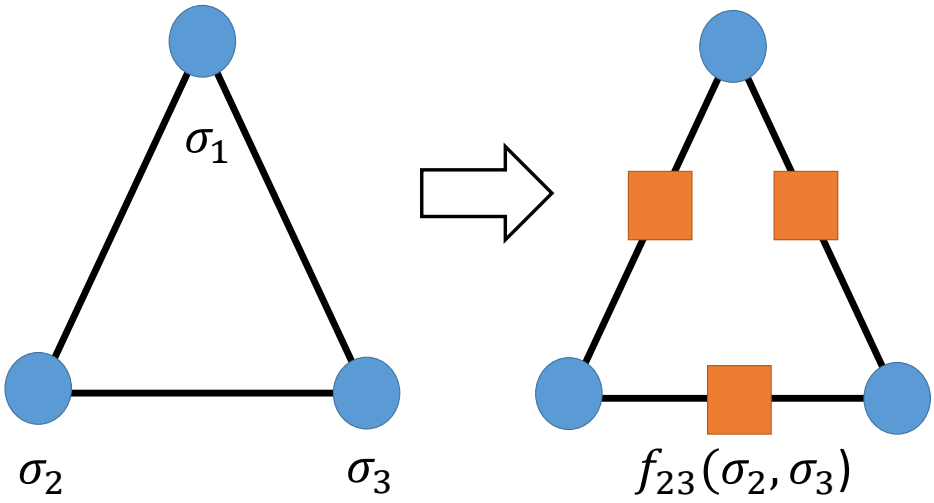


Figure 10.2: Factor Graph Representation for the (simple case) with pair-wise factors only. In the case of the Ising model: $f_{12}(x_1, x_2) = \exp(-J_{12}x_1x_2 + h_1x_1 + h_2x_2))$.

in terms of the general factor-graph framework/model

$$P(x) = Z^{-1} \prod_{a \in \mathcal{V}_f} f_a(x_a), \quad x_a := (x_i | i \in \mathcal{V}_n, \quad (i,a) \in \mathcal{E}), \tag{10.11}$$

where $(\mathcal{V}_f, \mathcal{V}_n, \mathcal{E})$ is the bi-partite graph built of factors and nodes.

The factor graph language (representation) is more general. We will see it next - discussing another problem which originates from the Information Theory, and specifically from the error-correction theory.

10.2.2 Decoding of Graphical Codes as a Factor Graph problem

First, let us discuss decoding of a graphical code. (Our description here is terse, and we advise interested reader to check the book by Richardson and Urbanke [24] for more details.) A message word consisting of L information bits is encoded in an N-bit long code word, $N > L$. In the case of binary, linear coding discussed here, a convenient representation of the code is given by $M \geq N - L$ constraints, often called parity checks or simply, checks. Formally, $\boldsymbol{\varsigma} = (\varsigma_i = 0, 1 | i = 1, \cdots, N)$ is one of the 2^L code words iff $\sum_{i \sim \alpha} \varsigma_i = 0 \ (\mod 2)$ for all checks $\alpha = 1, \cdots, M$, where $i \sim \alpha$ if the bit if the bit i contributes the check α, and $\alpha \sim i$ will indicate that the check α contains bit i. The relation between bits and checks is often described in terms of the $M \times N$ parity-check matrix $\boldsymbol{H}$ consisting of ones and

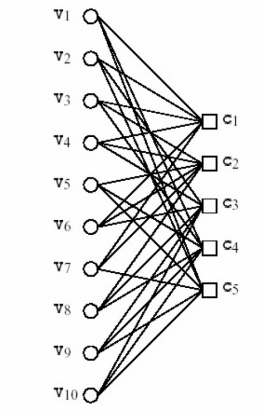


Figure 10.3: Tanner graph of a linear code, represented with $N = 10$ bits, $M = 5$ checks, and $L = N - M = 5$ information bits. This code selects 2^5 codewords from 2^10 possible patterns. This adjacency, parity-check matrix of the code is given by Eq. (10.12).

zeros: $H_{i\alpha} = 1$ if $i \sim \alpha$ and $H_{i\alpha} = 0$ otherwise. The set of the codewords is thus defined as $\Xi^{(cw)} = (\varsigma | \boldsymbol{H}\boldsymbol{\varsigma} = \mathbf{0} \ (\mathrm{mod}\ 2))$. A bipartite graph representation of $\boldsymbol{H}$, with bits marked as circles, checks marked as squares, and edges corresponding to respective nonzero elements of $\boldsymbol{H}$, is usually called (in the coding theory) the Tanner graph of the code, or parity-check graph of the code. (Notice that, fundamentally, code is defined in terms of the set of its codewords, and there are many parity check matrixes/graphs parameterizing the code. We ignore this unambiguity here, choosing one convenient parametrization $\boldsymbol{H}$ for the code.) Therefore the bi-partite Tanner graph of the code is defined as $\mathcal{G} = (\mathcal{G}_0, \mathcal{G}_1)$, where the set of nodes is the union of the sets associated with variables and checks, $\mathcal{G}_0 = \mathcal{G}_{0;v} \cup \mathcal{G}_{0;e}$ and only edges connecting variables and checks contribute $\mathcal{G}_1$.

For a simple example with 10 bits and 5 checks, the parity check (adjacency) matrix of the code with the Tanner graph shown in Fig. (10.3) is

$$\boldsymbol{H} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}. \tag{10.12}$$

Another example of a bigger code and respective parity check matrix are shown in Fig. (10.4). For this example, $N = 155$, $L = 64$, $M = 91$ and the Hamming distance, defined as the minimum l_0-distance between two distinct codewords, is 20.

Assume that each bit of the transmitted signal is changed (effect of the channel noise)

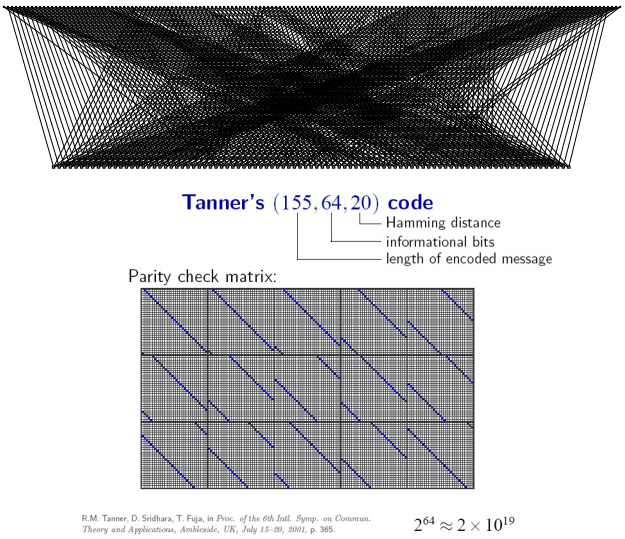


Figure 10.4: Tanner graph and parity check matrix of the $(155, 64, 20)$ Tanner code, where $N = 155$ is the length of the code (size of the code word), $L = 64$ and the Hamming distance of the code, $d = 20$.

independently of others. It is done with some known conditional probability, $p(x|\sigma)$, where $\sigma = 0, 1$ is the valued of the bit before transmission, and $x \in \mathbb{R}$ is its changed/distorted image. Once $\boldsymbol{x} = (x_i | i = 1, \cdots, N)$ is measured, the task of the Maximum-A-Posteriori (MAP) decoding becomes to reconstruct the most probable codeword consistent with the measurement:

$$\boldsymbol{\sigma}^{(MAP)} = \arg \min_{\boldsymbol{\sigma} \in \Xi^{(cw)}} \prod_{i=1}^{N} p(x_i|\sigma_i). \tag{10.13}$$

More generally, the probability of a codeword $\boldsymbol{\varsigma} \in \Xi^{(cw)}$ to be a pre-image for $\boldsymbol{x}$ is

$$\mathcal{P}(\boldsymbol{\varsigma}|\boldsymbol{x}) = (Z(\boldsymbol{x}))^{-1} \prod_{i \in \mathcal{G}_{0;v}} g^{(ch)}(x_i|\varsigma_i), \quad Z(\boldsymbol{x}) = \sum_{\boldsymbol{\varsigma} \in \Xi^{(cw)}} \prod_{i \in \mathcal{G}_{0;v}} g^{(ch)}(x_i|\varsigma_i), \tag{10.14}$$

where $Z(\boldsymbol{x})$ is thus the partition function dependent on the detected vector $\boldsymbol{x}$. One may also consider the signal (bit-wise) MAP decoder

$$\forall\ i: \quad \varsigma_i^{(s-MAP)} = \arg \max_{\varsigma_i} \sum_{\boldsymbol{\varsigma} \setminus \varsigma_i}^{\boldsymbol{\varsigma} \in \Xi^{(cw)}} \mathcal{P}(\boldsymbol{\varsigma}|\boldsymbol{x}). \tag{10.15}$$

10.2.3 Partition Function. Marginal Probabilities. Maximum Likelihood.

The partition function in Eq. (10.11) is the normalization factor

$$Z = \sum_{x} \prod_{a \in \mathcal{V}_f} f_a(x_a), \quad x_a := (x_i | i \in \mathcal{V}_n, \quad (i,a) \in \mathcal{E}), \tag{10.16}$$

where $x = (x_i \in \{0,1\} \in \mathcal{V}_n)$. Here, we assume that the alphabet of the elementary random variable is binary, however generalization to the case of a higher alphabet is straightforward.

We are interested to 'marginalize' Eq. (10.11) over a subset of variables, for example over all the elementary/nodal variables but one

$$P(x_i) := \sum_{x \setminus x_i} P(x). \tag{10.17}$$

Expectation of x_i computed with the probability Eq. (10.17) is also called (in physics) 'magnetization' of the variable.

Example 10.2.1. Is a partition function oracle sufficient for computing $P(x_i)$? What is the relation in the case of the Ising model between $P(x_i)$ and $Z(h)$?

Solution. There are two ways one can relate $P(x_i)$ to the partition function. First, introduce an auxiliary graphical model, derived from the original one simply by fixing the value at the node i to x_i. Then, $P(x_i)$ is simply the ratio of the partition function of the newly derived graphical model to the original graphical model. Second, we can modify the original graphical model introducing multiplicative factor, $\exp(x_i h_i)$, and denoting the resulting partition function by $Z(h)$. Then, $\log Z(h)$, is also a moment generating function of $P(x_i)$.

Another object of interest is the so-called Maximum Likelihood. Stated formally, is the most probable state of all the states represented in Eq. (10.11):

$$x_* = \arg\max_x P(x). \tag{10.18}$$

All these objects are difficult to compute. "Difficulty" - still stated casually - means that the number of operations needed is exponential in the system size (e.g. number of variables/spins in the Ising model). This is in general, i.e. for a GM of a general position. However, for some special cases, or even special classes of cases, the computations may be much easier than in the worst case. Thus, ML (10.18) for the case of the so-called ferromagnetic (attractive, sub-modular) Ising model can be computed with efforts polynomial in the system size. Note that the partition function computation (at any nonzero temperatures) is still exponential even in this case, thus illustrating the general statement - computing Z or $P(x_i)$ is a more difficult problem than computing x_*.

A curious fact. Ising model (ferromagnetic, anti-ferromagnetic or glassy) when the "magnetic field" is zero, $h = 0$, and the graph is planar, represents a very unique class of problems for which even computations of Z are easy. In this case the partition function is expressed via determinant of a finite matrix, while computing determinant of a size N matrix is a problem of $O(N^3)$ complexity (actually $O(N^{3/2})$ in the planar case).

In the general (difficult) case we will need to relay on approximations to make computations scalable. And some of these approximations will be discussed later in the lecture. However, let us first prepare for that - restating the most general problem discussed so far – computation of the partition function, Z – as an optimization problem.

10.2.4 Kullback-Leibler Formulation & Probability Polytope

We will go from counting (computing partition function is the problem of weighted counting) to optimization by changing description from states to probabilities of the states, which we will also call beliefs. $b(x)$ will be a belief - which is our probabilistic guess - for the probability of state x. Consider it on the example of the triangle system shown in Fig. (10.2). There are 2^3 states in this case: $(x_1 = \pm 1, x_2 = \pm 1, x_3 = \pm 1)$, which can occur with the probabilities, $b(x_1, x_2, x_3)$. All the beliefs are positive and together should some to unity. We would like to compare a particular assignment of the beliefs with $P(x)$, generally described by Eq. (10.11). Let us recall a tool which we already used to compare probabilities - the Kullback-Leibler (KL) divergence (of probabilities) discussed in Lecture #2:

$$D(b\|P) = \sum_x b(x) \log\left(\frac{b(x)}{P(x)}\right) \tag{10.19}$$

Note that the KL divergence (10.19) is a convex function of the beliefs (remember, there are 2^3 of the beliefs in the our enabling three node example) within the following polytope – domain in the space of beliefs bounded by linear constraints:

$$\forall x: \quad b(x) \geq 0, \tag{10.20}$$

$$\sum_x b(x) = 1. \tag{10.21}$$

Moreover, it is straightforward to check (please do it at home!) that the unique minimum of $D(b\|P)$ is achieved at $b = P$, where the KL divergence is zero:

$$P = \arg\min_b D(b\|P), \quad \min_b D(b\|P) = 0. \tag{10.22}$$

Substituting Eq. (10.11) into Eq. (10.22) one derives

$$\log Z = -\min_b \mathcal{F}(b), \quad \mathcal{F}(b) := \sum_x b(x) \log\left(\frac{\prod_a f_a(x_a)}{b(x)}\right), \tag{10.23}$$

where $F(b)$, considered as a function of all the beliefs, is called (configurational) free energy (where configuration is one of the beliefs). The terminology originates from statistical physics.

To summarize, we did manage to reduce counting problem to an optimization problem. Which is great, however so far it is just a reformulation – as the number of variational degrees of freedom (beliefs) is as much as the number of terms in the original sum (the partition function). Indeed, it is not the formula itself but (as we will see below) its further use for approximations which will be extremely useful.

10.2.5 Variational Approximation: Mean Field

The main idea is to reduce the search space from exploration of the 2^N-1 dimensional beliefs to their lower dimensional, i.e. parameterized with fewer variables, proxy/approximation. What kind of factorization can one suggest for the multivariate (N-spin) probabilities/beliefs? The idea of postulating independence of all the N variables/spins comes to mind:

$$b(x) \to b_{MF}(x) = \prod_i b_i(x_i) \tag{10.24}$$

$$\forall i \in \mathcal{V}_i, \quad \forall x_i : \quad b_i(x_i) \geq 0 \tag{10.25}$$

$$\forall i \in \mathcal{V}_i : \quad \sum_{x_i} b_i(x_i) = 1. \tag{10.26}$$

Clearly $b_i(x_i)$ is interpreted within this substitution as the single-node marginal belief (estimate for the single-node marginal probability).

Substituting b by b_{MF} in Eq. (10.23) one arrives at the MF estimation for the partition function

$$\log Z_{mf} = -\min_{b_{mf}} \mathcal{F}(b_{mf}),$$

$$\mathcal{F}(b_{mf}) := \sum_a \sum_{x_a} \left(\prod_{i \sim a} b_i(x_i) \right) \log f_a(x_a) - \sum_i \sum_{x_i} b_i(x_i) \log(b_i(x_i)). \tag{10.27}$$

To solve the variational problem (10.27) constrained by Eqs. (10.24,10.25,10.26) is equivalent to searching for the (unique) stationary point of the following MF Lagrangian

$$\mathcal{L}(b_{mf}) := \mathcal{F}(b_{mf}) + \sum_i \lambda_i \sum_{x_i} b_i(x_i) \tag{10.28}$$

Example 10.2.2. Show that $Z \geq Z_{mf}$, and that $\mathcal{F}(b_{mf})$ is a strictly convex function of its (vector) argument. Write down equations defining the stationary point of $\mathcal{L}(b_{mf})$.

Solution. Z_{mf} is low bound because we optimize over the class of belief functions, $b_{mf}(x)$ which is strictly within the class of allowed $b(x)$. Convexity is proven simply by utilizing the fact the optimization's objective is decomposed into a sum of the convex functions.

The fact that Z_{mf} (see the example above) gives a lower bound on Z is a good news. However, in general the approximation is very crude, i.e. the gap between the bound and the actual value is large. The main reason for that is clear - by assuming that the variables are independent we have ignored significant correlations.

In the next lecture we will analyze what, very frequently, provides a much better approximation for ML inference - the so called Belief Propagation approach.

We will mainly focus on the so-called Belief Propagation, related theory and techniques. In addition to discussing inference with Belief Propagation we will also have a brief discussions (pointers) to respective inverse problem – learning with Graphical Models.

10.2.6 Dynamic Programming for (Exact) Inference over Trees

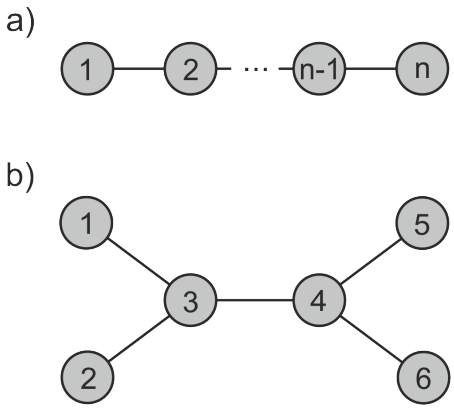


Figure 10.5: Exemplary interaction/factor graphs which are tree.

Consider Ising model over a linear chain of n spins shown in Fig. 10.5a, the partition function is

$$Z = \sum_{x_n} Z(x_n), \tag{10.29}$$

where $Z(x_n)$ is the newly introduced object representing sum over all but last spin in the chain, labeled by n. Z_n can be expressed as follows

$$Z(x_n) = \sum_{x_{n-1}} \exp(J_{n,n-1}x_n x_{n-1} + h_n x_n) Z_{(n-1)\to(n)}(x_{n-1}), \tag{10.30}$$

where $Z_{(n-1)\to(n)}(x_i)$ is the partial partition function for the subtree (a shorter chain in this case) rooted at $n-1$ and built excluding the branch/link directed towards n. The newly

introduced partially summed partition function contains summation over one less spins then the original chain. In fact, this partially sum object can be defined recursively

$$Z_{(i-1)\to(i)}(x_{i-1}) = \sum_{x_{i-2}} \exp(J_{i-1,i-2}x_{i-1}x_{i-2} + h_{i-1}x_{i-1}) Z_{(i-2)\to(i-1)}(x_{i-2}) \tag{10.31}$$

that is expressing one partially sum object via the partially sum object computed on the previous step. Advantage of this recursive approach is obvious – it allows to replace summation over the exponentially many spin configurations by summing up of only two terms at each step of the recursion.

What should also be obvious is that the method just described is adaptation of the Dynamic Programming (DP) methods we have discussed in the optimization part of the course to the problem of statistical inference.

It is also clear that the approach just explained allows generalization from the case of the linear chain to the case of a general tree. Then, in the general case $Z(x_i)$ is the partition function of the entire tree with a value of the spin at the site/node i fixed. We derive

$$Z(x_i) = e^{h_i x_i} \prod_{j\in\partial i} \left(\sum_{x_j} e^{J_{ij}x_i x_j} Z_{j\to i}(x_j) \right), \tag{10.32}$$

where ∂i denotes the set of neighbors of the i-th spin and

$$Z_{j\to i}(x_j) = e^{h_j x_j} \prod_{k\in\partial j\setminus i} \left(\sum_{x_k} e^{J_{kj}x_k x_j} Z_{k\to j}(x_j) \right) \tag{10.33}$$

is the partition function of the subtree rooted at the node j.

Let us illustrate the general scheme on example of the tree in Fig. (10.5b), one obtains

$$Z = \sum_{x_4} Z(x_4), \tag{10.34}$$

The partition function, partially summed and conditioned to the spin value at the spin, x_4, is

$$Z(x_4) = e^{h_4 x_4} \sum_{x_5} e^{J_{45}x_4 x_5} Z_{5\to 4}(x_5) \sum_{x_6} e^{J_{46}x_4 x_6} Z_{6\to 4}(x_6) \sum_{x_3} e^{J_{34}x_3 x_4} Z_{3\to 4}(x_3) \tag{10.35}$$

where

$$Z_{3\to 4}(x_3) = e^{h_3 x_3} \sum_{x_1} e^{J_{13}x_1 x_3} Z_{1\to 3}(x_1) \sum_{x_2} e^{J_{23}x_2 x_3} Z_{2\to 3}(x_2). \tag{10.36}$$

Exercise 10.3. Consider the Ising model on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with spins $\boldsymbol{x}$. Let $\mathcal{V}_0 \subset \mathcal{V}$, and let $\bar{\mathcal{V}}_0 = \{i \in \mathcal{V} \setminus \mathcal{V}_0 : (i,j) \in \mathcal{E} \text{ for some } j \in \mathcal{V}\}$. (You may think of $\bar{\mathcal{V}}_0$ as the "boundary" of $\mathcal{V}_0$ within $\mathcal{V}$.) Show that the spins on $\mathcal{V}_0$ are conditionally independent of all other spins, given values of spins on $\bar{\mathcal{V}}_0$.

10.2.7 Properties of Undirected Tree-Structured Graphical Models

It appears that in the case of a general pair-wise graphical model over trees the joint distribution function over all variables can be expressed solely via single-node marginals and pair-vise marginals over all pairs of the graph-neighbors. To illustrate this important factorization property, let us consider examples shown in Fig. 10.6. In the case of the two-nodes example of Fig. 10.6a the statement is obvious as following directly from the Bayes formula

$$P(x_1, x_2) = P(x_1)P(x_2|x_1), \tag{10.37}$$

or, equivalently, $P(x_1, x_2) = P(x_2)P(x_1|x_2)$.

For the pair-wise graphical model shown in Fig. 10.6b one obtains

$$\begin{aligned} P(x_1, x_2, x_3) = P(x_1, x_2)P(x_3|x_1, x_3) = P(x_1, x_2)P(x_3|x_2) = \\ = P(x_1)P(x_2|x_1)P(x_3|x_2) = \frac{P(x_1, x_2)P(x_2, x_3)}{P(x_2)}, \end{aligned} \tag{10.38}$$

where the conditional independence of x_3 on x_1, $P(x_3|x_1, x_2) = P(x_3|x_2)$, was used.

Next, let us work it out on the example of the pair-wise graphical model shown in Fig. 10.6

$$\begin{aligned} P(x_1, x_2, x_3, x_4) = P(x_1, x_2, x_3)P(x_4|x_1, x_3, x_4) = P(x_1, x_2, x_3)P(x_4|x_2) = \\ = P(x_1, x_2)P(x_3|x_1, x_2)P(x_4|x_2) = P(x_1, x_2)P(x_3|x_2)P(x_4|x_2) = \\ = P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_2) = \frac{P(x_1, x_2)P(x_2, x_3)P(x_2, x_4)}{P^2(x_2)}. \end{aligned} \tag{10.39}$$

Here one uses the following reductions, $P(x_4|x_1, x_3, x_4) = P(x_4|x_2)$ and $P(x_3|x_1, x_2) = P(x_3|x_2)$, related to respective independence properties.

Finally, it is easy to verify that the joint probability distribution corresponding to the model in Fig. 10.6d is

$$\begin{aligned} P(x_1, x_2, x_3, x_4, x_5, x_6) = P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_2)P(x_5|x_2)P(x_6|x_5) = \\ = \frac{P(x_1, x_2)P(x_2, x_3)P(x_2, x_4)P(x_2, x_5)P(x_5, x_6)}{P^3(x_2)P(x_5)}. \end{aligned} \tag{10.40}$$

Exercise 10.4. In the case of a general tree-like graphical model joint probability distribution can be stated in terms of the pair-wise and singleton marginals as follows

$$P(x_1, x_2, \ldots, x_n) = \frac{\prod_{(i,j)\in\mathcal{E}} P(x_i, x_j)}{\prod_{i\in\mathcal{V}} P^{q_i-1}(x_i)}, \tag{10.41}$$

where q_i is the degree of the i-th node. Use mathematical induction to prove Eq. (10.41).

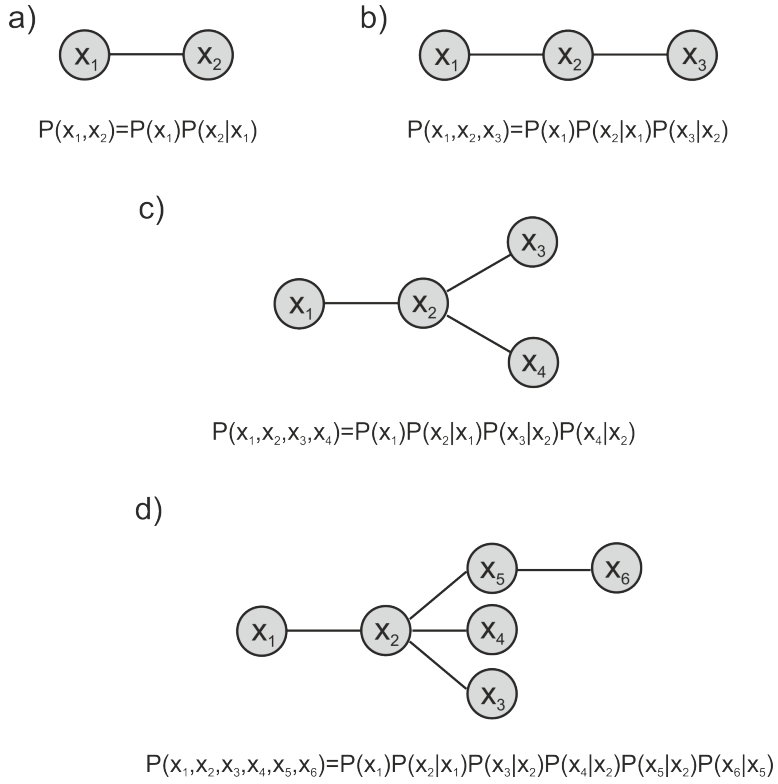


Figure 10.6: Examples of undirected tree-structured graphical models.

10.2.8 Bethe Free Energy & Belief Propagation

As discussed above Dynamic Programming is a provably exact approach for inference when the graph is a tree. It also provides an empirically good approximation for a very broad family of problems stated on loopy graphs.

The approximation is usually called Bethe-Peierls or Belief Propagation (BP is the abbreviation which works for both). Loopy BP is anothe popular term. See the original paper [25], a comprehensive review [26], and respective lecture notes, for an advanced/additional reading.

Instead of Eq. (10.24) one uses the following BP substitution

$$b(x) \to b_{bp}(x) = \frac{\prod_a b_a(x_a)}{\prod_i (b_i(x_i))^{q_i-1}} \tag{10.42}$$

$$\forall a \in \mathcal{V}_f, \quad \forall x_a : \quad b_a(x_a) \geq 0 \tag{10.43}$$

$$\forall i \in \mathcal{V}_n, \quad \forall a \sim i : \quad b_i(x_i) = \sum_{x_a \setminus x_i} b_a(x_a) \tag{10.44}$$

$$\forall i \in \mathcal{V}_n : \quad \sum_{x_i} b_i(x_i) = 1. \tag{10.45}$$

where q_i stands for degree of node i. The physical meaning of the factor $q_i - 1$ on the rhs of Eq. (10.42) is straightforward: by placing beliefs associated with the factor-nodes connected by an edge with a node, i, we over-count contributions of an individual variable q_i times and thus the denominator term in Eq. (10.42) comes as a correction for this over-counting.

Substitution of Eqs. (10.42) into Eq. (10.23) results in what is called Bethe Free Energy (BFE)

$$\mathcal{F}_{bp} := E_{bp} - \mathcal{H}_{bp}, \tag{10.46}$$

$$E_{bp} := -\sum_a \sum_{x_a} b_a(x_a) \log f_a(x_a) \tag{10.47}$$

$$\mathcal{H}_{bp} = \sum_a \sum_{x_a} b_a(x_a) \log b_a(x_a) - \sum_i \sum_{x_i} (q_i - 1) b_i(x_i) \log b_i(x_i), \tag{10.48}$$

where E_{bp} is the so-called self-energy (physics jargon) and $\mathcal{H}_{bp}$ is the BP-entropy (this name should be clear in view of what we have discussed about entropy so far). Thus the BP version of the KL-divergence minimization becomes

$$\arg\min_{b_a, b_i} \mathcal{F}_{bp}\Big|_{\text{Eqs. (10.43,10.44,10.45)}}, \tag{10.49}$$

$$\min_{b_a, b_i} \mathcal{F}_{bp}\Big|_{\text{Eqs. (10.43,10.44,10.45)}} \tag{10.50}$$

Question: Is $\mathcal{F}_{bp}$ a convex function (of its arguments)? [Not always, however for some graphs and/or some factor functions the convexity holds.]

The ML (zero temperature) version of Eq. (10.49) results from the following optimization

$$\min_{b_a,b_i} E_{bp}\Big|_{\text{Eqs. (10.43,10.44,10.45)}} \tag{10.51}$$

Note the optimization is a Linear Programming (LP) — minimizing linear objective over set of linear constraints.

Belief Propagation & Message Passing

Let us restate Eq. (10.49) as an unconditional optimization. We use the standard method of Lagrangian multipliers to achieve it. The resulting Lagrangian is

$$\begin{aligned}\mathcal{L}_{bp}(b,\eta,\lambda) &:= \sum_a \sum_{x_a} b_a(x_a)\log f_a(x_a) - \sum_a \sum_{x_a} b_a(x_a)\log b_a(x_a)\\ &+ \sum_i \sum_{x_i} (q_i-1) b_i(x_i)\log b_i(x_i)\\ &- \sum_i \sum_{a\sim i} \sum_{x_i} \eta_{ia}(x_i)\left(b_i(x_i) - \sum_{x_a\setminus x_i} b_a(x_a)\right) + \sum_i \lambda_i\left(\sum_{x_i} b_i(x_i) - 1\right),\end{aligned} \tag{10.52}$$

where η and λ are the dual (Lagrangian) variables associated with the conditions Eqs. (10.44,10.45) respectively. Then Eq. (10.49) become the following min-max problem

$$\min_b \max_{\eta,\lambda} \mathcal{L}_{bp}(b,\eta,\lambda). \tag{10.53}$$

Changing the order of optimizations in Eq. (10.53) and then minimizing over η one arrives at the following expressions for the beliefs via messages (check the derivation details)

$$\begin{aligned}\forall a,\ \forall x_a: \qquad b_a(x_a) &\sim f_a(x_a)\exp\left(\sum_{i\sim a}\eta_{ia}(x_i)\right) := f_a(x_a)\prod_{i\sim a} n_{i\to a}(x_i)\\ &:= f_a(x_a)\prod_{i\sim a}\prod_{b\sim i}^{b\neq a} m_{b\to i}(x_i)\end{aligned} \tag{10.54}$$

$$\forall i,\ \forall x_i: \qquad b_i(x_i) \sim \exp\left(\frac{\sum_{a\sim i}\eta_{ia}(x_i)}{q_i-1}\right) := \prod_{a\sim i} m_{a\to i}(x_i), \tag{10.55}$$

where, as usual, $\sim$ for beliefs means equality up to a constant which guarantees that the sum of respective beliefs is unity, and we have also introduce the auxiliary variables , m and n, called messages, related to the Lagrangian multipliers η as follows

$$\forall i,\ \forall a\sim i: \quad n_{i\to a}(x_i) := \exp\left(\eta_{ia}(x_i)\right) \tag{10.56}$$

$$\forall a,\ \forall i\sim a: \quad m_{a\to i}(x_i) := \exp\left(\frac{\eta_{ia}(x_i)}{q_i-1}\right). \tag{10.57}$$

Combining Eqs. (10.54,10.55,10.56,10.57) with Eq. (10.44) results in the following BP-equations stated in terms of the message variables

$$\forall i,\ \forall a \sim i,\ \forall x_i: \quad n_{i\to a}(x_i) = \prod_{b\sim i}^{b\neq a} m_{a\to i}(x_i) \tag{10.58}$$

$$\forall a,\ \forall i \sim a,\ \forall x_i: \quad m_{a\to i}(x_i) = \sum_{x_a\setminus x_i} f_a(x_a) \prod_{j\sim a}^{j\neq i} n_{j\to a}(x_j). \tag{10.59}$$

Note that if the Bethe Free Energy (10.46) is non-convex there may be multiple fixed points of the Eqs. (10.58,10.59). The following iterative, so called Message Passing (MP), algorithm (10) is used to find a fixed point solution of the BP Eqs. (10.42,10.43)

Algorithm 10 Message Passing, Sum-Product Algorithm [factor graph representation]

Input: The graph. The factors.

1: $\forall i,\ \forall a \sim i, \forall x_i: \quad m_{a\to i} = 1$ [initialize variable-to-factor messages]
2: $\forall a,\ \forall i \sim a, \forall x_i: \quad n_{i\to 1} = 1$ [initialize factor-to-variable messages]
3: **loop**Till convergence within an error [or proceed with a fixed number of iterations]
4: $\quad \forall i,\ \forall a \sim i,\ \forall x_i: \quad n_{i\to a}(x_i) \leftarrow \prod_{b\sim i}^{b\neq a} m_{a\to i}(x_i)$
5: $\quad \forall a,\ \forall i \sim a,\ \forall x_i: \quad m_{a\to i}(x_i) \leftarrow \sum_{x_a\setminus x_i} f_a(x_a) \prod_{j\sim a}^{j\neq i} n_{j\to a}(x_j)$
6: **end loop**

10.3 Theory of Learning: Sufficient Statistics and Maximum Likelihood Estimation

10.3.1 Sufficient Statistics: infinitely many samples

So far we have been discussing direct (inference) GM problem. In the remainder of this lecture we will briefly talk about inverse problems. This subject will also be discussed (on example of the tree) in the following.

Stated casually - the inverse problem is about 'learning' GM from data/samples. Think about the two room setting. In one room a GM is known and many samples are generated. The samples, but not GM (!!!), are passed to the second room. The task becomes to reconstruct GM from samples.

The first question we should ask is if this is possible in principle, even if we have an infinite number of samples. A very powerful notion of *sufficient statics* helps to answer this question.

Consider the Ising model (not the first time in this course)

$$P(x) = \frac{1}{Z(\theta)} \exp\left\{\sum_{i\in V} h_i x_i + \sum_{\{i,j\}\in E} J_{ij} x_i x_j\right\} = \exp\{\theta^T \phi(x) - \log Z(\theta)\}, \tag{10.60}$$

where $x_i \in \{-1, 1\}$, $\theta := h \cup J = (h_i | i \in \mathcal{V}) \cup (J_{ij} | \{i, j\} \in \mathcal{E})$ and the *partition function* $Z(\theta)$ serves to normalize the probability distribution. In fact, Eq. (10.60) describes what is called the *exponential family* - emphasizing 'exponential' dependence on the factors θ. Notice that any pairwise GM over binary variables can be represented as an Ising model.

Consider collection of all first and second moments (but only these two, and not higher moments) of the spin variables, $\mu^{(1)} := (\mu_i = \mathbb{E}[x_i], i \in V)$ and $\mu^{(2)} := (\mu_{ij} = \mathbb{E}[x_i x_j], \{i, j\} \in E)$. The *sufficient statistics* statement is that to reconstruct θ, fully defining the GM, it is *sufficient* to know $\mu^{(1)}$ and $\mu^{(2)}$.

10.3.2 Maximum-Likelihood Estimation/Learning of Graphical Models

Let us turn the *sufficiency* into a constructive statement – the *Maximum-likelihood estimation* over an exponential family of GMs.

First, notice that (according to the definition of μ)

$$\forall i: \quad \partial_{h_i} \log Z(\theta) = -\mu_i, \quad \forall i, j: \quad \partial_{J_{ij}} \log Z(\theta) = -\mu_{ij}. \tag{10.61}$$

This leads to the following statement: if we know how to compute log-partition function for any values of θ - reconstructing 'correct' θ is a convex optimization problem (over θ):

$$\theta^* = \arg\max_{\theta} \{\mu^T \theta - \log Z(\theta)\} \tag{10.62}$$

If P represents the empirical distribution of a set of independent identically-distributed (i.i.d.) samples $\{x^{(s)}, s = 1, \ldots, S\}$ then μ are the corresponding empirical moments, e.g. $\mu_{ij} = \frac{1}{S} \sum_s x_i^{(s)} x_j^{(s)}$.

General Remarks about GM Learning. The ML parameter Estimation (10.62) is the best we can do. It is fundamental for the task of Machine Learning, and in fact it generalizes beyond the case of the Ising model.

Unfortunately, there are only very few nontrivial cases when the partition function can be calculated efficiently for any values of θ (or parametrization parameters if we work with more general class of GM than described by the Ising models).

Therefore, to make the task of parameter estimation practical one may utilize one of the following approaches:

- Limit consideration to the class of functions for which computation of the partition function can be done efficiently for any values of the parameters. We will discuss such case below – this will be the so-called tree (Chow-Lou) learning. (In fact, the partition function can also be computed efficiently in the case of the Ising model over planar graphs and generalizations, see [27] and references therein for details.)

- Rely on approximations, e.g. such as variational approximation (MF, BP, and other), MCMC or approximate elimination (approximate Dynamical Programming).

- There exists a very innovative new approach - which allows to learn GM efficiently however using more information than suggested by the notion of the *sufficient statistics*. How one of the scientists contributing to this line of research put it – 'the sufficient statistics is not sufficient'. This is a fascinating novel subjects, which is however beyond the scope of this course. Check [28] and references therein.

10.3.3 Learning Spanning Tree

Eq. (10.41) suggests that knowing the structure of the tree-based graphical model allows to express the joint probability distribution in terms of the single-(node) and pairwise (edge-related) marginals. Below we will utilize this statement to pose and solve an inverse problem. Specifically, we attempt to reconstruct a tree representing correlations between multiple (ideally, infinitely many) snapshots of the discrete random variables $x_1, x_2, \ldots, x_n$?

A straightforward strategy to achieve this goal is as follows. First, one estimates all possible single-node and pairwise marginal probability distributions, $P(x_i)$ and $P(x_i, x_j)$, from the infinite set of the snapshots. Then, we may similarly estimate the joint distribution function and verify for a possible tree layout if the relations (10.41) hold. However, this strategy is not feasible as requiring (in the worst unlucky case) to test exponentially many, n^{n-2}, possible spanning threes. Luckily a smart and computationally efficient way of solving the problem was suggested by Chow and Liu in 1968 [29].

Consider the candidate probability distribution, $P_T(x = (x_1, \ldots, x_n))$ over a tree, $T = (\mathcal{V}, \mathcal{E})$ (where $\mathcal{V}$ and $\mathcal{E}$ are the sets of nodes and edges of the tree, respectively) which is tree-factorized according to Eq. (10.41) via marginal (pair-wise and single-variable) probabilities as follows

$$P_T(x_1, x_2, \ldots, x_n) = \frac{\prod_{(i,j)\in\mathcal{E}^F} P(x_i, x_j)}{\prod_{i\in\mathcal{V}^F} P(x_i)^{q_i-1}(x_i)}. \tag{10.63}$$

"Distance" between the actual (correct) joint probability distribution P and the candidate tree-factorized probability distribution, P_T, can be measured in terms of the Kullback-

Leibler (KL) divergence

$$D(P \parallel P_T) = -\sum_x P(x) \log \frac{P(x)}{P_T(x)}. \tag{10.64}$$

As discussed in Section 8.5, the KL divergence is always positive if P and P_T are different, and is zero if these distributions are identical. Then, we are looking for a tree that minimizes the KL divergence.

Substituting (10.63) into Eq. (10.64) one arrives at the following chain of explicit transformations

$$\begin{aligned}
&\sum_x P(x) \left(\log P(x) - \sum_{(i,j)\in\mathcal{E}} \log P(x_i, x_j) + \sum_{i\in\mathcal{V}} (q_i - 1) \log P(x_i) \right) = \\
&= \sum_x P(x) \log P(x) - \sum_{(i,j)\in\mathcal{E}^{\mathcal{F}}} \sum_{x_i, x_j} P(x_i, x_j) \log P(x_i, x_j) + \\
&+ \sum_{i\in\mathcal{V}} (q_i - 1) \sum_{x_i} P(x_i) \log P(x_i) = - \sum_{(i,j)\in\mathcal{E}^{\mathcal{F}}} \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i) P(x_j)} + \\
&+ \sum_x P(x) \log P(x) - \sum_{i\in\mathcal{V}^{\mathcal{F}}} \sum_{x_i} P(x_i) \log P(x_i),
\end{aligned} \tag{10.65}$$

where the following nodal and edge marginalization relations were used, $\forall i \in \mathcal{V}^{\mathcal{F}}: \quad P(x_i) = \sum_{x\setminus x_i} P(x)$, and, $\forall (i,j) \in \mathcal{E}^{\mathcal{F}}: \quad P(x_i, x_j) = \sum_{x\setminus x_i, x_j} P(x)$, respectively. One observes that the Kullback-Leibler divergence becomes

$$D(P \parallel P_F) = - \sum_{(i,j)\in\mathcal{E}^{\mathcal{F}}} I(X_i, X_j) + \sum_{i\in\mathcal{V}^{\mathcal{F}}} S(x_i) - S(x), \tag{10.66}$$

where

$$I(X_i, X_j) := \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i) P(x_j)} \tag{10.67}$$

is the mutual information of the pair of random variables x_i and x_j.

Since the entropies $S(X_i)$ and $S(X)$ do not depend on the tree choice, minimizing the Kullback-Leibler divergence is equivalent to maximizing the following sum over branches of a tree

$$\sum_{(i,j)\in\mathcal{E}^{\mathcal{F}}} I(X_i, X_j). \tag{10.68}$$

Based on this observation, Chow and Liu have suggested to use the following (standard in computer science) Kruskal maximum tree reconstruction algorithm (notice that the algorithm is greedy, i.e. of the Dynamic Programming type):

Table 10.1: Information available about an exemplary probability distribution of four binary variables discussed in the Exercise 10.5.

$x_1x_2x_3x_4$	$P(x_1, x_2, x_3, x_4)$	$P(x_1)P(x_2\|x_1)P(x_3\|x_2)P(x_4\|x_1)$	$P(x_1)P(x_2)P(x_3)P(x_4)$
0000	0.100	0.130	0.046
0001	0.100	0.104	0.046
0010	0.050	0.037	0.056
0011	0.050	0.030	0.056
0100	0.000	0.015	0.056
0101	0.000	0.012	0.056
0110	0.100	0.068	0.068
0111	0.050	0.054	0.068
1000	0.050	0.053	0.056
1001	0.100	0.064	0.056
1010	0.000	0.015	0.068
1011	0.000	0.018	0.068
1100	0.050	0.033	0.068
1101	0.050	0.040	0.068
1110	0.150	0.149	0.083
1111	0.150	0.178	0.083

- (step 1) Sort the edges of $\mathcal{G}$ into decreasing order by weight = **Mutual Information**, i.e. $I(X_i, X_j)$ for the candidate edge (i, j). Let $\mathcal{E}_T$ be the set of edges comprising the maximum weight spanning tree. Set $\mathcal{E}_T = \emptyset$.
- (step 2) Add the first edge to $\mathcal{E}_T$
- (step 3) Add the next edge to $\mathcal{E}_T$ if and only if it does not form a cycle in $\mathcal{E}_T$.
- (step 4) If $\mathcal{E}_T$ has $n-1$ edges (where n is the number of nodes in $\mathcal{G}$) stop and output $\mathcal{E}_T$. Otherwise go to step 3.

Eq. (10.41) is exact only in the case when it is guaranteed that the Graphical Model we attempt to recover forms a tree. However, the same tree ansatz can be used to recover the best tree approximation for a graphical model defined over a graph with loops. How to choose the optimal (best approximation) tree in this case? To answer this question within the aforementioned Kullback-Leibler paradigm one needs to compare the tree ansatz (10.41) and the empirical joint distribution. This reconstruction of the optimal tree is based on the

Chow-Liu algorithm.

Exercise 10.5. Find the Chow-Liu optimal spanning tree approximation for the joint probability distribution of four random binary variables with statistical information presented in the Table 10.1. [*Hint:* Estimate the empirical (i.e. based on the data), pair-wise mutual information and then utilize the Chow-Liu-Kruskal algorithm (see description above in the lecture notes) to reconstruct the optimal tree.]

10.4 Function Approximation with Neural Networks

Material presented in this lecture is relatively new. Some of the theoretical results discussed below are 30 years old, but practical power of these results, and of course the flow of new results and approaches, are very recent – 5 ears old, or even younger. In this situation there are not that many books written on the topics yet, especially books focusing on the applied mathematics aspects of the function approximation with Neural Networks (NN). The book of Gilbert Strang [30] on "Linear Algebra and Learning from Data", which we highly recommend, stands alone. Part VII of the book, entitled "Learning from Data", is especially relevant to this Section.

NN, and especially Deep Neural Networks (DNN), is the newest most important tool of applied mathematics which can be used universally to fit a function.

Mathematical foundations of the methodology is established by the following Theorem:

Theorem 10.4.1 (Universal Approximation Theorem (Cybenko 1989; Hornik 1991; Pinkus 1999))**.** Let $\rho : \mathbb{R} \to \mathbb{R}$ be any continuous function (called activation function). Let $\mathcal{NN}_n^\rho$ represent the class of feed-forward NN with activation function, ρ, with n neurons in the input layer, one neuron in the output layer, and one arbitrary layer with an arbitrary number of neurons. Let $K \subset \mathbb{R}$ be a compact. Then $\mathcal{NN}_n^\rho$ is dense in $C(K)$ (class of differentiable function on K if and only if ρ is not polynomial

This famous theorem was an inhibitor of the NN revolution. In its original form, stated above, it therefore applied to the regime of bounded depth and arbitrary width, was also extended recently to the complementary case of the bounded width and arbitrary depth (See [31] and references therein. Notice that the term "deep" in the name of Deep Learning refers to the large depth of respective NN.)

Even though, the Theorem 10.4.1 is agnostic to the type of the activation function some activation functions are used more frequently in practice. The choice which has succeeded far beyond expectations is the nonlinear function called Rectified Linear Unit, or simply, $\mathrm{ReLU}(x) = x_+ = \max(x, 0)$.

Obviously, NNs are not limited to representing an arbitrary $\rho : \mathbb{R} \to \mathbb{R}$ map. For example, we can use ReLU to construct the following piecewise-linear function mapping a p-dimensional data vector, v, to an m-dimensional output:

- Choose, (q, p)-dimensional matrix A_1 and q-dimensional vector b_1 and set to zero all negative components of $A_1 v + b_1$, i.e. introduce $\mathrm{ReLU}(A_1 v + b_1) = (A_1 v + b_1)_+$ acting component-wise, where each component is associated with a neuron of a hidden layer.
- Choose (m, q) dimensional matrix A_2 and apply it to the hidden layer vector, $A_2(A_1 v + b_1)_+$.

Introducing depth in NN allows to construct more and more expressive, as containing more piece-wise linear pieces, continuous piece-wise linear functions. One standard way to do it is to use *composition*:

$$\mathcal{NN}(v) = F_L(F_{L-1}(\cdots F_2(F_1(v)))) = (F_L \circ F_{L-1} \circ \cdots \circ F_2 \circ F_1)(v), \tag{10.69}$$

where $l = 1, \cdots, L: \ F_l(x) = (A_l x + b_l)_+$.

While the composition may be considered as the key operation in construction of the Deep NN, the *loss function*, the *chain rule* and associated *automatic differentiation* and *back-propagation* are the other key ingredients of DNN, which we discuss in the following, one after another.

Example 10.4.2 (Counting Number of Pieces)**.** Consider an example of $\mathcal{NN}(v)$ with ReLU activation function, one dimensional input layer, one 5-dimensional hidden layer, and one dimensional output layer, and count the number of pieces in the resulting continuous piece-wise linear function.

10.4.1 Fitting a Function with NN as an Optimization

Key element of fitting a function with NN is the so-called *loss function*, which is the term commonly used in data science to describe objective of the underlying optimization formulation. Standard choice of the loss function is a norm, e.g. l_1, l_2 or l_∞ norm, of the error between the function, $F(v)$, and its NN-approximation evaluated at the available samples, $s = 1, \cdots, S$. Then, for the example of the l_p norm min-error, the resulting optimization becomes

$$\min_{\theta} \sum_{s=1}^{S} \|F(v_s) - \mathcal{NN}(v_s|\theta)\|_p, \tag{10.70}$$

where θ is the vector of the NN parameters, e.g. $\theta = (A_L, b_L, \cdots, A_1, b_1)$ in the case of the continuous piece-wise NN given by Eq. (10.69).

A standard way to solve the optimization is to evaluate partial derivatives (components of the gradient) of the loss function over the parameters and require that all of them are zero. This zero gradient solution of the optimization problem is normally found by the gradient descent algorithm (by one of its many versions).

In all of its variants, and according to Eq. (10.70) the gradient descent needs to compute derivatives of $\mathcal{NN}(v_s|\theta)$ over the components of the parameter vector θ. *Automatic differentiation*, or its particular version *back-propagation*, is a method to compute the derivatives quickly.

10.4.2 Automatic Differentiation, Back-Propagation and the Chain Rules

Computational efficiency of the Automatic Differentiation is associated with the so-called chain rule, illustrated on the example

$$\frac{dg}{dx} = \frac{d}{dx}(g_3(g_2(g_1(x)))) = \left(\frac{dg_3}{dg_2}(g_2(g_1(x)))\right)\left(\frac{dg_3}{dg_1}(g_1(x))\right)\left(\frac{dg_1}{dx}(x)\right).$$

Steps in the automatic differentiation can be arranged in two ways – forward mode and backwards mode. We choose forward mode (as much faster ones) if we have many functions g depending on a few inputs (components of x), and vice versa we choose backward mode if we have fewer function and higher dimensional input.

Since in Deep Learning we have one loss function depending on many weights, the choice of the back-propagation mode (of the automatic differentiation) is natural for this application.

x above is the vector of weights consisting of all the matrices, $A_1, \cdots, A_L$, where L is the number of layers, and all the bias vectors, $b_1, \cdots, b_L$. The input-output pair of vectors, (v, w), are associated with the 0'th and L's layer of the NN. In the super-vised learning discussed here, $v = v_0$, and, $w = v_L$, are the training data represented through S samples, $(v^{(s)}, w^{(s)})$, $s = 1, \cdots, S$. NN maps inputs to outputs, $w = F(x, v_0)$. Each new layer is a map from the previous layer, $v_n = R_n(b_n + A_n v_{n-1})$, where R_n is the activation function of the n-th layer, for example ReLu.

Consider example of one hidden layer, $L = 2$, with R_2 set as the identity function

$$w = v_2 = b_2 + A_2 v_1 = b_2 + A_2 R(b_1 + A_1 v_0).$$

Let us compute derivatives over A_1 following the chain rule:

$$\frac{\partial w}{\partial A_1} = A_2 R'(b_1 + A_1 v_0)\frac{\partial(b_1 + A_1 v_0)}{\partial A_1}.$$

Notice that computing the derivative we go backwards from $w = v_2$ to $v = v_0$ – that is we follow the backward propagation rule starting from the output and moving to the input.

10.4.3 Avoiding Over-fitting

Over-fitting occurs when a trained network performs very accurately on the given data, but cannot generalize well to new data. If F does purely on the training set we say that the NN is under-fit. Over-fitting occurs when F does well on the training set but gives much larger error on the validation data. Colloquially, it occurs when our function is not sufficiently smooth – filling gaps (unnecessarily) between the training data points. One standard recommendation: to avoid over-fitting introduce regularization to the loss function. Another recommendation: stop training before you over-fit.

Example 10.4.3. Consider a Neural Network (NN) with $L = 2$, with one hidden layers each consisting of a single node (neuron), and with the tanh activation functions. Therefore, the model is

$$w = v_2 = \tanh\Big(a_2 \tanh\big(a_1 v + b_1\big) + b_2\Big),$$

and assume that the weights are currently set at $(a_1, b_1) = (1.0, 0.5)$ and $(a_2, b_2) = (-0.5, 0.3)$. What is the gradient of the Mean Square Error (MSE) cost for the observation $(x, y) = (2, -0.5)$? What is the optimal MSE and optimal values of the parameters?

Solution. Evaluating the function (forward pass) with the initial parameters gives:

$$w = \tanh(-0.5 \tanh(1.0(2) + 0.5) + 0.3) = -0.1909$$

Define and compute the intermediary variables z_1, a_1, z_2 as follows:

$$\begin{aligned} z_1 &:= a_1 v + b_1 = 2.500,\ v_1 := \tanh(a_1 v + b_1) = 0.9866 \\ z_2 &:= a_2 v_1 + b_2 = -.1933,\ \hat{w} = v_2 = \tanh(a_2 v_1 + b_2) = -0.1909, \end{aligned}$$

where $\hat{w}$ stands for the NN estimate of the output, as opposed to the actual, i.e. sample, output, w.

The MSE loss function is $\mathcal{L} = (\hat{w} - w)^2$. The gradient of the loss function is

$$\nabla \mathcal{L} = \begin{pmatrix} \frac{\partial \mathcal{L}}{\partial a_2} \\ \frac{\partial \mathcal{L}}{\partial b_2} \\ \frac{\partial \mathcal{L}}{\partial a_1} \\ \frac{\partial \mathcal{L}}{\partial b_1} \end{pmatrix} = \begin{pmatrix} \frac{\partial L}{\partial \hat{w}} \frac{\partial \hat{w}}{\partial z_2} \frac{\partial z_2}{\partial a_2} \\ \frac{\partial L}{\partial \hat{w}} \frac{\partial \hat{w}}{\partial z_2} \frac{\partial z_2}{\partial b_2} \\ \frac{\partial L}{\partial \hat{w}} \frac{\partial \hat{w}}{\partial z_2} \frac{\partial z_2}{\partial a_1} \frac{\partial a_1}{\partial z_1} \frac{\partial z_1}{\partial a_1} \\ \frac{\partial L}{\partial \hat{w}} \frac{\partial \hat{w}}{\partial z_2} \frac{\partial z_2}{\partial a_1} \frac{\partial a_1}{\partial z_1} \frac{\partial z_1}{\partial b_1} \end{pmatrix}$$

Evaluating each partial derivative gives:

$$\begin{aligned}
\frac{\partial L}{\partial \hat{w}} &= 2(\hat{w} - w) = 2((-0.1909) - (-0.5)) = 0.6182 \\
\frac{\partial \hat{w}}{\partial z_2} &= 1 - (-.1933)^2 = 0.9626 \\
\frac{\partial z_2}{\partial a_2} &= a_1 = 0.9866 \\
\frac{\partial z_2}{\partial b_2} &= 1 \\
\frac{\partial z_2}{\partial a_1} &= a_2 = -0.5 \\
\frac{\partial a_1}{\partial z_1} &= 1 - (.9866)^2 = .0266 \\
\frac{\partial z_1}{\partial a_1} &= x = 2.0 \\
\frac{\partial z_1}{\partial b_1} &= 1
\end{aligned}$$

Putting this all together, we get:

$$\nabla \mathcal{L} = \begin{pmatrix} \frac{\partial \mathcal{L}}{\partial a_2} \\ \frac{\partial \mathcal{L}}{\partial b_2} \\ \frac{\partial \mathcal{L}}{\partial a_1} \\ \frac{\partial \mathcal{L}}{\partial b_1} \end{pmatrix} = \begin{pmatrix} \frac{\partial L}{\partial \hat{w}} \frac{\partial \hat{w}}{\partial z_2} \frac{\partial z_2}{\partial a_2} \\ \frac{\partial L}{\partial \hat{w}} \frac{\partial \hat{w}}{\partial z_2} \frac{\partial z_2}{\partial b_2} \\ \frac{\partial L}{\partial \hat{w}} \frac{\partial \hat{w}}{\partial z_2} \frac{\partial z_2}{\partial a_1} \frac{\partial a_1}{\partial z_1} \frac{\partial z_1}{\partial w_1} \\ \frac{\partial L}{\partial \hat{w}} \frac{\partial \hat{w}}{\partial z_2} \frac{\partial z_2}{\partial a_1} \frac{\partial a_1}{\partial z_1} \frac{\partial z_1}{\partial b_1} \end{pmatrix} = \begin{pmatrix} 0.5951 \\ 0.5871 \\ -0.0079 \\ -0.0158 \end{pmatrix}$$

(b) There is only one data point, and four parameters. Therefore the problem is underdetermined. Iterating gradient descent-type methods does not guarantee return of a unique solution.

Exercise 10.6. Consider the following two-layer ($L = 2$) NN-map, $v \to w$ where $v, w \in \mathbb{R}$, built from three ReLU neurons :

$$v_{1i} = \text{ReLU}(a_i v + b_i),\ \forall i = 1, 2,\ \boldsymbol{v}_1 = (v_{11}, v_{12}) \in \mathbb{R}^2,\ w = v_2 = \text{ReLU}(\boldsymbol{A}_3 \cdot \boldsymbol{v}_1^T + b_3),$$

where thus, $\boldsymbol{A}_3 \in \mathbb{R}^{1\times 2}$, and, $a_1, a_2, b_1, b_2, b_3 \in \mathbb{R}$, are the parameters.

(a) Describe the complexity of the class of functions representing this NN.

(b) What is the minimal number P of non-degenerate samples, $(v^{(p)}, w^{(p)})$, $p = 1, \cdots, P$, needed for exact (!) reconstruction of the NN's parameters?

(c) Build an example where this NN outputs continuous piece-wise linear function with two linear intervals.

Appendix A

Convex and Non-Convex Optimization *

This Appendix was originally prepared by Dr. Yury Maximov from Los Alamos National Laboratory (and edited by MC). The material was presented in 2020 in 6 lectures cross-cut between Math 581 (then Math 583), Math 584 (then Math 527) and Math 589 (then Math 575). In 2021 material of the Appendix was mainly covered in Math 584 (then Math 527), with a brief summary included in Math 581 (then Math 583) via 1.5 lecture and two exercises. In 2022 we do not include the material in Math 581 at all. The Appendix should thus be viewed as a $*$-Section of the main text – suggested for an extra-curriculum study *.

The Appendix is split into four Subsections. Sections A.1 and A.2 will be discussing basic convex and non-convex finite dimensional optimizations. Then in Sections A.3 and A.4 we will turn to discussing iterative optimization methods for the optimization formulations, set in Sections A.1 and A.2, which are of constrained and unconstrained types.

The most general problem we will start our discussion from in Section A.1 consists in minimization of a function, $f : S \subseteq \mathbb{R}^n \to \mathbb{R}$:

$$\begin{aligned} & f(x) \to \min \\ & \texttt{s.t.:} \quad x \in S \subseteq \mathbb{R}^n. \end{aligned} \tag{A.1}$$

Notice variability in notations – an absolutely equivalent alternative expression is

$$\min_{x \in S \subseteq \mathbb{R}^n} f(x).$$

Section A.1 should be viewed as introductory (setting notations) leading us to discussion of the notion of (optimization) duality in Section A.2.

*This auxiliary Appendix can be dropped at the first reading. Material form the Subsection will not contribute midterm and final exams of Math 581.

Iterative algorithms, discussed in Sections A.3 and A.4, will be designed to solve Eq. (A.1). Each step of such an algorithm will consist in updating the current estimate, x_k, using x_j,$f(x_j)$, $j \leq k$, possibly its vector of derivatives $\nabla f(x)$, and possibly the Hessian matrix, $\nabla^2 f(x)$, such that the optimum is achieved in the limit, $\lim_{k\to+\infty} f(x_k) = \inf_{x\in S\subseteq\mathbb{R}^n} f(x)$.

Different iterative algorithms can be classified depending on the information available, as follows:

- *Zero-order algorithm*, where at each iteration step one has an access to the value of $f(x)$ at a given point x (but no information on $\nabla f(x)$ and $\nabla^2 f(x)$ is available);
- *First-order optimization*, where at each iteration step one has an access to the value of $f(x)$ and $\nabla f(x)$;
- *Second-order algorithm*, where at each iteration step one has an access to the value of $f(x), \nabla f(x)$ and $\nabla^2 f(x)$;
- *Higher-order algorithm* where at each iteration step one has an access to the value of the objective function, its first, second and higher-order derivatives.

We will not discuss in these notes second-order and higher-order algorithm, focusing in Sections A.3 and A.4 primarily on the first-order and second order algorithms.

A.1 Convex Functions, Sets and Optimizations

Calculus of Convex Functions and Sets

An important class of functions one can efficiently minimize are convex functions, that were introduced earlier in Definition 6.7.3. We restate it here for convenience.

Definition A.1.1 (Definition 6.7.3)**.** A function, $f : \mathbb{R}^n \to \mathbb{R}$ is convex if

$$\forall x, y \in \mathbb{R}^n,\ \lambda \in (0,1): \quad f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

If a function is smooth, one can give an equivalent definition of convexity.

Definition A.1.2. A smooth function $f(x) : \mathbb{R}^n \to \mathbb{R}$ is convex, if

$$\forall x, y \in \mathbb{R}^n:\ f(y) \geq f(x) + \nabla f(x)^\top (y - x).$$

Definition A.1.3. Let function $f : \mathbb{R}^n \to \mathbb{R}$ has smooth gradient. Then f is convex iff

$$\forall x:\ \nabla^2 f(x) (:= \partial_{x_i}\partial_{x_j} f(x); \forall i, j = 1, \cdots, n) \succeq 0,$$

that is the Hessian of the function is a positive semi-definite matrix at any point. (Remind that real symmetric $n \times n$ matrix H is positive semi-definite iff $x^T H x \geq 0$ for any $x \in \mathbb{R}^n$.)

Lemma A.1.4. Prove that the definitions above are equivalent for sufficiently smooth functions.

Proof. Assume that the function is convex according to the Definition A.1.1. Then for any $h \in \mathbb{R}^n$, $\lambda \in [0,1]$, one has according to the Definition A.1.1:

$$f(\lambda(x+h)+(1-\lambda)x) - f(x) = f(x+\lambda h) - f(x) \leq \lambda(f(x+h)-f(x)).$$

That is

$$f(x+h) - f(x) \geq f(x+\lambda h) - f(x) = \nabla f(x)^\top h + O(\lambda) \qquad \forall \lambda \in [0,1]$$

Then taking the limit for $\lambda \to 0$ one has $\nabla f(x)^\top h \leq f(x+h) - f(x),\ \forall h \in \mathbb{R}^n$ which is exactly Def. A.1.2. Vice versa, if $\forall x, y:\ f(y) \leq \nabla f(x)^\top (y-x)$, one has for $z = \lambda x + (1-\lambda)y$, and any $\lambda \in [0,1]$:

$$\begin{aligned} f(y) &\geq f(z) + \nabla f(z)^\top (y-z) = f(z) + \lambda \nabla f(z)^\top (y-x), \\ f(x) &\geq f(z) + \nabla f(z)^\top (x-z) = f(z) + (1-\lambda)\nabla f(z)^\top (x-y) \end{aligned}$$

summing up the inequalities above with the quotients $1-\lambda$ and λ one gets $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$. Thus Def. A.1.1 and Def. A.1.2.

Further, if f is sufficiently smooth, one has according to the Taylor expansion:

$$f(y) = f(x) + \nabla f(x)^\top (y-x) + \frac{1}{2}(y-x)\nabla^2 f(x)(y-x) + o(\|y-x\|_2^2).$$

Taking $y \to x$ one gets from the Definition A.1.2 to the Definition A.1.3 and vice versa. □

Definition A.1.5. Function $f(x)$ is *concave* iff $-f(x)$ is convex.

Definition A.1.2 is probably the most practical. To generalize it to non-smooth functions, we introduce the notion of *sub-gradient*.

Definition A.1.6. Vector $g \in \mathbb{R}^n$ is a sub-gradient of the convex function f, $f : \mathbb{R}^n \to \mathbb{R}$, at point x iff

$$\forall y \in \mathbb{R}^n : f(y) \geq f(x) + g^\top (y-x).$$

Set $\partial f(x)$ is a set of all sub-gradients for the function f at point x.

To establish some properties of the sub-gradients (which can also be called sub-differentials) let us introduce the notion of convex set, i.e. a segment between any point of the set which belongs to the set as well.

Definition A.1.7. Set S is convex, iff for any $x_1, x_2 \in S$, and $\theta \in [0,1]$ one has $x_1\theta + x_2(1-\theta) \in S$. In other words, set S is convex if for any points x_1, x_2 in it, the set contains a line segment $[x_1, x_2]$.

Theorem A.1.8. For any convex function f, $f : \mathbb{R}^n \to \mathbb{R}$, and any point $x \in \mathbb{R}^n$ the sub-differential $\partial f(x)$ is a convex set. In other words, for any $g_1, g_2 \in \partial f(x)$ one has $\theta g_1 + (1-\theta)g_2 \in \partial f(x)$. Moreover, $\partial f(x) = \{\nabla f(x)\}$ if f is smooth.

Proof. Let $g_1, g_2 \in \partial f(x)$, then $f(y) \geq f(x) + g_1^\top (y-x)$, and $f(y) \geq f(x) + g_2^\top (y-x)$. That is for any $\lambda \in [0,1]$ one has $f(y) \geq f(x) + (\lambda g_1 + (1-\lambda)g_2)^\top (y-x)$ and $\lambda g_1 + (1-\lambda)g_2$ is a sub-gradient as well. We conclude that the set of all the sub-gradients is convex. Moreover, if f is smooth, according to the Taylor expansion formula one has $f(x+h) = f(x) + \nabla f(x)^\top h + O(\|h\|)_2^2$. Assume that there exists sub-gradient $g \in \partial f(x)$ other than $\nabla f(x)$ (as $\nabla f(x) \in \partial f(x)$ by the definition of convex functions A.1.2). Then $f(x) + g^\top h \leq f(x+h) = f(x) + \nabla f(x)^\top h + O(\|h\|_2^2)$ and similarly $f(x) - g^\top h \leq f(x-h) = f(x) - \nabla f(x)^\top h + O(\|h\|_2^2)$, and

$$g^\top h \leq \nabla f^\top h + O(\|h\|_2^2) \quad \text{and} \quad g^\top h \geq \nabla f^\top h + O(\|h\|_2^2)$$

which implies $g = \nabla f(x)$, therefore concluding the proof. □

Let us illustrate the sub-gradient calculus on the following examples:

- Sub-differential of $|x|$ is

$$\partial f(x) = \begin{cases} 1, & \text{if } x > 0 \\ -1, & \text{if } x < 0 \\ [-1,1], & \text{if } x = 0. \end{cases}$$

- Sub-differential of $f(x) = \max\{f_1(x), f_2(x)\}$ is

$$\partial f(x) = \begin{cases} \nabla f_1(x), & \text{if } f_1(x) > f_2(x) \\ \nabla f_2(x), & \text{if } f_1(x) < f_2(x) \\ \{\theta \nabla f_1(x) + (1-\theta)\nabla f_2(x), \theta \in [0,1]\}, & \text{if } f_1(x) = f_2(x) \end{cases}$$

  if f_1 and f_2 are smooth functions on $\mathbb{R}^n$.

Exercise A.1. Consider $f(x,y) = \sqrt{x^2 + 4y^2}$. Prove that f is convex. Sketch level curves of f. Find the sub-differential $\partial f(0,0)$.

Example A.1.9. Examples of convex functions include:

a) $x^p, p \geq 1$ or $p \leq 0$ is convex; $x^p, 0 \leq p \leq 1$ is concave;

b) $\exp(x), x \in \mathbb{R}$ and $-\log x, x \in \mathbb{R}_{++}$, are convex;

c) $f(h(x))$, where $f : \mathbb{R} \to \mathbb{R}$, $h : \mathbb{R} \to \mathbb{R}$ is convex if

   (a) $f(x)$ is convex and non-decreasing, and $h(x)$ is convex;

   (b) Or $f(x)$ is convex and non-increasing, $h(x)$ is concave;

   To prove the statement for smooth functions we consider

$$g''(x) = f''(h(x))(h'(x))^2 + f'(h(x))h''(x)$$

   One can also extend the statement to non-smooth and multidimensional functions.

d) LogSumMax, also called soft-max, $\log\left(\sum_{i=1}^n \exp(x_i)\right)$, is convex in $x \in \mathbb{R}^n$ as a composition of a convex non-decreasing and a convex function. The soft-max function plays a very important role because it bridges smooth and non-smooth optimizations:

$$\max(x_1, x_2, \ldots, x_n) \approx \frac{1}{\lambda} \log\left(\sum_{i=1}^n \exp(\lambda x_i)\right), \ \lambda \to 0, \ \lambda > 0. \tag{A.2}$$

e) Ratio of the quadratic function of on variable to a linear function of another variable, e.g. $f(x, y) = x^2/y$, is jointly convex in x and y at $y > 0$;

f) Vector norm: $\|x\|_p := \left(|x_i|^p\right)^{1/p}, x \in \mathbb{R}^n$, also called p-norm, or ℓ_p-norm when $p \geq 1$, is convex.

g) Dual norm $\|\cdot\|_*$ to $\|\cdot\|$ is $\|y\|_* := \sup_{\|x\| \leq 1} x^\top y$. The dual norm is always convex.

h) Indicator function of a convex set, $I_S(x)$, is convex:

$$I_S(x) = \begin{cases} 0, & x \in S \\ +\infty, & x \notin S \end{cases}$$

Example A.1.10. Examples of convex sets:

1. If (any number of) sets $\{S_i\}_i$ are convex, then $\bigcap_i S_i$ is convex;

2. Affine image of a convex set:

$$\bar{S} = \{x : Ax + b, x \in S\}$$

3. Image (and inverse image) of a convex set S under perspective mapping $P : \mathbb{R}^{n+1} \to \mathbb{R}^n$, $P = x/t$, $\operatorname{dom} P = \{(x,t) : t > 0\}$.
   Indeed, consider $y_1, y_2 \in P(S)$ so that $y_1 = x_1/t_1$ and $y_2 = x_2/t_2$. We need to prove that for any $\lambda \in [0,1]$
   $$y = \lambda y_1 + (1-\lambda)y_2 = \lambda\frac{x_1}{t_1} + (1-\lambda)\frac{x_2}{t_2} = \frac{\theta x_1 + (1-\theta)x_2}{\theta t_1 + (1-\theta)t_2}$$
   which holds for $\theta = \lambda t_2/(\lambda t_1 + (1-\lambda)t_2)$. The proof of the inverse statement is similar.

4. Image of a convex set under the linear-fractional function, $f : \mathbb{R}^{n+1} \to \mathbb{R}^n$, $f(x) = \frac{Ax+b}{c^\top x+d}$, $\operatorname{dom} f = \{x : c^\top x + d > 0\}$. Indeed, $f(x)$ is a perspective transform of an affine function.

Exercise A.2. Check that all functions and all sets above are convex using Definition A.1.1 of the convex function (or equivalent Definitions A.1.2,A.1.3) and the Definition A.1.7 of the complex set.

In further analysis, we introduce a special subclass of convex functions for which one can guarantee much faster convergence than for minimization of a general convex function.

Definition A.1.11. Function $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex with respect to norm $\|\cdot\|$ for some $\mu > 0$, iff

1. $\forall x, y:\ f(y) \geq f(x) + \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|y-x\|^2$
2. if f is sufficiently smooth, the strong convexity condition in ℓ_2 norm is equivalent to $\forall x:\ \nabla^2 f(x) \succeq \mu$.

As we will see later, generalization of the strong convexity definition A.1.11 to a general ℓ_p norm allows to design more efficient algorithms in various cases. (Concavity, strong concavity and convexity in ℓ_p are defined by analogy.)

Exercise A.3. Find a subset of $\mathbb{R}^3$ containing $(0,0,0)$ such that $f(u) = \sin(x+y+z)$ is (a) convex; (b) strongly convex.

Exercise A.4. Is it true that the functions, $f(x) = x^2/2 - \sin x$ and $g(x) = \sqrt{1 + x^\top x}$, $x \in \mathbb{R}^n$, are convex. Are the functions strongly convex?

Exercise A.5. Check if the function $\sum_{i=1}^n x_i \log x_i$ defined on $\mathbb{R}^n_{++}$ is

- convex/concave/strongly convex/strongly concave?
- strongly convex/concave in ℓ_1, ℓ_2, ℓ_∞ ?

Hint: to prove that the function is strongly convex in ℓ_p norm it is sufficient to show that
$$h^\top \nabla^2 f(x)\ h\ \geq\ \|h\|_p^2$$

Convex Optimization Problems

The optimization problem

$$f(x) \to \min_{x \in S \subseteq \mathbb{R}^n}$$

is convex if $f(x)$ and S are convex. *Complexity* of an iterative algorithm initiated with x_0 to solve the optimization problem is measured in the number of iterations required to get a point x_k such that $|f(x_k) - \inf_{x \in S \subseteq \mathbb{R}^n} f(x)| < \varepsilon$. Each iteration means an update of x_k. Complexity classification is as follows

- linear, that is the number of iterations $k = O(\log(1/\varepsilon))$, and in other words $f(x_{k+1}) - \inf_{x \in S} f(x) \leq c(f(x_k) - \inf_{x \in S} f(x))$ for some constant c, $0 < c < 1$. Roughly, after iteration we increase the number of correct digits in our answer by one.

- quadratic, that is $k = O(\log\log(1/\varepsilon))$, and $f(x_{k+1}) - \inf_{x \in S} f(x) \leq c(f(x_k) - \inf_{x \in S} f(x))^2$ for some constant c, $0 < c < 1$. That is, after iteration we double the number of correct digits in our answer.

- sub-linear, that is characterized by the rate slower than $O(\log(1/\varepsilon))$. In convex optimization, it is often the case that the convergence rate for different methods is $k = O(1/\varepsilon)$, $O(1/\varepsilon^2)$, or $O(1/\sqrt{\varepsilon})$ depending on the properties of function f.

Consider an optimization problem

$$\begin{aligned} f(x) &\to \min_{x \in \mathbb{R}^n} \\ \texttt{s.t.:} g(x) &\leq 0 \\ h(x) &= 0 \end{aligned}$$

If the inequality constraint $g(x)$ is convex and the equality constraint is affine, $h(x) = Ax + b$, a feasible set of this problem, $S = \{x : g(x) \leq 0 \text{ and } h(x) = 0\}$, is convex that follows immediately from definitions of a convex set and a convex function. As we will see later in the lectures, in contrast to non-convex problems the convex ones admit very efficient and scalable solutions.

Exercise A.6. Let $\Pi_C^{\ell_p}(x)$ be a projection of a point x to a convex compact set C in ℓ_p norm, if

$$\Pi_C^{\ell_p}(x) = \arg\min_{y \in C} \|x - y\|_p.$$

Find ℓ_1, ℓ_2, ℓ_∞ projections of $x = \{1, 1/2, 1/3, \ldots, 1/n\} \in \mathbb{R}^n$ on the unit simplex $S = \{x : \sum_{i=1}^{n} |x_i| = 1\}$. Which of the ℓ_1, ℓ_2, ℓ_∞ projections of an arbitrary point $x \in \mathbb{R}^n$ to a unit simplex is easier to compute?

A.2 Duality

Duality is very powerful tool which allows (1) to design efficient (tractable) algorithms to approximate non-convex problems; (2) to build efficient algorithms to convex and non-convex problems with constraints (which are often of a much smaller dimensionality than the original formulations); (3) to formulate necessary and sufficient conditions of optimality for convex and non-convex optimization problems.

Lagrangian

Consider the following constrained (not necessary convex) optimization problem:

$$\begin{aligned} f(x) &\to \min \\ \texttt{s.t.}: g_i(x) &\le 0,\ 1 \le i \le m \\ h_j(x) &= 0,\ 1 \le j \le p \\ x &\in \mathbb{R}^n \end{aligned} \tag{A.3}$$

with the optimal value p^* (which is possibly $-\infty$). Let S be the feasible set of this problem, that is the set of all x for which all the constraints are satisfied.

Compose the so-called Lagrangian function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$:

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x) = f(x) + \lambda^\top g(x) + \mu^\top h(x), \qquad \lambda \ge 0 \tag{A.4}$$

which is a weighted combination of the objective and the constraints. Lagrange multipliers, λ and μ, can be viewed as penalties for violation of inequality and equality constraints.

The Lagrangian function (A.4) allows us to formulate the constrained optimization, Eq. (A.3), as a min-max (also called saddle point) optimization problem:

$$p^* = \min_{x \in S \subseteq \mathbb{R}^n} \max_{\lambda \ge 0, \mu} \mathcal{L}(x, \lambda, \mu) \tag{A.5}$$

where the optimum of Eq. (A.3) is achieved at p_*.

Weak and Strong Duality

Let us consider the saddle point problem (A.5) in greater details. For any feasible point $x \in S \subseteq \mathbb{R}^n$ one has $f(x) \ge \mathcal{L}(x, \lambda, \mu)$, $\lambda \ge 0$. Thus

$$L(\lambda, \mu) = \min_{x \in S} \mathcal{L}(x, \lambda, \mu) \le \min_{x \in S} f(x) = p^* \Rightarrow \max_{\lambda \ge 0, \mu} \underbrace{\min_{x \in S} \mathcal{L}(x, \lambda, \mu)}_{L(\lambda, \mu)} \le p^* = \min_{x \in S} \max_{\lambda \ge 0, \mu} \mathcal{L}(x, \lambda, \mu),$$

where $L(\lambda,\mu) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x,\lambda,\mu) = \inf_{x\in\mathbb{R}^n} \left\{f(x) + \lambda^\top g(x) + \mu^\top h(x)\right\}$ is called the Lagrange dual function. One can restate it as

$$d^* = \max_{\lambda\geq 0,\mu}\min_{x\in S} \mathcal{L}(x,\lambda,\mu) \leq \min_{x\in S}\max_{\lambda\geq 0,\mu} \mathcal{L}(x,\lambda,\mu) = p^*$$

The original optimization, $\min_{x\in S} f(x) = \min_{x\in S}\max_{\lambda\geq 0,\mu} \mathcal{L}(x,\lambda,\mu)$, is called Lagrange primal optimization, while , $\max_{\lambda\geq 0,\mu} L(\lambda,\mu) = \max_{\lambda\geq 0,\mu}\min_{x\in S} \mathcal{L}(x,\lambda,\mu)$, is called the Lagrange dual optimization.

Note that, $\max_{\lambda\geq 0,\mu}\min_{x\in S} \mathcal{L}(x,\lambda,\mu) = \max_{\lambda\geq 0,\mu}\min_{x\in\mathbb{R}^n} \mathcal{L}(x,\lambda,\mu)$, regardless of what S is. This is because $\hat{x}\notin S$ one has $\max_{\lambda\geq 0,\mu}\mathcal{L}(\hat{x},\lambda,\mu) = +\infty$, thus allowing us to perform unconstrained minimization of $\mathcal{L}(x,\lambda,\mu)$ over x much more efficiently.

Let us describe a number of important features of the dual optimization:

1. *Concavity of the dual function.* The dual function $L(\lambda,\mu)$ is always concave. Indeed for $(\bar{\lambda},\bar{\mu}) = \theta(\lambda_1,\mu_1) + (1-\theta)(\lambda_2,\mu_2)$ one has

   $$\begin{aligned} L(\bar{\lambda},\bar{\mu}) &= \min_x \mathcal{L}(x,\bar{\lambda},\bar{\mu}) = \min_x\{\theta\mathcal{L}(x,\lambda_1,\mu_1) + (1-\theta)\mathcal{L}(x,\lambda_2,\mu_2)\} \\ &\geq \theta\min_x \mathcal{L}(x,\lambda_1,\mu_1) + (1-\theta)\min_x \mathcal{L}(x,\lambda_2,\mu_2) = \theta L(\lambda_1,\mu_1) + (1-\theta)L(\lambda_2,\mu_2) \end{aligned}$$

   The dual (maximization) problem $\max_{\lambda\geq 0,\mu} L(\lambda,\mu)$ is equivalent to the minimization of the convex function $-L(\lambda,\mu)$ over the convex set $\lambda\geq 0$.

2. *Lower bound property.* $L(\lambda,\mu)\leq p^*$ for any $\lambda\geq 0$.

3. *Weak duality:* For any optimization problem $d^*\leq p^*$. Indeed, for any feasible (x,λ,μ) we have $f(x)\geq\mathcal{L}(x,\lambda,\mu)\geq L(\lambda,\mu)$, thus $p^* = \min_{x\in\mathbb{R}^n} f(x)\geq\max_{\lambda\geq 0,\mu} L(\lambda,\mu) = d^*$.

4. *Strong duality:* We say that strong duality holds if $p^* = d^*$. Convexity of the objective function and convexity of the feasible set S is neither sufficient nor necessary condition for strong duality (see the example following).

Example A.2.1. Convexity alone is not sufficient for the strong duality. Find the dual problem and the duality gap $p^* - d^*$ for the following optimization

$$\begin{aligned} \exp(-x) &\to \min_{y>0,x} \\ \texttt{s.t.}: x^2/y &\leq 0. \end{aligned}$$

The optimal problem is $p^* = 1$, which is achieved at $x = 0$ and any positive y. The dual problem is

$$L(\lambda) = \inf_{y>0,x}(\exp(-x) + \lambda x^2/y) = 0.$$

That is the dual problem is $\max_{\lambda\geq 0} 0 = 0$, and the duality gap is $p^* - d^* = 1$.

Theorem A.2.2 (Slater (sufficient) conditions)**.** Consider the optimization (A.3) where all the equality constraints are affine and all the inequality constraints and the objective function are convex. The strong duality holds if there exists an x_* such that x_* is strictly feasible, i.e. all constraints are satisfied and the nonlinear constraints are satisfied with strict inequalities.

The Slater conditions imply that the set of optimal solutions of the dual problem, therefore making the conditions sufficient for the strong duality of the optimization.

Optimality Conditions

Another notable feature of the Lagrangian function is due to its role in establishing necessary and sufficient conditions for a triplet (x, λ, μ) to be the solution of the saddle-point optimization (A.5). First, let us formulate necessary conditions of optimality for

$$\begin{aligned} f(x) &\to \min \\ \texttt{s.t.:} g_i(x) &\leq 0,\ 1 \leq i \leq m \\ h_j(x) &= 0,\ 1 \leq j \leq p \\ x &\in S \subseteq \mathbb{R}^n. \end{aligned}$$

According to Eq. (A.5) the optimization is equivalent to

$$\min_{x \in S} \max_{\lambda \geq 0, \mu} \mathcal{L}(x, \lambda, \mu),$$

where the Lagrangian is defined in Eq. (A.4). The following conditions, called Karush-Kuhn-Tucker (KKT) conditions, are necessary for a triplet (x^*, λ^*, μ^*) to become optimal:

1. *Primal feasibility*: $x^* \in S$.
2. *Dual feasibility*: $\lambda^* \geq 0$.
3. *Vanishing gradient:* $\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0$ for smooth functions, and $0 \in \partial \mathcal{L}(x^*, \lambda^*, \mu^*)$ for non-smooth functions. Indeed for the optimal (λ^*, μ^*), $\mathcal{L}$ should attain its minimum at x^*.
4. *Complementary slackness conditions:* $\lambda_i^* g_i(x^*) = 0$. Otherwise if $g_i(x^*) < 0$ and $\lambda_i^* > 0$ one can reduce the Lagrange multiplier and increase the objective.

Note, that the KKT conditions generalize (the finite dimensional version of) the Euler-Lagrange conditions introduced in the variational calculus. Let us now investigate when the conditions are sufficient.

The KKT conditions are sufficient if the problem allows the strong duality, for which (as we saw above) the Slatter conditions are sufficient. Indeed, assume that the strong duality holds and a point (x^*, λ^*, μ^*) satisfies the KKT conditions. Then

$$L(\lambda^*, \mu^*) = f(x^*) + L(x^*)^\top \lambda^* + h(x^*)^\top \mu^* = f(x^*) \tag{A.6}$$

where the first equality holds because of the problem stationarity, and the second conditions holds because of the complementary slackness.

Example A.2.3. Find a duality gap and solve the dual problem for the following minimization

$$\min\left((x_1-3)^2 + (x_2-2)^2\right) \quad \begin{matrix} x_1 + 2x_2 = 4 \\ x_1^2 + x_2^2 \le 5 \end{matrix}$$

Note, that the problem is convex and the Slater's conditions are satisfied, therefore the minimum is unique and there is no duality gap. The Lagrangian is

$$\mathcal{L}(x, \lambda, \mu) = (x_1-3)^2 + (x_2-2)^2 + \mu(x_1 + 2x_2 - 4) + \lambda(x_1^2 + x_2^2 - 5), \quad \lambda \ge 0.$$

The dual problem becomes

$$L(\lambda, \mu) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda, \mu).$$

The KKT conditions are

$$\nabla \mathcal{L} = \begin{pmatrix} 2(x_1-3) + \mu + 2\lambda x_1 \\ 2(x_2-2) + 2\mu + 2\lambda x_2 \end{pmatrix} = 0$$

Therefore, $(1+\lambda)(2x_1 - x_2) = 4$, and using the primal feasibility constraint one derives, $x_1 = \frac{12+4\lambda}{5(1+\lambda)}$, $x_2 = \frac{4+8\lambda}{5(1+\lambda)}$. The dual problem becomes

$$L(\lambda) = 5 - \frac{9\lambda}{5} - \frac{16}{5(1+\lambda)} \to \max_{\lambda \ge 0}$$

Looking for a stationary point of $L(\lambda)$, we arrive at $\lambda = 1/3$ and $\lambda = -7/3$. However, given that $\lambda^* \ge 0$, we get $\lambda^* = 1/3$. Finally, the saddle point is $(x_1^*, x_2^*, \lambda^*, \mu^*) = (2, 1, 2/3, 1/3)$.

Example A.2.4. For the primal problem

$$\begin{aligned} 3x + 7y + z &\to \min \\ \texttt{s.t.}: x + 5y &= 2 \\ x + y &\ge 3 \\ z &\ge 0 \end{aligned}$$

find the dual problem, the optimal values of the primal and dual objectives, as well as optimal solutions for the primal variables and for the dual variables. Describe all the steps in details.

Solution:

1. Note, that the problem is equivalent to

$$\begin{aligned} 3x + 7y &\to \min \\ \texttt{s.t.}: x + 5y &= 2 \\ x + y &\geq 3 \end{aligned}$$

   as x, y are independent of z, and the objective attains its minimum at $z = 0$.

2. Introduce the Lagrangian:

$$\mathcal{L}(x, y, \mu, \lambda) = 3x + 7y + \mu(2 - x - 5y) + \lambda(3 - x - y)$$

3. State the KKT conditions for $\nabla\mathcal{L}(x, y, \mu, \lambda)$:

$$\begin{aligned} \frac{d}{dx}\mathcal{L}(x, y, \mu, \lambda) &= 3 - \mu - \lambda = 0 \\ \frac{d}{dy}\mathcal{L}(x, y, \mu, \lambda) &= 7 - 5\mu - \lambda = 0, \end{aligned}$$

   therefore resulting in $\mu = 1$, and $\lambda = 2$. One observes that the Lagrange multipliers are feasible, meaning that there exists at least one point on the intersection of the equality and inequality constraints.

4. The complimentary slackness condition (for the inequality) is

$$\lambda(3 - x - y) = 0.$$

   Since $\lambda = 2$, the respective inequality constraint is active $x + y = 3$.

5. Using the primal feasibility one derives:

$$x + 5y = 2 \qquad \text{and} \qquad x + y = 3,$$

   resulting in $y = -0.25$ and $x = 3.25$.

6. Optimal values of the primal variables are $(x, y, z) = (3.25, -0.25, 0)$.

Dual problem.

1. The Lagrangian function is

   $$\mathcal{L}(x,y,\mu,\lambda) = 3x+7y+\mu(2-x-5y)+\lambda(3-x-y) = 2\mu+3\lambda+x(3-\mu-\lambda)+y(7-5\mu-\lambda)$$

   Dual objective:

   $$L(\lambda,\mu) = \inf_{x,y}\mathcal{L}(x,y,\mu,\lambda) = \begin{cases} 2\mu+3\lambda, & \text{if } 3-\mu-\lambda=0 \text{ and } 7-5\mu-\lambda=0 \\ -\infty, & \text{otherwise} \end{cases}$$

2. Thus, the dual problem is

   $$\begin{aligned} 2\mu+3\lambda &\to \max \\ \texttt{s.t.:}\ 3-\lambda-\mu &= 0 \\ 7-5\mu-\lambda &= 0 \end{aligned}$$

3. The duality gap is 0 as this problem is linear (Slater's condition is satisfied by the definition).

Exercise A.7. For the primal optimization problems stated below find the dual problem, the optimal values of the primal and dual objectives, as well as optimal solutions for the primal variables and for the dual variables. Describe all the steps in details.

1. min $4x+5y+7z$, $\texttt{s.t.:}$ $2x+7y+5z+d=9$, and $x,y,z,d\geq 0$. [Hint: try to drop an inequality constraint, find the optimal value and check after finding the optimal solution if the dropped inequality is satisfied.]

2. min $\{(x_1-5/2)^2+7x_2^2-x_3^2\}$, $\texttt{s.t.:}$ $x_1^2-x_2\leq 0$, and $x_3^2+x_2\leq 4$

Examples of Duality

Example A.2.5 (Duality and Legendre-Fenchel Transform)**.** Let us discuss the relation between transformation from the Lagrange function to the dual (Lagrange) function and the Legendge-Fenchel (LF) transform (or a conjugate function),

$$f^*(y) = \sup_{x\in\mathbb{R}^n}(y^\top x - f(x)),$$

introduced in Variational Calculus Section 6 of the course. One of the principal conclusions of the LF analysis is $f(x)\geq f^{**}(x)$. The inequality is directly linked to the statement of duality, specifically to the fact that dual optimization low bounds the primal one. To illustrate the relationship between the maximization of f^{**} and the dual problem consider

$$\begin{aligned} f(x) &\to \min \\ \texttt{s.t.:} x &= b \end{aligned}$$

where b is a parameter. Then

$$\min_x \max_\mu \{f(x) + \mu^\top (b - x)\} \le \max_\mu \min_x \{f(x) + \mu(b - x)\}$$
$$= \max_\mu \{-\mu b - \max_x (\mu x - f(x))\} = \max_\mu \{-\mu b - f^*(\mu)\} = f^{**}(-b)$$

Minimizing the expression over all $b \in \mathbb{R}^n$ one arrives at $\min_{x\in\mathbb{R}^n} f(x) \ge \min_{x\in\mathbb{R}^n} f^{**}(x)$.

Example A.2.6. [Duality in Linear Programming (LP)] Consider the following problem:

$$c^\top x \to \min$$
$$\texttt{s.t.:} Ax \le b$$

We define Lagrangian $\mathcal{L}(x, \lambda) = c^\top x + \lambda^\top (Ax - b)$, $\lambda \ge 0$, and arrives at the following dual objective

$$L(\lambda) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x, \lambda)$$
$$= \inf_{x\in\mathbb{R}^n} \left\{x^\top (c + A^\top \lambda) - b^\top \lambda\right\} = \begin{cases} -b^\top \lambda, & \text{if } c + A^\top \lambda = 0 \\ -\infty, & \text{otherwise} \end{cases}$$

The resulting dual optimization is

$$L(\lambda) = -b^\top \lambda \to \max_{c+A^\top\lambda=0,\ \lambda\ge 0}$$

Example A.2.7 (Non-convex problems with strong duality)**.** Consider the following quadratic minimization:

$$x^\top Ax + 2b^\top x \to \min$$
$$\texttt{s.t.:} x^\top x \le 1$$

where $A \not\succeq 0$. Its dual objective is:

$$L(\lambda) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x, \lambda)$$
$$= \inf_{x\in\mathbb{R}^n} \left\{x^\top (A + \lambda I)x - 2b^\top x - \lambda\right\} = \begin{cases} -\infty, & A + \lambda I \not\succeq 0 \\ -\infty, & A + \lambda I \succeq 0, b \in Im(A + \lambda I) \\ -b^\top (A + \lambda I)^+ b - \lambda, & \text{otherwise} \end{cases}$$

The resulting dual optimization is

$$-b^\top (A + \lambda I)^+ b - \lambda \to \max$$
$$\texttt{s.t.:} A + \lambda I \succeq 0$$
$$b \in Im(A + \lambda I)$$

Let us restate the optimization in a convex form by introducing an extra variable t

$$\begin{aligned} &-t-\lambda \to \max \\ &\texttt{s.t.:} t \geq b^{\top}(A+\lambda I)^{+}b \\ &\qquad A+\lambda I \succeq 0 \\ &\qquad b \in Im(A+\lambda I) \end{aligned}$$

Finally one arrives at

$$\begin{aligned} &-t-\lambda \to \max \\ &\texttt{s.t.:} \begin{pmatrix} A+\lambda I & b \\ b^{\top} & t \end{pmatrix} \succeq 0 \end{aligned}$$

Example A.2.8 (Dual to binary Quadratic Programming (QP))**.** Consider the following binary quadratic optimization

$$\begin{aligned} &x^{\top}Ax \to \max \\ &\texttt{s.t.:} x_i^2 = 1,\ 1 \leq i \leq n \end{aligned}$$

with $A \succeq 0$. The dual optimization is

$$\min_{x\in\mathbb{R}^n}\left\{-x^{\top}Ax+\sum_{i=1}^{n}\mu_i(x_i^2-1)\right\} = \min_{x\in\mathbb{R}^n}\left\{x^{\top}(Diag(\mu)-A)x-\sum_{i=1}^{n}\mu_i\right\} \to \max_{\mu}$$

that is

$$\begin{aligned} &\sum_{i=1}^{n}\mu_i \to \min \\ &\texttt{s.t.:} Diag(\mu) \succeq A \end{aligned} \tag{A.7}$$

Note that the optimization (A.7 is convex and it provides a non-trivial lower bound to the primal optimization problem. The low bound is called *Semi-Definite Programming* (SDP) relaxation.).

Example A.2.9. Show that $\min_x \lambda^T x\Big|_{\|x\|_p\leq 1} = -\|\lambda\|_{p/(p-1)}$, where $x \in \mathbb{R}^d$, $p \geq 0$ and, $\|x\|_p := (|x_1|^p + \cdots + |x_d|^p)^{1/p}$, is the p-norm of x.

Solution: The dual formulation of the problem is

$$\min_x \max_{\mu} \left(\lambda^T x + \mu\left(\|x\|_p - 1\right)\right)_{\mu\geq 0}. \tag{A.8}$$

The original formulation is convex and Slater's condition is obviously satisfied (any x which is sufficiently small in the p-norm is feasible), therefore the strong duality holds and we can reverse the order of optimizations in Eq. (A.8)

$$\max_{\mu} \min_{x} \left(\lambda^T x + \mu \left(\|x\|_p - 1\right)\right)_{\mu \geq 0}. \tag{A.9}$$

Next let us write the KKT conditions. Stationary point condition over the primal variable for the objective in Eq. (A.9) is

$$\forall i = 1, \cdots, d: \ \lambda_i + \mu x_i^* \frac{|x_i^*|^{p-2}}{\left((x_1^*)^p + \cdots + (x_d^*)^p\right)^{1-1/p}} = 0. \tag{A.10}$$

The complementary slackness condition is

$$\mu \left(\|x^*\|_p - 1\right) = 1. \tag{A.11}$$

Assume that $\lambda \neq 0$ (if otherwise the result is trivially zero), then $\mu \neq 0$ according to Eq. (A.10), $\|x^*\|_p = 1$, and Eq. (A.10) becomes, $\forall i = 1, \cdots, d: \ \lambda_i = -\mu x_i^* |x_i^*|^{p-2}$. Combining the two equations we derive

$$\mu = -\left(\sum_i |\lambda_i|^{p/(p-1)}\right)^{(p-1)/p} = -\|\lambda\|_{(p-1)/p}, \tag{A.12}$$

$$\lambda^T x^* = -\mu \sum_i |x_i^*|^p = -\mu, \tag{A.13}$$

therefore proving the relation after substitution the optimal values back into the objective.

Exercise A.8. For the quadratic constraint optimization problem

$$\min_{x \in \mathbb{R}^d} -\frac{1}{2} x^\top L x + b^\top x \tag{A.14}$$

$$\text{s.t.: } \|x\|_\infty \leq 1, \tag{A.15}$$

- (a) Describe conditions on L and b guaranteeing convexity of Eq. (A.14).
- (b) Find the dual of Eq. (A.14), restating the l_∞-constraint as a convex quadratic constraint. Is the duality gap zero at $L \preceq 0$?
- (c) Show that if $bb^\top \succeq \varepsilon L \succeq 0$ for some $\varepsilon > 0$, then $L = cbb^\top$, where c is a constant.
- (d) Assuming that conditions in (c) are satisfied, solve Eq. (A.14) analytically. (Hint: Transition to a scalar variable and show that the problem reduces to a one-dimensional quadratic, concave optimization over a bounded domain.)

Conic Duality (additional material)

Standard formulation of the conic optimization is:

$$\begin{aligned} c^\top x &\to \min_x \\ \texttt{s.t.}: Ax &= b \\ x &\in \mathcal{K} \end{aligned} \tag{A.16}$$

where $\mathcal{K}$ is a proper cone, i.e. a set which satisfies

1. $\mathcal{K}$ is a convex cone, that is for any $x, y \in \mathcal{K}$ one has $\alpha x + \beta y \in \mathcal{K}$, $\alpha, \beta \geq 0$;
2. $\mathcal{K}$ is closed;
3. $\mathcal{K}$ is solid, meaning it has nonempty interior;
4. $\mathcal{K}$ is pointed, meaning if $x \in \mathcal{K}$, and $-x \in \mathcal{K}$ then $x = 0$.

Conic optimization problems are important in optimization. In Example A.2.8 you already see the (dual to the binary quadratic optimization) problem which is a conic optimization problem over the cone of positive semi-definite matrices.

$\mathcal{K}^*$ defines a dual cone of $\mathcal{K}$ $\mathcal{K}^* = \{c : \ c^\top \langle c, x\rangle \geq 0, x \in \mathcal{K}\}$.

Exercise A.9. Show that the following sets are self-dual cones (that is, $\mathcal{K}^* = \mathcal{K}$).

1. Set of positive semi-define matrices, $\mathbb{S}^n_+$;
2. Positive orthant, $\mathbb{R}^n_+$;
3. Second-order cone, $Q^n = \{(x, t) \in \mathbb{R}^n_+ : \ t \geq \|x\|_2\}$

Note, that in the case of the semi-definite matrices $c^\top x = \sum_{i,j=1}^n c_{ij} x_{ij}$ (e.g. Hadamard product of matrices). The Lagrangian to the Problem A.16 is given by

$$\mathcal{L} = c^\top x + \mu^\top (b - Ax) - \lambda x$$

where the last term stands for $x \in \mathcal{K}$. From the definition of the dual cone one derives

$$\max_{\lambda \in \mathcal{K}^*} -\lambda^\top x = \begin{cases} 0, & x \in \mathcal{K} \\ +\infty, & x \notin \mathcal{K} \end{cases}$$

Therefore

$$\begin{aligned} p^* = \min_{x \in \mathcal{K}} \max_{\lambda \in \mathcal{K}^*, \mu} \mathcal{L}(x, \lambda, \mu) \geq \\ d^* = \max_{\lambda \in \mathcal{K}^*, \mu} \min_{x \in \mathcal{K}} \mathcal{L}(x, \lambda, \mu) \end{aligned}$$

And the dual problem is

$$L(\lambda,\mu)=\min_{x\in\mathcal{K}}\{c^\top x+\mu^\top(b-Ax)-\lambda^\top x\}=\begin{cases}\mu^\top b, & \text{if } c-A^\top\mu-\lambda=0\\ -\infty, & \text{otherwise}\end{cases}$$

And finally

$$\begin{aligned} d^* &= \max \mu^\top b\\ \texttt{s.t.:} &= c-A^\top\mu-\lambda=0\\ &\lambda\in\mathcal{K}^* \end{aligned}$$

Finally, eliminating λ one has

$$\begin{aligned} &\mu^\top b\to\max\\ &\texttt{s.t.:}c-A^\top y\in\mathcal{K}^*. \end{aligned}$$

Exercise A.10. Find a dual problem (see Example A.2.8) to

$$\begin{aligned} 1^\top\mu=\sum_{i=1}^n\mu_i &\to \min\\ \texttt{s.t.:}Diag(\mu) &\succeq A. \end{aligned}$$

Ensure, that your dual problem is equivalent to

$$\begin{aligned} \langle A,X\rangle &\to\max\\ \texttt{s.t.:}X &\in\mathbb{S}^n_+\\ X_{ii} &= 1\ \forall i \end{aligned}$$

In the remainder of the Section we will study iterative algorithms to solve the optimization problems discussed so far. It will be convenient to think about iterations in terms of "discrete (algorithmic) time", and also consider the "continuous time" limit when changes in the values per iteration is sufficiently small and the number of iterations is sufficiently large. In the continuous time analysis of the algorithms we utilize the language of differential equations, as it helps both for intuition (familiar from first semester studies of the differential equations) and also for analysis. However, to reach some of the rigorous conclusions we may also get back to the original, discrete, language.

A.3 Unconstrained First-Order Convex Minimization

In this lecture, we will consider an unconstrained convex minimization problem

$$f(x) \to \min_{x\in\mathbb{R}^n},$$

and focus on the first-order optimization methods. That is we assume that the objective function, as well as the gradient of the objective function, can both be evaluated efficiently. Note that the first order methods described in this Section are most popular methods/algorithms currently in use to resolve majority of practical machine learning, data science and more generally applied mathematics problems.

We assume that function f is smooth, that is

$$\forall x, y: \ \|\nabla f(x) - \nabla f(y)\|_* \le \beta\|x - y\|, \tag{A.17}$$

for some positive constant β. Choosing the ℓ_2 norm for $\|\cdot\| = \|\cdot\|_2$, one derives

$$f(y) \le f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^n.$$

To simplify description we will thus omit in the following "w.r.t. to norm $\|\cdot\|$" when discussing the ℓ_2 norm.

Smooth Optimization

Gradient Descent. Gradient Descent (GD) is the simplest and arguably most popular method/algorithm for solving convex (and non-convex) optimization problems. Iteration of the GD algorithm is

$$x_{k+1} = x_k - \eta_k \nabla f(x_k) = \arg\min_x \underbrace{\left\{ f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{1}{2\eta_k}\|x - x_k\|_2^2 \right\}}_{h_{\eta_k}(x)}, \qquad \eta_k \le 1/\beta$$

where we assume that f is β smooth with respect to ℓ_2 norm. If $\eta_k \leq 1/\beta$, each step of the GD becomes equivalent to minimization of the convex quadratic upper bound $h_{\eta_k}(x)$ of $f(x)$.

Definition A.3.1. Function $f : \mathbb{R}^n \to \mathbb{R}$ is β-smooth w.r.t. to a norm $\|\cdot\|$ if

$$\|\nabla f(x) - \nabla f(y)\|_* \leq \|x - y\| \qquad \forall x, y.$$

If $\|\cdot\| = \|\cdot\|_2$, we call the function β-smooth.

Theorem A.3.2. Assume that a function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and β-smooth. Then repeating the GD step k times/iterations with a fixed step-size, $\eta \leq 1/\beta$, results in $f(x_k)$ which satisfies:

$$f(x_k) - f(x^*) \leq \frac{\|x_1 - x^*\|_2^2}{2\eta k}, \qquad \eta \leq 1/\beta, \tag{A.18}$$

where x^* is the optimal solution.

We will provide the continuous time proof of the Theorem, as well as its discrete time version, where the former will rely on the notion of the Lyapunov function.

Definition A.3.3. Lyapunov function, $V(x(t))$, of the differential equation, $\dot{x}(t) = f(x(t))$, is a function that

1. decreases monotonically along (discrete or continuous time) trajectory, $\dot{V}(x(t)) < 0$.
2. converges to zero at $t \to \infty$, i.e. $V(x(\infty)) = 0$, where $x^* = x(\infty)$.

From now on, we will use capitalized notation, $X(t)$, for the continuous time version of $(x_k | k = 1, \cdots)$.

Proof of Theorem A.3.2: Continuous time. The GD algorithm can be viewed as a discretization of the first-order differential equation:

$$\dot{X}(t) = -\nabla f(X(t)).$$

Introduce the following Lyapunov's function for this ODE, $V(X(t)) = \|X(t) - x^*\|_2^2/2$. Then

$$\frac{d}{dt}V(t) = (X(t) - x^*)^\top \dot{X}(t) = -\nabla f(X(t))^\top (X(t) - x^*) \leq -(f(X(t)) - f^*), \tag{A.19}$$

where the last inequality is due to the convexity of f. Integrating Eq. (A.19 over time, one derives

$$V(X(t)) - V(X(0)) \leq tf^* - \int_0^t f(X(t))dt$$

Utilizing (a) Jensen's inequality

$$f\left(\frac{1}{t}\int_0^t X(\tau)d\tau\right) \leq \frac{1}{t}\int_0^t f(X(\tau))d\tau,$$

which is valid for all convex functions, and (b) non-negativity of $V(t)$ one derives

$$f\left(\frac{1}{t}\int_0^t X(\tau)d\tau\right) - f^* \leq \frac{1}{t}\int_0^t X(\tau)d\tau - f^* \leq \frac{V(X(0))}{t}.$$

The prove is complete after setting, $t \approx k/\beta$, and recalling that f is smooth. □

Proof of Theorem A.3.2: Discrete time. Condition of smoothness applied to, $y = x - \eta\nabla f(x)$, results in

$$\begin{aligned}
f(y) &\leq f(x) + \nabla f(x)^\top (y-x) + \frac{\beta_2}{2}\|y-x\|_2^2 \\
&= f(x) + \nabla f(x)^\top (x - \eta\nabla f(x) - x) + \frac{1}{2}\beta_2^2\|x - \eta\nabla f(x) - x\|_2^2 \\
&= f(x) - \eta\|\nabla f(x)\|_2^2 + \frac{\beta_2}{2}\|\nabla f(x)\|_2^2 \\
&= f(x) - \left(1 - \frac{\beta_2\eta}{2}\right)\eta\|\nabla f(x)\|_2^2.
\end{aligned}$$

As $\eta \leq 1/\beta$, one derives, $1 - \beta\eta/2 \leq -1/2$, and

$$f(y) \leq f(x) - \frac{\eta}{2}\|\nabla f(x)\|_2^2. \tag{A.20}$$

Note, that Eq. A.20 does not require convexity of the function, however if the function is convex one derives

$$f(x^*) \geq f(x) + \nabla f(x)^\top (x^* - x),$$

by choosing $y = x^*$. Plugging the last inequality into the smoothness inequality, one derives for $y = x - \eta\nabla f(x)$:

$$\begin{aligned}
f(y) - f(x^*) &\leq \nabla f(x)^\top (x - x^*) - \frac{\eta}{2}\|\nabla f(x)\|_2^2 \\
&= \frac{1}{2\eta}\left\{\|x - x^*\|_2^2 - \|x - \eta\nabla f(x) - x^*\|_2^2\right\} \\
&= \frac{1}{2\eta}\left\{\|x - x^*\|_2^2 - \|y - x^*\|_2^2\right\} \\
\sum_{j\leq k}(f(x_j) - f(x^*)) &\leq \frac{1}{2\eta}\sum_{j\leq k}\left(\|x_j - x^*\|_2^2 - \|x_{j+1} - x^*\|_2^2\right) \\
&= \frac{1}{2\eta}\left(\|x_1 - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2\right) \\
&\leq \frac{R_2^2}{2\eta} = \frac{\beta_2 R_2^2}{2},
\end{aligned}$$

where $R_2^2 \geq \|x_1 - x^*\|_2^2$ and the step-size $\eta = 1/\beta$. Finally

$$\min_j f(x_j) - f(x^*) \leq f(\bar{x}) - f(x^*) \leq \frac{\beta R_2^2}{2},$$

where $\bar{x} = \sum_{j \leq k} x_j / k$. □

One obviously would like to choose the step size in GD which results in the fastest convergence. However, this problem – of choosing best, or simply good step size – is hard and remains open. The statement also means that finding a good stopping criterion for the iterations is hard as well. Here are practical/empirical strategies for choosing the step size in GD:

- *Exact line search.* Choose η_k so that

$$\eta_k = \arg\min_{\eta} \{f(x_k - \eta \nabla f(x_k))\}$$

- *Backtracking line search.* Choose the step-size η_k so that:

$$f(x_k - \eta_k \nabla f(x_k)) \leq f(x_k) - \frac{\eta_k}{2} \|\nabla f(x_k)\|_2^2$$

  As the difference between the right-hand side and the left-hand size of the inequality above is monotone in η_k, one can start with some η and then update, $\eta \to b\eta$, $0 < b < 1$.

- *Polyak's step-size rule.* If the optimal value f^* of the function is known, one can suggest a better step-size policy. Minimization of the right-hand side of:

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\|_2^2 - 2\eta_k (f(x_k) - f(x^*)) + \eta_k^2 \|g_k\|_2^2 \to \min_{\eta_k},$$

  results in the Polyak's rule, $\eta_k = (f(x_k) - f(x^*))/\|g_k\|_2^2$, which is known to be useful, in particular, for solving an undetermined system of linear equations, $Ax = b$.

Exercise A.11. Recall that GD minimizes the convex quadratic upper bound $h_{\eta_k}(x)$ of $f(x)$. Consider a modified GD, where the step size is, $\eta = (2+\varepsilon)/\beta$, with ε chosen positive. (Notice that the step size used in the conditions of the Theorem A.3.2 was $\eta \leq 1/\beta$.) Derive modified version of Eq. (A.18). Can one find a quadratic convex function for which the modified algorithm fails to converge?

Exercise A.12. Consider minimization of the following (non-convex) function f:

$$f(x) \to \min$$
$$\texttt{s.t.:} \|x - x^*\| \leq \varepsilon,$$
$$x \in \mathbb{R}^n$$

where x^* is a global and unique minimum of the β-smooth function f. Moreover, let

$$\forall x \in \mathbb{R}^n : \frac{1}{2}\|\nabla f(x)\|_2^2 \geq \mu(f(x) - f(x^*)).$$

Is it true, that for some small $\varepsilon > 0$ the GD with a step-size $\eta_k = 1/\beta$ converges to the optimum? How ε depends on β and μ?

Exercise A.13 (not graded - difficult)**.** In many optimization problems, it is often the case that exact value of the gradient is polluted, i.e. only its noise version is observed. In this case one may consider the following "inexact oracle" optimization: $f(x) \to \min, x \in R^n$, assuming that for any x one can compute $\hat{f}(x)$ and $\hat{\nabla} f(x)$ so that

$$\forall x : \ |f(x) - \hat{f}(x)| \leq \delta, \ \text{ and } \ \|\nabla f(x) - \nabla \hat{f}(x)\|_2 \leq \varepsilon,$$

and seek for an algorithm to solve it. Propose and analyze modification of GD solving the "inexact oracle" optimization?

Gradient Descent in ℓ_p. GD in ℓ_p norm

$$x_{k+1} = \arg\min_{x \in S \subset \mathbb{R}^n} \left\{ f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{1}{2\eta_k}\|x - x_k\|_p^2 \right\},$$

where $\eta_k \leq 1/\beta_p$, $\beta_p \geq \sup_x \|g(x)\|_p$, with a properly chosen p can converge much faster than in ℓ_2. GD in ℓ_1 is particularly popular.

Exercise A.14. Restate and prove discrete time version of the Theorem A.3.2 for GD in ℓ_p norm. (Hint: Consider the following Lyapunov function: $\|x - x^*\|_p^2$.)

Gradient Descent for Strongly Convex, Smooth Functions.

Theorem A.3.4. GD for a strongly convex function f and a fixed step-size policy

$$x_{k+1} = x_k - \eta \nabla f(x_k), \quad \eta = 1/\beta$$

converges to the optimal solution as

$$f(x_{k+1}) - f(x^*) \leq c^k (f(x_1) - f(x^*)),$$

where $c \leq 1 - \mu/\beta$.

Exercise A.15. Extend proof of the Theorem A.3.2 to Theorem A.3.4.

Fast Gradient Descent. GD is simple and efficient in practice. However, it may also be slow if the gradient is small. It may also oscillate about the point of optimality if the gradient is pointed in a direction with a small projection to the optimal direction (pointing at the optimum). The following two modifications of the GD algorithm were introduced to cure the problems

(1964) Polyak's heavy-ball rule:

$$x_{k+1} = x_k + \eta_k \nabla f(x_k) + \mu_k (x_k - x_{k-1}) \tag{A.21}$$

(1983) Nesterov Fast Gradient Method (FGM):

$$x_{k+1} = x_k + \eta_k \nabla f(x_k + \mu(x_k - x_{k-1})) + \mu_k (x_k - x_{k-1}). \tag{A.22}$$

The last term in Eqs. (A.21,A.22) is called "momentum" or "inertia" term to emphasize relation to respective phenomena in classical mechanics. The inertia terms, added to the original GD term, which may be associated with "damping" or "friction", aims to force the hypothetical "ball" rolling towards optimum faster. In spite of their seemingly minor difference, convergence rate of FGM and of the heavy-ball method differ rather dramatically, as the heavy ball can lead to an overshoot (not enough "friction").

Exercise A.16. Construct a convex function f with a piece-wise linear gradient such that the heavy ball algorithm (A.21) with some fixed μ and ηfails to converge.

Consider a slightly modified (less general, two-step recurrence) version of the FGM (A.22):

$$x_k = y_{k-1} - \eta \nabla f(y_{k-1}), \qquad y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}), \tag{A.23}$$

which can be re-stated in continuous time as follows

$$\ddot{X}(t) + \frac{3}{t}\dot{X}(t) + \nabla f(X) = 0. \tag{A.24}$$

Indeed, assuming $t \approx k\sqrt{\eta}$ and re-scaling one derives from Eq. (A.23)

$$\frac{x_{k+1} - x_k}{\sqrt{\eta}} = \frac{k-1}{k+2}\frac{x_k - x_{k-1}}{\sqrt{\eta}} - \sqrt{\eta}\nabla f(y_k). \tag{A.25}$$

Let $x_k \approx X(k\sqrt{\eta})$, then

$$X(t) \approx x_{t/\sqrt{\eta}} = x_k, \qquad X(t+\sqrt{\eta}) \approx x_{(t+\sqrt{\eta})/\sqrt{\eta}} = x_{k+1}$$

and utilizing the Taylor expansion

$$\frac{x_{k+1} - x_k}{\sqrt{\eta}} = \dot{X}(t) + \frac{1}{2}\ddot{X}(t)\sqrt{\eta} + o(\sqrt{\eta})$$
$$\frac{x_k - x_{k-1}}{\sqrt{\eta}} = \dot{X}(t) - \frac{1}{2}\ddot{X}(t)\sqrt{\eta} + o(\sqrt{\eta})$$

one arrives at

$$\dot{X}(t)+\frac{1}{2}\ddot{X}(t)+o(\sqrt{\eta}) = (1-3\sqrt{\eta}/t)\left(\dot{X}(t)-\frac{1}{2}\ddot{X}(t)\sqrt{\eta}+o(\sqrt{\eta})\right)-\sqrt{\eta}f(X(t))+o(\sqrt{\eta}) = 0,$$

resulting in Eq. (A.24).

To analyze convergence rate of the FGM (A.24) we introduce the following Lyapunov function:

$$V(X(t)) = t^2(f(X(t))-f^*)+2\|X+t\dot{X}/2-x^*\|_2^2.$$

Time derivative of the Lyapunov function is

$$\dot{V}(X(t)) = 2t(f(X(t))-f^*)+t^2\nabla f(X(t))^\top\dot{X}(t)+4(X(t)+t\dot{X}(t)/2-x^*)^\top(3\ddot{X}(t)/2+t\ddot{X}/2).$$

Given that, $\dot{X}+t\ddot{X}/2 = -t\nabla f(X)/2$, and also utilizing convexity of f one derives

$$\dot{V} = 2t(f(X)-f^*)-4(X-x^*)^\top(t\nabla f(X)/2) = 2t(f(X)-f^*)-2t(X-x^*)^\top\nabla f(X)\le 0.$$

Making use of the monotonicity of V and of the non-negativity of $\|X+t\dot{X}/2-x^*\|$ one finds

$$f(X(t))-f^* \le \frac{V(t)}{t^2}\le\frac{V(0)}{t^2} = \frac{2\|x_0-x^*\|_2^2}{t^2}.$$

Finally, substituting, $t\approx k\sqrt{\eta}$, one derives

$$f(x_k)-f^*\le\frac{2\|x_0-x^*\|_2^2}{\eta k^2},\qquad \eta\le 1/\beta.$$

We have just sketched a proof of the following statement.

Theorem A.3.5. Fast GD for, $f(x)\to\min_{x\in\mathbb{R}^n}$, where $f(x)$ is a β-smooth convex function, with an update rule

$$x_k = y_{k-1}-\eta\nabla f(y_{k-1})),\quad y_k = x_k+\frac{k-1}{k+2}(x_k-x_{k-1})$$

converges to the optimum as

$$f(x_{k+1})-f^*\le\frac{2\|x_0-x^*\|_2^2}{\eta k^2}.$$

As always, turning the continuous time sketch of the proof into the actual (discrete time) proof takes some additional technical efforts.

Exercise A.17. Consider the following differential equation

$$\ddot{X}(t)+\frac{r}{t}\dot{X}(t)+\nabla f(X) = 0,$$

at some positive r. Derive respective discrete time algorithm, analyze its convergence and show that if $r\le 2$, the convergence rate of the algorithm is $O(1/k^2)$.

Exercise A.18. Show that the FGM method, described by Eq. (A.23), transitions to Eq. (A.22) at some η_k.

Non-Smooth Problems

Sub-Gradient Method. We start discussion of the Sub-Gradient (SG) methods with the simplest, and arguably most-popular, SG algorithm:

$$x_{k+1} = x_k - \eta_k g_k, \quad g_k \in \partial F(x_k), \tag{A.26}$$

which is just the original GD with the gradient replaced by the sub-gradient to deal with non-smooth f. Note, however, that it is not proper to call the algorithm (A.26) SG descent because in the process of iterations $f(x_{k+1})$ may become larger than $f(x_k)$. To fix the problem one may keep track of the best point, or substitute the result by an average of the points seen in the iterations so far (a finite horizon portion of the past). For example, one may augment Eq. (A.26) at each k with

$$f_{best}^{(k)} = \min\{f_{best}^{(k-1)}, f(x_k)\}.$$

We assume that SG of $f(x)$ is bounded, that is

$$\forall x: \ \|g(x)\| \leq L, \ g(x) \in \partial f(x).$$

This condition follows, for example, from the Lipshitz condition, $|f(x) - f(y)| \leq L\|x - y\|$, imposed on f. Let x^* be the optimal point of, $f(x) \to \min_{x\in\mathbb{R}^n}$, then

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 = \|x_k - \eta_k g_k - x^*\|_2^2 &= \|x_k - x^*\|_2^2 - 2\eta_k g_k^\top (x_k - x^*) + \eta_k^2 \|g_k\|_2^2 \\ &\leq \|x_k - x^*\|_2^2 - 2\eta_k (f(x_k) - f(x^*)) + \eta_k^2 \|g_k\|_2^2, \end{aligned} \tag{A.27}$$

where the last inequality is due to convexity of f, i.e. $f(x^*) \geq f(x_k) + g_k^\top (x^* - x_k)$. Applying the inequality (A.27) recursively,

$$\|x_{k+1} - x^*\|_2^2 \leq \|x_1 - x^*\|_2^2 - 2\sum_{j\leq k} \eta_j (f(x_j) - f(x^*)) + \sum_{j\leq k} \eta_j^2 \|g_j\|_2^2,$$

one derives,

$$\left(2\sum_{j\leq k} \eta_j\right) (f(_{best}^{(k)}) - f(x^*)) \leq 2\sum_{j\leq k} \eta_j (f(x_j) - f(x^*)) \leq \|x^{(1)} - x^*\|_2^2 + \sum_{j\leq k} \eta_j^2 \|g_j\|_2^2,$$

which becomes

$$f(_{best}^{(k)}) - f(x^*) = \min_{j\leq k} f(x_j) - f^* \leq \frac{\|x_1 - x^*\|_2^2 + L_2^2 \sum_{j\leq k} \eta_j^2}{2\sum_{j\leq k} \eta_j},$$

where we assume that the SG of f are bounded by L_2 in the ℓ_2 norm. Therefore, if $R_2^2 \geq \|x_1 - x^*\|_2^2$, one arrives at

$$\min_{j\leq k} f(x_j) - f^* \leq \min_{\eta} \frac{R_2^2 + L_2^2 \sum_{j\leq k} \eta_j^2}{2\sum_{j\leq k} \eta_j} = \frac{RL}{\sqrt{k}}, \tag{A.28}$$

where the step-size is $\eta_k = R/(L\sqrt{k})$. Note, that the $\sim 1/\sqrt{k}$ scaling in Eq. (A.28 is much worse than the one we got above, $\sim 1/k^2$, for smooth functions. In the following we discuss this result in more details and suggest a number of ways to improve the convergence.

Proximal Gradient Method. In multiple machine learning (and more generally statistics) applications we deal with a function built as a sum over samples. Inspired by this application consider the following *composite optimization*

$$f(x) = g(x) + h(x) \to \min_{x\in\mathbb{R}^n}, \tag{A.29}$$

where we assume that $g : \mathbb{R} \to \mathbb{R}^n$ is a convex and smooth function on $\mathbb{R}^n$, and $h : \mathbb{R} \to \mathbb{R}^n$ is closed, convex and possibly non-smooth function on $\mathbb{R}^n$. One of the most frequently used composite optimization is the Lasso minimization:

$$f(x) = \|Ax - b\|_2^2 + \lambda\|x\|_1 \to \min_{x\in\mathbb{R}^n}. \tag{A.30}$$

Notice that the $\|x\|_1$ term is not smooth at $x = 0$.

Let us now introduce the so-called proximal operator

$$\mathrm{prox}_h(x) = \arg\min_{u\in\mathbb{R}^n} \left(h(u) + \frac{1}{2}\|u - x\|_2^2 \right),$$

which will soon be linked to the composite optimization. Standard examples of the proximal operator/function are

1. $h(x) = I_C(x)$, that is $h(x)$ is an indicator of a convex set C. Then the proximal function is
$$\mathrm{prox}_h(x) = \arg\min_{u\in C} \|x - u\|_2^2$$
is a projection of x on C.

2. $h(x) = \lambda\|x\|_1$, then the proximal function acts as a soft threshold:
$$\mathrm{prox}_h(x)_i = \begin{cases} x_i - \lambda, & x_i \geq \lambda, \\ x_i + \lambda, & x_i \leq -\lambda, \\ 0, & \text{otherwise} \end{cases}$$

The examples suggest using the proximal operator to smooth out non-smooth functions entering respective optimizations. Having this use of the proximal operator in mind we introduce the Proximal Gradient Descent (PGD) algorithm

$$x_{k+1} = \text{prox}_{\eta_k h}(x_k - \eta_k \nabla g(x_k)) = \arg\min_u \left(\frac{1}{2}\|x_k - \eta_k \nabla g(x_k) - u\|_2^2 + \eta_k h(u) \right)$$
$$= \arg\min_u \left(g(x_k) + \nabla g(x_k)^\top (u - x_k) + \frac{1}{2\eta_k}\|u - x_k\|_2^2 + h(u) \right)$$

where $\eta_k \leq \beta$, and g is a β-smooth function in ℓ_2 norm.

Note, that as in the case of the GD algorithm, at each step of the PGD we minimize a convex upper bound of the objective function. We find out that the PGD algorithm has the same convergence rate (measured in the number of iterations) as the GD algorithm.

Finally, we are ready to connect PGD algorithm to the composite optimization (A.29).

Theorem A.3.6. PGD algorithm,

$$x_{k+1} = \text{prox}_h(x_k - \eta \nabla g(x_k)), \quad \eta \leq 1/\beta,$$

with a fixed step size policy converges to the optimal solution f^* of the composite optimization (A.29) according to

$$f(x_{k+1}) - f^* \leq \frac{\|x_0 - x^*\|_2^2}{2\eta k}.$$

Proof of the Theorem (A.3.6) repeats the logic we use to prove Theorem A.3.2 for the GD algorithm. Moreover, one can also accelerate the PGD, similarly to how we have accelerated GD. The accelerated version of the PGD is

$$x_k = \text{prox}_{\eta_k}(y_{k-1} - \eta_k \nabla f(y_{k-1})) \qquad y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}).$$

We naturally arrives at the PGD version of the Theorem A.3.5:

Theorem A.3.7. PGD for a convex optimization

$$f(x) \to \min_{x \in \mathbb{R}^n}$$

with an update rule

$$x_k = \text{prox}_{h\eta}(y_{k-1} - \eta \nabla f(y_{k-1})), \quad y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1})$$

converges as

$$f(x_{k+1}) - f^* \leq \frac{2\|x_0 - x^*\|_2^2}{\eta k^2},$$

for any β-smooth convex function f.

PGD is one possible approach developed to deal with non-smooth objectives. Another sound alternative is discussed next.

Smoothing Out Non-Smooth Objectives

Consider the following min-max optimization

$$\max_{1\le i\le n} f_i(x) \to \min_{x\in\mathbb{R}^n}$$

which is one of the most common non-smooth optimizations. Recall, that a smooth and convex approximation to the maximum function is provided by the soft-max function (A.2) which can then be minimized by the accelerated GD (that has a convergence rate $O(1/\sqrt{\varepsilon})$ in contrast to $1/\varepsilon^2$ for non-smooth functions). Accurate choice of λ (parameter within the soft-max) normally allows to speed up algorithms to $O(1/\varepsilon)$.

A.4 Constrained First-Order Convex Minimization

Projected Gradient Descent

The Projected Gradient Descent (PGD) is

$$\begin{aligned} x_{k+1} &= \Pi_C(x_k - \eta_k \nabla f(x_k)) \\ &= \arg\min_{y\in C}\left(f(x_k) - \nabla f(x_k)^\top (y - x) + \frac{1}{2\eta_k}\|x - y\|_2^2 + I_C(y)\right) \\ &= \mathrm{prox}_{I_C}(x_k - \eta_k \nabla f(x_k)), \end{aligned} \tag{A.31}$$

where Π_C is an Euclidean projection to the convex set C, $\Pi_C(y) = \arg\min_{x\in C}\|x - y\|_2^2$. PGD has the same convergence rate as GD. The proof is similar to the one of the gradient descent taking into account that projection does not lead to an expansion, i.e.

$$\|x_{k+1} - x^*\|_2^2 \le \|x_k - \eta_k \nabla f(x_k) - x^*\|_2^2 \text{ as } x^* \in C.$$

Exercise A.19. (Alternating Projections.) Consider two convex sets $C, D \subseteq R^n$ and pose the question of finding $x \in C \cap D$. One starts from, $x_0 \in C$, and applies PGD

$$y_k = \Pi_C(x_k) \qquad x_{k+1} = \Pi_D(y_k).$$

How many iterations are required to guarantee

$$\max\{\inf_{x\in C}(x_k, x), \inf_{x\in D}(x_k, x)\} \le \varepsilon?$$

Frank-Wolfe Algorithm (Conditional Gradient)

Frank-Wolfe algorithm solves the following optimization problem

$$f(x) \to \min, \quad \texttt{s.t.:} x \in S \tag{A.32}$$

In contrast to the PGD algorithm (A.31) making projection at each iteration, the Frank-Wolfe (FW) algorithm solves the following linear problem on C:

$$y_k = \arg\min_{y \in C} y^\top \nabla f(x_k), \quad x_{k+1} = (1-\gamma_k)x_k + \gamma_k y_k, \quad \gamma_k = 2/(k+1). \tag{A.33}$$

To illustrate, consider the case when C is a simplex:

$$f(x) \to \min \quad \texttt{s.t.}: x \in S = \{x : x \geq 0, x^\top 1 = 1\}.$$

In this case the update y_k of the FW algorithm is a unit vector correspondent to the maximal coordinate of the gradient. Overall time to update x_k is $O(n)$ therefore resulting in a significant acceleration in comparison with the PGD algorithm.

FW algorithm has an edge over other algorithms considered so far because it has a reliable stopping criteria. Indeed, convexity of the objective guarantees that

$$f(y) \geq f(x_k) + \nabla f(x_k)^\top (y - x_k),$$

minimizing both sides of the inequality over $y \in C$ one derives that

$$f^* \geq f(x_k) + \min_{y \in C} \nabla f(x_k)^\top (y - x_k),$$

where f^* is the optimal solution of Eq. (A.32), then leading to

$$\max_{y \in C} \nabla f(x_k)^\top (x_k - y) \geq f(x_k) - f^*. \tag{A.34}$$

The value on the left of the inequality, $\max_{y \in C} \nabla f(x_k)^\top (x_k - y)$, gives us an easy to compute stopping criterion.

The following statement characterizes convergence of the FW algorithm.

Theorem A.4.1. Given that $f(x)$ in Eq. (A.32) is a convex β-smooth function and C is a bounded, convex, compact set, Eq. A.33 converges to the optimal solution, f^*, of Eq. (A.32) as

$$f(x_k) - f^* \leq \frac{2\beta D^2}{k+2},$$

where $D^2 \geq \max_{y,y \in C} \|x - y\|_2^2$.

Proof. Convexity of f means that

$$f(x) \geq f(x_k) + \nabla f(x_k)^\top (x - x_k), \ \forall x \in C.$$

Minimizing both sides of the inequality one derives

$$f(x^*) \geq f(x_k) + \nabla f(x_k)^\top (y_k - x_k).$$

That is $f(x_k) - f(x^*) \leq \nabla f(x_k)^\top (x_k - x^*)$. This inequality, in combination with the second sub-step in the FW algorithm, $x_{k+1} = \gamma_k y_k + (1-\gamma_k)x_k$, results in the following transformations

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq f(x_{k+1}) - f(x_*) \\ &\leq f(x_k) + \nabla f(x_k)^\top (x_{k+1} - x_k) + \frac{\beta}{2}\|x_{k+1} - x_k\|_2^2 - f(x^*) \\ &\leq f(x_k) + \gamma_k \nabla f(x_k)(y_k - x_k) + \frac{\beta\gamma_k^2}{2}\|y_k - x_k\|_2^2 - f(x^*) \\ &\leq f(x_k) - f(x^*) - \gamma_k(f(x_k) - f(x^*)) + \frac{\beta\gamma_k^2}{2}D^2, \end{aligned}$$

and finally

$$f(x_{k+1}) - f^* \leq (1-\gamma_k)(f(x_k) - f^*) + \frac{\beta\gamma_k^2 D^2}{2}.$$

Utilizing the inequality in a chain of inductive relations over k, starting from $k = 1$, one can show that $f(x_k) - f^* \leq 2\beta D^2/(k+2)$. □

The conditional GD is slower than the FGM method in terms of the number of iterations. However, it is often favorable in practice especially when minimizing a convex function over sufficiently simple objects (like the norm-ball or a polytope) as it does not require implementing explicit projection to the constraining set.

Primal-Dual Gradient Algorithm

Consider the following smooth convex optimization problem:

$$\begin{aligned} f(x) &\to \min \\ Ax &= b, x \in \mathbb{R}^n \end{aligned}$$

It is a good practice to work with the equivalent *augmented problem*:

$$\begin{aligned} f(x) + \frac{\rho}{2}\|Ax - b\|_2^2 &\to \min \\ \texttt{s.t.}: Ax &= b \end{aligned}$$

where $\rho > 0$. Let us define *augmented Lagrangian*

$$\mathcal{L}(x, \mu) = f(x) + \mu^\top (Ax - b) + \frac{\rho}{2}\|Ax - b\|_2^2.$$

We say that a point (in the extended, augmented space), (x, μ), is primal-dual optimal iff

$$\begin{aligned} 0 &= \nabla_x \mathcal{L}(x, \mu) = \nabla f(x) + A^\top \mu + \rho A^\top (Ax - b), \\ 0 &= -\nabla_\mu \mathcal{L}(x, \mu) = b - Ax. \end{aligned}$$

One can also re-state the primal-dual optimality condition as,

$$T(x,\mu)=0,\quad T(x,\mu)=\begin{pmatrix}\nabla_x\mathcal{L}(x,\mu)\\ -\nabla_\mu\mathcal{L}(x,\mu)\end{pmatrix}$$

. Operator/function, T, is often called the Karush-Kuhn-Tucker (KKT) operator. (We may call T operator to emphasize that it maps a function, $f(x)$, to another function, $\nabla_x\mathcal{L}$.)

We are now ready to state the Primal-Dual Gradient (PDG) algorithm

$$\begin{pmatrix}x\\ \mu\end{pmatrix}_{k+1}=\begin{pmatrix}x\\ \mu\end{pmatrix}_{k}-\eta_k T(x_k,\mu_k).$$

Similar construction works if inequality constraints are added:

$$\begin{aligned}&f(x)\to\min\\ &\texttt{s.t.:}g_i(x)\le 0,\ 1\le i\le m.\end{aligned}$$

The augmented problem, accounting for the inequalities, becomes

$$\begin{aligned}f(x)+\frac{\rho}{2}\sum_{i=1}^m(g_i(x))_+^2&\to\min\\ \texttt{s.t.:}g_i(x)&\le 0,\ 1\le i\le m.\end{aligned}$$

Respective augmented Lagrangian is

$$\mathcal{L}(x,\lambda)=f(x)+\lambda^\top F(x)+\frac{\rho}{2}\|F(x)\|_2^2,$$

where $F(x)_i=(g_i(x))_+$. We say that the pair (x,λ) is primal-dual optimal iff

$$\begin{aligned}0=-\nabla_x\mathcal{L}(x,\lambda)&=\nabla f(x)+\sum_{i=1}^m(\lambda_i+\rho g_i(x)_+)(\nabla g_i(x))_+\\ 0=-\nabla_\lambda\mathcal{L}(x,\lambda)&=-F(x).\end{aligned}$$

PDG algorithm accounting for the inequality constraints is

$$\begin{pmatrix}x\\ \lambda\end{pmatrix}_{k+1}=\begin{pmatrix}x\\ \lambda\end{pmatrix}_{k}-\eta_k T(x_k,\lambda_k)$$

Convergence analysis of PDG algorithm repeats all steps involved in analysis of the original GD. The Lypunov exponent here is , $V(x,\lambda)=\|x_0-x^*\|_2^2+\|\lambda_0-\lambda^*\|_2^2$.

Exercise A.20. Analyze convergence of the PDG algorithm for convex optimization with inequality constraints assuming that all the functions involved (in the objective, f, and in the constraints, g_i) are convex and β-smooth.

Mirror Descent Algorithm

Our previous analysis was mostly focused on the case, where the objective function f is smooth in ℓ_2 norm and the distance from the starting point, where we initiate the algorithm, to the optimal point is measured in the ℓ_2 norm as well. From the perspective of the GD, the optimization over a unit simplex and the optimization over a unit Euclidean sphere are equivalent computational complexity-wise. On the other hand, the volume of the unit simplex is exponentially smaller than the volume of the unit sphere. Mirror Descent (MD) algorithm allows to explore geometry of the domain thus providing a faster algorithm for the case of the simplex. The acceleration is up to the $\sim \sqrt{d}$ factor, where d is the dimensionality of the underlying space.

We start with an unconstrained convex optimization problem:

$$f(x) \to \min$$
$$\texttt{s.t.:}\ x \in S \subseteq \mathbb{R}^n$$

Consider in more details an elementary iteration of the GD algorithm

$$x_{k+1} = x_k - \eta_k \nabla f(x_k).$$

From the mathematical perspective we sum up objects from different spaces: x belongs to the primal space, while the space where $\nabla f(x)$ resides, called the dual (conjugate) space may be different. To overcome this "inconsistency", Nemirovski and Yudin have proposed in 1978 the following algorithm:

$$\begin{aligned}
y_k &= \nabla\phi(x_k),\ \text{– map the point to a point in the dual space}\\
y_{k+1} &= y_k - \eta_k \nabla f(x_k),\ \text{– update the point in the dual space}\\
\bar{x}_{k+1} &= (\nabla\phi)^{-1}(y_{k+1}) = \nabla\phi^*(y_{k+1}),\ \text{– project the point back to the primal space}\\
x_{k+1} &= \Pi_C^{D_\phi}(\bar{x}_{k+1}) = \arg\min_{x\in C} D_\phi(x, \bar{x}_{k+1}),\ \text{project the point to a feasible set}
\end{aligned}$$

where $\phi(x)$ is a strongly convex function defined on $\mathbb{R}^n$ and $\nabla\phi(\mathbb{R}^n) = \mathbb{R}^n$; and $\phi^*(y) = \sup_{x\in\mathbb{R}^n}(y^\top x - \phi(x))$ is the Legendre Fenchel (LF) transform (conjugate function) of $\phi(x)$. Function ϕ is also called the *mirror map* function. $D_\phi(u,v) = \phi(u) - \phi(v) - \nabla\phi(v)^\top(u-v)$ is the so-called Bregman divergence

$$D_\phi(u,v) = \phi(u) - \phi(v) - \nabla\phi(v)^\top(u-v),$$

which measures (for strictly convex function ϕ) the distance between $\phi(u)$ and its linear approximation $\phi(v) - \nabla\phi(v)^\top(u-v)$ evaluated at v.

Exercise A.21. Let $\phi(x)$ be a strongly convex function on $\mathbb{R}^n$. Using the definition of the conjugate function prove that $\nabla\phi^*(\nabla\phi(x)) = x$, where ϕ^* is a conjugate function to ϕ.

The Bregman divergence has a number of attractive properties:

- *Non-negativity.* $D_\phi(u, v) \geq 0$ for any convex function ϕ.
- *Convexity in the first argument.* The Bregman divergence $D_\phi(u, v)$ is convex in its first argument. (Notice that is not necessarily convex in the second argument.)
- *Linearity* with respect to the non-negative coefficients. In other words, for any strictly convex ϕ and ψ we observe:

$$D_{\lambda\phi+\mu\psi}(u, v) = \lambda D_\phi(u, v) + \mu D_\psi(u, v).$$

- *Duality.* Let function ϕ has a convex conjugate ϕ^*, then

$$D_{\phi^*}(u^*, v^*) = D_\phi(u, v), \text{ with } u^* = \nabla\phi(u), \text{ and } v^* = \nabla\phi(v).$$

Examples of the Bregman divergence are

- *Euclidean norm.* Let $\phi = \|x\|_2^2$, then $D_\phi(x, y) = \|x\|_2^2 - \|y\|_2^2 - 2y^\top(x - y) = \|x - y\|_2^2$.
- *Negative entropy.* $\phi(x) = \sum_{i=1}^n x_i \ln x_i$, $f : \mathbb{R}^n_{++} \to \mathbb{R}$. Then

$$D_\phi(x, y) = \sum_{i=1}^n x_i \ln(x_i/y_i) - \sum_{i=1}^n x_i + \sum_{i=1}^n y_i = D_{KL}(x||y),$$

  where $D_{KL}(x||y)$ is the so called Kullback-Leibler (KL) divergence.
- *Lower and upper bounds.* Let ϕ be a μ-strongly convex function with respect to a norm $\|\cdot\|$ then

$$D_\phi(x, y) \geq \frac{\mu}{2}\|x - y\|^2, \qquad D_\phi(x, y) \leq \frac{\beta}{2}\|x - y\|^2$$

The following statement represents an important fact which will be used below to analyze the MD algorithm.

Theorem A.4.2 (Pinsker Inequality)**.** For any x, y, such that $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i = 1$, $x \geq 0$, $y \geq 0$ one get the following KL divergence estimate, $D_{KL}(x||y) \geq \frac{1}{2}\|x - y\|_1^2$.

An immedite corollary of the Theorem is that $\phi(x) = \sum_{i=1}^n x_i \ln x_i$ is 1-strongly convex in ℓ_1 norm:

$$\phi(y) \geq \phi(x) + \nabla\phi(x)^\top (y-x) + D_{KL}(y||x) \geq \phi(x) + \nabla\phi(x)^\top (y-x) + \frac{1}{2}\|x-y\|_1^2$$

The proximal form of the MD algorithm is

$$x_{k+1} = \Pi_C^{D_\phi}\left(\arg\min_{x\in\mathbb{R}^n}\left\{f(x_k) + \nabla f(x_k)^\top (x-x_k) + \frac{1}{\eta_k} D_\phi(x, x_k)\right\}\right),$$

where $\Pi_S^{D_\phi}(y) = \arg\min_{x\in S} D_\phi(x,y)$.

Example A.4.3. Consider the following optimization problem over the unit simplex:

$$f(x) \to \min_{x\in\mathbb{R}^n}$$
$$\texttt{s.t.:} x \in S = \{x : x^\top 1 = 1, x \in \mathbb{R}^n_{++}\}.$$

Let the distance generating function $\phi(x)$ be a negative entropy, $\phi(x) = \sum_{i=1}^n x_i \ln x_i$. Then the MD algorithm update becomes

$$x_{k+1} = \Pi_S^{D_\phi}\left\{\arg\min_x\left\{f(x_k) + \nabla f(x_k)^\top (x-x_k) + \frac{1}{\eta_k} D_\phi(x, x_k)\right\}\right\},$$

where $D_\phi(x,y) = \sum_{i=1}^n x_i \ln(x_i/y_i) - (x_i - y_i)$. The resulting optimal x is

$$\nabla\phi(x) = \nabla\phi(x_k) - \eta_k \nabla f(x_k), \text{ that is } y_i = (x_k)_i \exp(-\eta_k \nabla f(x_k)_i).$$

One observes that the Bregman projection onto the simplex is a renormalization: $\Pi_S^{D_\phi} = y/\|y\|_1$. This results in the following expression for the MD update:

$$(x_k)_i = \frac{(x_k)_i \exp(-\eta_k \nabla f(x_k)_i)}{\sum_{j=1}^n (x_k)_j \exp(-\eta_k \nabla f(x_k)_j)}.$$

Let us sketch the continuous time analysis of the MD algorithm in the case of the β-smooth convex functions. In contrast with the GD analysis, it is more appropriate to work in this case with the Lyapunov's function in the dual space:

$$V(Z(t)) = D_{\phi^*}(Z(t), z^*), \qquad Z(t) = \nabla\phi(X(t)),$$

where ϕ is a strongly convex distance generating function. According to the definition of the Bregman divergence, one derives

$$\begin{aligned}\frac{d}{dt}V(Z(t)) &= \frac{d}{dt}D_{\phi^*}(Z(t), z^*) = \frac{d}{dt}\left\{\phi^*(Z(t)) - \phi^*(z^*) - \nabla\phi^*(z^*)^\top (Z(t) - z^*)\right\} \\ &= (\nabla\phi^*(Z(t)) - \nabla\phi^*(z^*), \dot{Z}(t)) = (X(t) - x^*)^\top \dot{Z}(t).\end{aligned}$$

Given that $\dot{Z}(t) = -\nabla f(X)$ one derives

$$\frac{d}{dt}V(Z(t)) = -\nabla f(X(t))^{\top}(X(t) - x^*) \leq -(f(X(t)) - f^*).$$

Integrating both sides of the inequality one arrives at

$$V(Z(t)) - V(Z(0)) \geq \int_0^t f(X(\tau))d\tau - tf^* \geq t\left(f\left(\frac{1}{t}\int_0^t X(\tau)d\tau\right) - f^*\right),$$

where the last transformation is due to the Jensen inequality. Therefore, similarly to the case of GD, the convergence rate of the MD algorithm is $O(1/k)$. The resulting MD ODE is

$$\begin{cases} X(t) & = \nabla\phi^*(Z(t)) \\ \dot{Z}(t) & = -\nabla f(X(t)) \\ X(0) & = x_0, Z(0) = z_0 \text{ with } \nabla\phi^*(z_0) = x_0. \end{cases}$$

Behavior of the MD, when applied to a non-smooth convex function, repeats the one of the GD: the convergence rate is $O(1/\sqrt{k})$ in this case.

Bibliography

[1] M. Tabor, *Principles and Methods of Applied Mathematics.* University of Arizona Press, 1999.

[2] V. Arnold, *Ordinary Differential Equations.* The MIT Press, 1973.

[3] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[4] J. Calder, "The calculus of variations (lecture notes)," http://www-users.math.umn.edu/~jwcalder/CalculusOfVariations.pdf, 2019.

[5] B. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1 – 17, 1964.

[6] Y. E. Nesterov, "A method for solving the convex programming problem with convergence rate $o(1/k^2)$," *Dokl. Akad. Nauk SSSR*, vol. 269, pp. 543–547, 1983.

[7] W. Su, S. Boyd, and E. J. Candes, "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights," *arXiv:1503.01243*, 2015.

[8] A. C. Wilson, B. Recht, and M. I. Jordan, "A Lyapunov Analysis of Momentum Methods in Optimization," *arXiv:1611.02635*, 2016.

[9] M. Levi, *Classical Mechanics with Calculus of Variations and Optimal Control: An Intuitive Introduction.* AMS, 2014.

[10] H. Touchette, "Legendre-fenchel transforms in a nutshell," 2014.

[11] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical Imaging and Vision*, vol. 20, pp. 89–97, 2004.

[12] R. K. P. Zia, E. F. Redish, and S. R. McKay, "Making sense of the legendre transform," *American Journal of Physics*, vol. 77, no. 7, p. 614–622, Jul 2009. [Online]. Available: http://dx.doi.org/10.1119/1.3119512

[13] L. Pontryagin, V. Boltayanskii, R. Gamkrelidze, and E. Mishchenko, *The mathematical theory of optimal processes (translated from Russian in 1962).* Wiley, 1956.

[14] A. T. Fuller, "Bibliography of pontryagm's maximum principle," *Journal of Electronics and Control*, vol. 15, no. 5, pp. 513–517, 1963.

[15] R. Bellman, "On the theory of dynamic programming," *PNAS*, vol. 38, no. 8, p. 716, 1952.

[16] C. Moore and S. Mertens, *The Nature of Computation.* New York, NY, USA: Oxford University Press, 2011.

[17] A. Sinclair, "Uc berkley, cs271 "randomness & computation" course," 2020. [Online]. Available: https://people.eecs.berkeley.edu/~sinclair/cs271/n13.pdf

[18] C. E. Shannon, "Prediction and entropy of printed english," *The Bell System Technical Journal*, vol. 30, no. 1, pp. 50–64, 1951.

[19] D. J. C. Mackay, *Information theory, inference, and learning algorithms.* Cambridge University Press, 2003.

[20] N. V. Kampen, *Stochastic processes in physics and chemistry.* North Holland, 2007.

[21] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* The MIT Press, 2018.

[22] F. Kelly and E. Yudovina, *Stochastic Networks*, ser. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2014.

[23] D. B. Wilson, "Perfectly random sampling with markov chains," http://www.dbwilson.com/exact/.

[24] T. Richardson and R. Urbanke, *Modern Coding Theory.* USA: Cambridge University Press, 2008.

[25] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2282–2312, 2005.

[26] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1, pp. 1–305, 2008.

[27] V. Likhosherstov, Y. Maximov, and M. Chertkov, "Inference and sampling of k_{33}-free ising models," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 3963–3972. [Online]. Available: http://proceedings.mlr.press/v97/likhosherstov19a.html

[28] A. Y. Lokhov, M. Vuffray, S. Misra, and M. Chertkov, "Optimal structure and parameter learning of ising models," *Science Advances*, vol. 4, no. 3, 2018. [Online]. Available: https://advances.sciencemag.org/content/4/3/e1700791

[29] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[30] G. Strang, *Linear Algebra and Learning from Data*. Wellesley-Cambridge Press, 2019.

[31] P. Kidger and T. Lyons, "Universal Approximation with Deep Narrow Networks," in *Proceedings of Thirty Third Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, J. Abernethy and S. Agarwal, Eds., vol. 125. PMLR, 09–12 Jul 2020, pp. 2306–2327. [Online]. Available: https://proceedings.mlr.press/v125/kidger20a.html